Information Contraction and Decomposition

by

Anuran Makur

B.S., Electrical Engineering and Computer Sciences
University of California, Berkeley, 2013

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2015

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Signature of Author: _____

Anuran Makur
Department of Electrical Engineering and Computer Science
May 21, 2019

Certified by: _____

Lizhong Zheng
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by: _____

Yury Polyanskiy
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: _____

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Information Contraction and Decomposition

by

Anuran Makur

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2019 in partial fulfillment of the requirements for the degree of
Doctor of Science in Electrical Engineering and Computer Science

Abstract

Information contraction is one of the most fundamental concepts in information theory as evidenced by the numerous classical converse theorems that utilize it. In this dissertation, we study several problems aimed at better understanding this notion, broadly construed, within the intertwined realms of information theory, statistics, and discrete probability theory.

In information theory, the contraction of f-divergences, such as Kullback-Leibler (KL) divergence, χ^2-divergence, and total variation (TV) distance, through channels (or the contraction of mutual f-information along Markov chains) is quantitatively captured by the well-known data processing inequalities. These inequalities can be tightened to produce "strong" data processing inequalities (SDPIs), which are obtained by introducing appropriate channel-dependent or source-channel-dependent "contraction coefficients." We first prove various properties of contraction coefficients of source-channel pairs, and derive linear bounds on specific classes of such contraction coefficients in terms of the contraction coefficient for χ^2-divergence (or the Hirschfeld-Gebelein-Rényi maximal correlation). Then, we extend the notion of an SDPI for KL divergence by analyzing when a q-ary symmetric channel dominates a given channel in the "less noisy" sense. Specifically, we develop sufficient conditions for less noisy domination using ideas of degradation and majorization, and strengthen these conditions for additive noise channels over finite Abelian groups. Furthermore, we also establish equivalent characterizations of the less noisy preorder over channels using non-linear operator convex f-divergences, and illustrate the relationship between less noisy domination and important functional inequalities such as logarithmic Sobolev inequalities.

Next, adopting a more statistical and machine learning perspective, we elucidate the elegant geometry of SDPIs for χ^2-divergence by developing modal decompositions of bivariate distributions based on singular value decompositions of conditional expectation operators. In particular, we demonstrate that maximal correlation functions meaningfully decompose the information contained in categorical bivariate data in a local information geometric sense and serve as suitable embeddings of this data into

Euclidean spaces. Moreover, we propose an extension of the well-known alternating conditional expectations algorithm to estimate maximal correlation functions from training data for the purposes of feature extraction and dimensionality reduction. We then analyze the sample complexity of this algorithm using basic matrix perturbation theory and standard concentration of measure inequalities. On a related but tangential front, we also define and study the information capacity of permutation channels.

Finally, we consider the discrete probability problem of broadcasting on bounded indegree directed acyclic graphs (DAGs), which corresponds to examining the contraction of TV distance in Bayesian networks whose vertices combine their noisy input signals using Boolean processing functions. This generalizes the classical problem of broadcasting on trees and Ising models, and is closely related to results on reliable computation using noisy circuits, probabilistic cellular automata, and information flow in biological networks. Specifically, we establish phase transition phenomena for random DAGs which imply (via the probabilistic method) the existence of DAGs with logarithmic layer size where broadcasting is possible. We also construct deterministic DAGs where broadcasting is possible using expander graphs in deterministic quasi-polynomial or randomized polylogarithmic time in the depth. Lastly, we show that broadcasting is impossible for certain two-dimensional regular grids using techniques from percolation theory and coding theory.

Thesis Supervisor: Lizhong Zheng
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Yury Polyanskiy
Title: Associate Professor of Electrical Engineering and Computer Science

Dedicated to my parents,
Anamitra and Anindita Makur,
and my little sister,
Anyatama Makur.

Acknowledgments

Expressing my sincere gratitude to everyone who has contributed indispensably to my arduous graduate school life and this dissertation is a task so gargantuan that I choose not to undertake it. To be honest, I feel the individuals who have mattered most to me would not derive much meaning from a painfully effusive "thank you note" or a soulless enumeration of their contributions, and are most certainly already aware of my deep appreciation and respect for them. So, I will modestly and succinctly recognize the most important people with the hope that you understand how much you have meant to me.

First, I thank my dear parents, Anamitra and Anindita Makur, and my little sister, Anyatama Makur. To put it simply, I am because you are. You are my spine, and to you, I dedicate this dissertation.

Second, I thank my doctoral thesis advisers, Profs. Yury Polyanskiy and Lizhong Zheng. You have been my parents within the academic sphere at MIT. I do not use this metaphor lightly; it is the highest compliment for a mentor in my book. It was an honor to be your student.

Third, I thank Prof. Elchanan Mossel, my doctoral thesis committee member, and Prof. Gregory Wornell, under whose warm guidance I served as a teaching assistant. You have both contributed immeasurably to my development as a researcher, and for that I am greatly indebted to you.

Fourth, I acknowledge several professors with whom I have had many meaningful and stimulating interactions. Within the MIT faculty, I thank Profs. Guy Bresler, Alan Edelman, Muriel Médard, Alan Oppenheim, and Devavrat Shah. Outside of MIT, I thank Profs. Venkat Anantharam and Afonso Bandeira. In particular, I am grateful to Venkat for inspiring me to pursue fundamental theoretical research during my undergraduate days at UC Berkeley.

Fifth, I acknowledge the friends who made my grueling stint as a doctoral student not only bearable, but also quite enjoyable (albeit slightly less productive). I thank my friends at MIT who have been with me every step of the way: Ganesh Ajjanagadde, Mohamed AlHajri, Nirav Bhan, Austin Collins, Eren Kizildag, Suhas Kowshik, Fabián Kozynski, Tarek Lahlou, Dheeraj Nagaraj, James Noraky, David Qiu, Ankit Rawat, Arman Rezaee, Amir Salimi, Tuhin Sarkar, James Thomas, and Christos Thrampoulidis. I thank my undergraduate friends from UC Berkeley who have kept in

touch with me for so many years: Joyjit Daw, Ankush Gupta, Govind Ramnarayan, Aniket Soneji, Sibi Venkatesan, Aditya Venkatramani, and Eric Zhan. I thank my old friends from high school who have walked the doctoral path with me: Sidharth Gupta, Gaurav Kankanhalli, Ashwin Kumar, and SangJin Lee. (If I have forgotten someone in my haste, forgive my weary mind.)

Finally, I acknowledge the outstanding community of graduate students, postdoctoral researchers, and administrative staff associated with the Polyanskiy and Zheng labs during my years at MIT. Specifically, I thank Ziv Goldfeld, Shao-Lun Huang, Wasim Huleihel, Or Ordentlich, and Hajir Roozbehani for a myriad of intellectually refreshing conversations.

Contents

List of Figures

Notation and Abbreviations

General Notation

$\mathbb{R}$	set of all real numbers
$\mathbb{C}$	set of all complex numbers
$\mathbb{N}$	set of all natural numbers $\{1, 2, 3, \dots\}$
$\mathbb{F}_2$	Galois field of order 2
$\mathbb{R}^n$	set of all n-dimensional real column vectors with $n \in \mathbb{N}$
$\mathbb{C}^n$	set of all n-dimensional complex column vectors with $n \in \mathbb{N}$
$\mathbb{F}_2^n$	set of all n-dimensional binary column vectors with $n \in \mathbb{N}$
$\mathbf{0}$	column vector (of appropriate dimension) with all 0 entries
$\mathbf{1}$	column vector (of appropriate dimension) with all 1 entries
e_i	ith standard basis column vector in $\mathbb{R}^n$ with $i \in \{1, \dots, n\}$ for some $n \in \mathbb{N}$, which has unity at the ith entry and zero elsewhere
$\lceil \cdot \rceil$	ceiling function
$\lfloor \cdot \rfloor$	floor function
$\mathrm{Re}\{\cdot\}$	real part of input argument
$\mathrm{Im}\{\cdot\}$	imaginary part of input argument
$\mathcal{U}$, $\mathcal{X}$, $\mathcal{Y}$	finite sets with cardinality greater than or equal to 2
$[n]$	set of first $n \in \mathbb{N}$ non-negative integers $\{0, \dots, n-1\}$
$\log(\cdot)$	natural logarithm with base e
$\exp(\cdot)$	natural exponential with base e
$\mathsf{conv}(\cdot)$	convex hull of input set of vectors
$\mathrm{supp}(\cdot)$	support of a real-valued function
$\operatorname{ess\,supp}(\cdot)$	essential support of a Borel measurable real-valued function with respect to the Lebesgue measure on $\mathbb{R}$
∇f	gradient $\nabla f : S \to \mathbb{R}^n$ of a function $f : S \to \mathbb{R}$ with open domain $S \subseteq \mathbb{R}^n$ in denominator layout notation, where $n \in \mathbb{N}$
$\nabla^2 f$	Hessian $\nabla^2 f : S \to \mathbb{R}^{n \times n}$ of a function $f : S \to \mathbb{R}$ with open domain $S \subseteq \mathbb{R}^n$ in denominator layout notation, where $n \in \mathbb{N}$
$(\mathsf{mod}\ n)$	modulo operation with modulus $n \in \mathbb{N}$

Matrix Theory

$(\mathbb{R}^n)^*$	set of all n-dimensional real row vectors with $n \in \mathbb{N}$
$\mathbb{R}^{m\times n}$	set of all real $m \times n$ matrices with $m, n \in \mathbb{N}$
$\mathbb{R}^{n\times n}_{\mathsf{sym}}$	set of all real $n \times n$ symmetric matrices with $n \in \mathbb{N}$
$\mathbb{R}^{n\times n}_{\succeq 0}$	closed convex cone of all real $n \times n$ symmetric positive semidefinite matrices with $n \in \mathbb{N}$
$\mathcal{V}_k(\mathbb{R}^n)$	Stiefel manifold of all orthonormal k-frames in $\mathbb{R}^n$ (i.e. all matrices in $\mathbb{R}^{n\times k}$ with orthonormal columns) with $k \in \{1, \ldots, n\}$ and $n \in \mathbb{N}$; $\mathcal{V}_n(\mathbb{R}^n)$ is the orthogonal group
$\mathbb{C}^{m\times n}$	set of all complex $m \times n$ matrices with $m, n \in \mathbb{N}$
$\mathbb{C}^{n\times n}_{\mathsf{Herm}}$	set of all complex $n \times n$ Hermitian matrices with $n \in \mathbb{N}$
$\mathcal{V}_k(\mathbb{C}^n)$	Stiefel manifold of all orthonormal k-frames in $\mathbb{C}^n$ (i.e. all matrices in $\mathbb{C}^{n\times k}$ with orthonormal columns) with $k \in \{1, \ldots, n\}$ and $n \in \mathbb{N}$; $\mathcal{V}_n(\mathbb{C}^n)$ is the unitary group
$\mathbb{F}_2^{m\times n}$	set of all binary $m \times n$ matrices with $m, n \in \mathbb{N}$
A^T	transpose of a matrix A
A^H	Hermitian transpose/conjugate transpose/adjoint of a matrix A
A^{-1}	inverse of a square non-singular matrix A
$A^\dagger$	Moore-Penrose pseudoinverse of a matrix A
$A^{\frac{1}{2}}$	unique positive semidefinite square root matrix of a positive semidefinite matrix $A \in \mathbb{R}^{n\times n}_{\succeq 0}$
$[A]_{j,k}$	(j,k)th entry of a matrix $A \in \mathbb{C}^{m\times n}$ with $j \in \{1, \ldots, m\}$ and $k \in \{1, \ldots, n\}$
$\rho(\cdot)$	spectral radius (or largest eigenvalue modulus) of a square matrix
$\mu(\cdot)$	second largest eigenvalue modulus of a square matrix
$\sigma_i(\cdot)$	ith largest singular value of a matrix in $\mathbb{R}^{m\times n}$ for any $i \in \{1, \ldots, \min\{m, n\}\}$
$\mathrm{tr}(\cdot)$	trace operator of a square matrix
$\mathcal{R}(\cdot)$	range/image/column space of a matrix
$\mathcal{K}(\cdot)$	kernel/nullspace of a matrix
I	identity matrix (of appropriate dimension)
I_n	$n \times n$ identity matrix with $n \in \mathbb{N}$
$\succeq_{\mathsf{PSD}}$	Löwner partial order over Hermitian matrices; for any two matrices $A, B \in \mathbb{C}^{n\times n}_{\mathsf{Herm}}$, $A \succeq_{\mathsf{PSD}} B$ if and only if $A - B \succeq_{\mathsf{PSD}} 0$ (where 0 is the zero matrix) if and only if $A - B$ is positive semidefinite
$\mathsf{diag}(\cdot)$	diagonal (square) matrix with input (row or column) vector on its principal diagonal
$\|\cdot\|_p$	ℓ^p-norm of a row or column vector, or induced operator norm of a matrix where the input and output vector spaces are equipped with the ℓ^p-norm, where $p \in [1, \infty]$

$\|\cdot\|_{\mathsf{Fro}}$	Frobenius or Hilbert-Schmidt norm of a matrix
$\|\cdot\|_{\mathsf{op}}$	operator or spectral norm of a matrix or linear map
$\|\cdot\|_{(p,k)}$	(p,k)-norm of a matrix in $\mathbb{R}^{m\times n}$ where $p \in [1,\infty]$ and $k \in \{1,\ldots,\min\{m,n\}\}$, as defined in (C.3) in appendix C.1

Probability Theory

pmf	probability mass function
pdf	probability density function with respect to Lebesgue measure on $\mathbb{R}$
a.s.	almost surely (with respect to appropriate probability measure)
P-a.s.	almost surely with respect to the probability measure associated with P (where P could be a pdf or a random variable)
i.i.d.	independent and identically distributed
X_i^j	subsequence of random variables $(X_i,\ldots,X_j)$ for $i,j \in \mathbb{N}$ and $i \leq j$
$\mathcal{P}_{\mathcal{X}}$	probability simplex of pmfs of a random variable X on the alphabet $\mathcal{X}$, i.e. $\{P_X \in (\mathbb{R}^{\lvert\mathcal{X}\rvert})^* : P_X \geq 0$ entry-wise and $P_X\mathbf{1} = 1\}$
$\mathcal{P}_{\mathcal{X}}^{\circ}$	relative interior of $\mathcal{P}_{\mathcal{X}}$, i.e. $\{P_X \in \mathcal{P}_{\mathcal{X}} : P_X > 0$ entry-wise$\}$
$\mathcal{P}_n$	probability simplex in $(\mathbb{R}^n)^*$ with $n \in \mathbb{N}$, i.e. $\{P \in (\mathbb{R}^n)^* : P \geq 0$ entry-wise and $P\mathbf{1} = 1\}$
$\mathcal{P}_n^{\circ}$	relative interior of $\mathcal{P}_n$ with $n \in \mathbb{N}$, i.e. $\{P \in \mathcal{P}_n : P > 0$ entry-wise$\}$
$\mathcal{P}_{\mathcal{Y}\lvert\mathcal{X}}$	convex set of $\lvert\mathcal{X}\rvert\times\lvert\mathcal{Y}\rvert$ row stochastic matrices in $\mathbb{R}^{\lvert\mathcal{X}\rvert\times\lvert\mathcal{Y}\rvert}$ (whose rows belong to $\mathcal{P}_{\mathcal{Y}}$)
$\mathbb{R}_{\mathsf{sto}}^{m\times n}$	convex set of $m\times n$ row stochastic matrices with $m,n \in \mathbb{N}$ (whose rows belong to $\mathcal{P}_n$)
$\mathcal{L}^2(\mathcal{X},P_X)$	separable Hilbert space of square integrable real-valued functions on the domain $\mathcal{X}$ endowed with an inner product defined by the probability distribution $P_X \in \mathcal{P}_{\mathcal{X}}$
$\mathbf{u}$	uniform probability mass function (pmf)
Δ_x	Kronecker delta pmf in $\mathcal{P}_{\mathcal{X}}$ with $x \in \mathcal{X}$ such that $\Delta_x(x) = 1$ and $\Delta_x(x') = 0$ for $x' \in \mathcal{X}\backslash\{x\}$
$\mathbb{1}\{\cdot\}$	indicator/characteristic function which equals 1 if its input proposition is true and 0 if its input proposition is false
$\mathbb{P}(\cdot)$	probability measure
$\mathbb{E}[\cdot]$	expectation operator
$\mathbb{E}_P[\cdot]$	expectation operator with respect to distribution P
$\mathbb{VAR}(\cdot)$	variance operator
$\mathbb{COV}(\cdot,\cdot)$	covariance operator
$X \perp\!\!\!\perp Y$	random variables X and Y are independent
$\mathcal{N}(\mu,\sigma^2)$	Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \geq 0$

$\mathsf{Bernoulli}(p)$	distribution of a Bernoulli random variable that equals 1 with probability $p \in [0,1]$ and 0 with probability $1-p$
$\mathsf{binomial}(n,p)$	distribution of a binomial random variable (which is a sum of n i.i.d. $\mathsf{Bernoulli}(p)$ random variables) with $n \in \mathbb{N}$ trials and success probability $p \in [0,1]$

Information Theory

$\mathsf{BSC}(\delta)$	binary symmetric channel with crossover probability $\delta \in [0,1]$ defined in (2.42) in chapter 2
$\mathsf{DSBS}(\delta)$	doubly symmetric binary source with parameter $\delta \in [0,1]$, which is the distribution of two uniform Bernoulli random variables (X,Y) where X is passed through a $\mathsf{BSC}(\delta)$ to produce Y
$\mathsf{BEC}(\delta)$	binary erasure channel with erasure probability $\delta \in [0,1]$ defined in (4.145) in chapter 4
w_δ	q-ary symmetric channel noise pmf in $\mathcal{P}_q$ with total crossover probability $\delta \in [0,1]$ defined in (3.18) in chapter 3, where $q \in \mathbb{N}$ and $q \geq 2$
W_δ	q-ary symmetric channel matrix in $\mathbb{R}_{\mathsf{sto}}^{q \times q}$ with total crossover probability $\delta \in [0,1]$ defined in (3.19) in chapter 3, where $q \in \mathbb{N}$ and $q \geq 2$
E_ϵ	$\lvert\mathcal{X}\rvert$-ary erasure channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{X} \cup \{\mathsf{e}\}$ that either erases its input and outputs the erasure symbol e with erasure probability $\epsilon \in [0,1]$ or copies its input with probability $1-\epsilon$

Bachmann-Landau Asymptotic Notation

Consider any two real-valued functions $f(\cdot)$ and $g(\cdot)$ with domain $\mathbb{R}$ or $\mathbb{N}$ (in which case f and g are sequences) such that g is strictly positive.

Little-o notation: $f(\epsilon) = o(g(\epsilon))$ if and only if:

$$\lim_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} = 0\,,$$

where $\epsilon \in \mathbb{R}$;
$f(n) = o(g(n))$ if and only if:

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} = 0\,,$$

where $n \in \mathbb{N}$ and we sometimes use the indices $K, k, q \in \mathbb{N}$ instead of n.

Little-ω notation: $f(n) = \omega(g(n))$ if and only if:

$$\lim_{n\to\infty} \frac{g(n)}{f(n)} = 0\,,$$

where we also assume that f is strictly positive.

Big-O notation: $f(n) = O(g(n))$ if and only if:

$$\limsup_{n\to\infty} \frac{|f(n)|}{g(n)} < +\infty\,.$$

Big-Ω notation: $f(n) = \Omega(g(n))$ if and only if:

$$\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0\,,$$

where we also assume that f is strictly positive.

Big-Θ notation: $f(n) = \Theta(g(n))$ if and only if:

$$0 < \liminf_{n\to\infty} \frac{f(n)}{g(n)} \leq \limsup_{n\to\infty} \frac{f(n)}{g(n)} < +\infty\,,$$

where we also assume that f is strictly positive.

Chapter 1

Introduction

INFORMATION contraction is a foundational concept in information theory and other related fields. For instance, the concept, broadly construed, is intimately related to the second law of thermodynamics in statistical physics, which is one of the most fundamental laws of nature (see [53, Section 4.4], [200, Section 3.3], [230, Section 1.3]). This dissertation attempts to develop a finer understanding of information contraction from various perspectives in information theory, statistics, and discrete probability theory.

The notion of information contraction is quantitatively captured by the well-known data processing inequality (DPI). The most rudimentary form of this inequality states that mutual information contracts across any Markov chain $U \to X \to Y$:

$$I(U;Y) \leq I(U;X) \tag{1.1}$$

where $I(\cdot;\cdot)$ denotes the mutual information between its two input random variables. Although this form of the DPI suffices for introductory expositions of information theory, in this dissertation, we consider the following more general version of the DPI to enable a more sophisticated mathematical treatment of information contraction. For every fixed channel or Markov transition kernel $W \in P_{\mathcal{Y}|\mathcal{X}}$, the f-divergence between any two probability distributions $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$ contracts after the distributions are pushed forward through W:

$$D_f(R_X W||P_X W) \leq D_f(R_X||P_X) \tag{1.2}$$

where $R_X W, P_X W \in \mathcal{P}_{\mathcal{Y}}$, and $D_f(\cdot||\cdot)$ denotes the f-divergence between its two input distributions (see Definition 2.1 in chapter 2). Intuitively, the f-divergence is a general notion of "distance" between distributions, and the general DPI in (1.2) depicts that distinguishing (or binary hypothesis testing) between R_X and P_X based on observations becomes more difficult when the observations are corrupted by noise from the channel W.

The DPI is undoubtedly one the most fundamental inequalities in information theory as evidenced by the numerous classical converse theorems that rely on it, cf. [200, Section 3.3]. Yet, there are several simple questions it does not answer. For example, when W represents the stochastic transition probability matrix of a discrete-time ergodic Markov chain, a natural question to pose is: What is the rate at which the distribution over the

states converges to the stationary distribution of the Markov chain? Evidently, the DPI cannot even establish convergence to stationarity, let alone determine the rate of this convergence. To address such questions, we require sharpened versions of DPIs known as "strong" data processing inequalities (SDPIs), which have recently been rediscovered and applied in several disciplines (see e.g. [231, Section 2]). In particular, SDPIs are obtained by maximally tightening DPIs for f-divergences using either source-channel-dependent or only channel-dependent constants known as "contraction coefficients." These astonishingly deep and important constants have inspired several key questions in this dissertation.

1.1 Organization of Thesis

We now concisely delineate the contents of the different chapters in this thesis. The majority of the results in this thesis can be divided into three complementary parts. Each part studies the notion of information contraction from the lens of a different discipline. Furthermore, each part can be construed as delving into the SDPIs and contraction coefficients corresponding to one of three salient specializations of f-divergences: Kullback-Leibler (KL) divergence, χ^2-divergence, and total variation (TV) distance.

1.1.1 Information Theory

The first part of this dissertation, which is comprised of chapters 2 and 3, adopts an information theoretic perspective. It studies the relationships between contraction coefficients for different f-divergences, and specifically, extends the SDPI for KL divergence by exploiting its relation to well-known information theoretic preorders over channels.

In chapter 2, we first provide a fairly self-contained survey of the classical literature on contraction coefficients, and then present some relevant new results and variations of known results. For instance, it is well-known that for any discrete source-channel pair, the contraction coefficients for a large class of f-divergences are lower bounded by the contraction coefficient for χ^2-divergence, cf. [189, Theorem 5], [234, Theorem III.3], [231, Theorem 2]. We elucidate that this lower bound can be achieved by driving the input f-divergences of the contraction coefficients to zero. Furthermore, we establish a linear upper bound on the contraction coefficients for a certain class of f-divergences using the contraction coefficient for χ^2-divergence, and refine this upper bound for the salient special case of KL divergence. Lastly, we present an alternative proof of the fact that the contraction coefficients for KL and χ^2-divergences are equal for a Gaussian source with an additive Gaussian noise channel, cf. [82, Theorem 7], [217, p.2], [152, Section 5.2, part 5], where we additionally allow the former contraction coefficient to be power constrained. Several proofs of the results in chapter 2 are contained in appendix A.

Recently, it was shown that the contraction coefficient for KL divergence of a given channel $V \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ can be characterized by the extremal q-ary erasure channel that dominates V in the precise sense of being "less noisy" (where $q = |\mathcal{X}|$) [231, Section 6]. Inspired by this result, chapter 3 extends the notion of SDPIs by studying the basic

question of whether a given channel V can be dominated by a q-ary symmetric channel. The concept of less noisy ordering between channels originated in network information theory (in the context of broadcast channels) and is defined in terms of mutual information or KL divergence [156]. We provide an equivalent characterization of it in terms of any f-divergence corresponding to a non-linear operator convex function f. This generalizes the well-known result that contraction coefficients of channels are equal for all f-divergences with non-linear operator convex functions f [46, Theorem 1]. Furthermore, we develop a simple criterion for domination by a q-ary symmetric channel in terms of the minimum entry of the stochastic matrix defining the channel V. The criterion is strengthened for the special case of additive noise channels over finite Abelian groups. Finally, we prove that domination by a q-ary symmetric channel implies (via comparison of Dirichlet forms) a logarithmic Sobolev inequality for the original channel. This develops a concrete connection between information theoretic preorders over channels and important functional inequalities which are cornerstones of the modern approach to studying ergodicity and hypercontractivity of Markov processes, isoperimetry, and concentration of measure. We note that several proofs of the results in chapter 3, as well as pertinent supplementary results for chapter 3, are contained in appendix B.

■ 1.1.2 Statistics and Machine Learning

The second part of this dissertation, which is comprised of chapter 4, adopts a statistical and machine learning perspective. It unveils the elegant geometry underlying the SDPI for χ^2-divergence, and demonstrates how this geometric understanding enables us to decompose mutual χ^2-information and perform feature extraction and dimensionality reduction for high-dimensional inference tasks.

One of the central ideas examined and extended in chapter 4 is the notion of Hirschfeld-Gebelein-Rényi maximal correlation [96, 125, 236, 242], which is a measure of true statistical dependence between two random variables. It is well-known in the literature that for a fixed source-channel pair $P_X \in \mathcal{P}_{\mathcal{X}}$ and $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, the squared maximal correlation between the random variables X and Y is equal to the contraction coefficient for χ^2-divergence [242]. Furthermore, the maximal correlation between X and Y is a singular value of the conditional expectation operator, or equivalently, the divergence transition matrix (DTM), cf. Definition 4.1, defined by the joint distribution $P_{X,Y}$ [125, 236]. Therefore, the underlying singular value decomposition structure of the conditional expectation operator or the DTM corresponding to $P_{X,Y}$ can be used to understand the geometry of the SDPI for χ^2-divergence. In chapter 4, we explicate this "modal decomposition" structure of the bivariate distribution $P_{X,Y}$. Although such modal decompositions have formed the basis of correspondence analysis [24, 125] and the theory of Lancaster distributions [165], our development reveals their vital role in modern data science and machine learning applications with categorical bivariate data. Specifically, we illustrate that maximal correlation functions (i.e. singular vectors of conditional expectation operators) serve as feature functions that meaningfully decom-

pose information in categorical data in a local information geometric sense and suitably embed categorical data into Euclidean spaces. In the process of demonstrating this, we also establish important properties of conditional expectation operators and DTMs, and characterize the set of all DTMs. Finally, we propose an extension of the so called alternating conditional expectations algorithm, cf. [35], to efficiently learn maximal correlation functions from training data, and perform various kinds of sample complexity analysis for this algorithm. We note that proofs of auxiliary results for chapter 4, as well as relevant compendiums of matrix perturbation theory and exponential concentration of measure inequalities, are contained in appendix C.

It is worth mentioning that in section 4.8, we digress from the statistical flavor of the rest of chapter 4 and tackle an information theoretic question. The modal decomposition structure of bivariate distributions intuitively suggests a natural way to encode information and combat noise in permutation channels. However, we demonstrate that this intuition is misleading, and present an elegant solution to the (ultimately different) problem of reliably communicating through permutation channels. In particular, the permutation channel model constitutes a discrete memoryless channel (DMC) followed by a random permutation block that reorders the output codeword of the DMC. This model naturally emerges in the context of communication networks, and coding theoretic aspects of the model have been widely studied (see e.g. [159–161]). In contrast to the bulk of this literature, we analyze the information theoretic aspects of the model by defining an appropriate notion of "permutation channel capacity." We consider two special cases of the permutation channel model: the binary symmetric channel (BSC) and the binary erasure channel. We establish the permutation channel capacity of the former, and prove bounds on the permutation channel capacity of the latter. Somewhat surprisingly, our results illustrate that permutation channel capacities are generally agnostic to the parameters that define the DMCs. Furthermore, our achievability proof yields a conceptually simple, computationally efficient, and capacity achieving coding scheme for the BSC permutation channel.

■ 1.1.3 Discrete Probability Theory

The third part of this dissertation, which is comprised of chapter 5, considers the broader setting of Bayesian networks rather than point-to-point channels or Markov chains, and analyzes the discrete probability problem of broadcasting on bounded indegree directed acyclic graphs (DAGs). Specifically, it studies the following generalization of the well-known problem of broadcasting on trees, cf. [83]. Consider an infinite DAG with a unique source vertex X, where every non-source vertex has $d \geq 2$ incoming edges. Let the collection of vertices at distance k from X be called the kth layer of the DAG, and have cardinality L_k. At time $k = 0$, the source vertex is given a uniform bit. At time $k \geq 1$, each vertex in the kth layer receives d bits from its parents in the $(k-1)$th layer. These bits are transmitted along edges that are independent BSCs with common crossover probability $\delta \in \left(0, \frac{1}{2}\right)$. Each vertex at the kth layer then combines its d input bits using a deterministic d-ary Boolean processing function that generates the value at the vertex.

The goal is to be able to reconstruct the original bit with probability of error better than $\frac{1}{2}$ from the values of all vertices at an arbitrarily deep layer k. Guided by the information theorist's basic tenet of understanding fundamental limits of models, we establish phase transition results for δ that determine when reconstruction is possible, and derive the optimal growth of L_k that permits reconstruction. We note that reconstruction of X is impossible if and only if the TV distance between the two conditional distributions of the vertices at the kth layer (given $X = 0$ and $X = 1$, respectively) vanishes as $k \to \infty$ (due to Le Cam's relation). Therefore, from this perspective, the broadcasting on DAGs problem is entirely a question about contraction of TV distance.

Besides its canonical broadcast interpretation, broadcasting on DAGs is a natural model of reliable computation and storage, cf. [85, 86, 115, 286]. Indeed, the model can be construed as a noisy circuit constructed to remember a bit, where the edges are wires that independently make errors, and the Boolean processing functions at the vertices are perfect logic gates. Furthermore, the broadcasting model on certain DAGs, such as trees or two-dimensional (2D) regular grids, can also be perceived as ferromagnetic Ising models, cf. [83, Section 2.2], or one-dimensional (1D) probabilistic cellular automata (or more generally, discrete-time statistical mechanical spin-flip systems on 1D lattices—see e.g. [107–109]) with boundary conditions that limit the number of sites at each time k to $L_k = k + 1$, respectively. Other special cases of the model represent information flow in biological networks (see e.g. [63, 212, 213, 238]).

In chapter 5, we demonstrate the existence of DAGs with bounded indegree and layers of size $L_k = \Omega(\log(k))$ that permit reconstruction provided that δ is sufficiently small. We show this via a probabilistic argument using random DAGs, where for each incoming edge to a vertex at layer k, its starting vertex is chosen uniformly from all vertices at layer $k - 1$ and independently of all other edges. In particular, for random DAGs with $d \geq 3$ and all majority processing functions, we identify the (degree and function dependent) critical threshold for δ below which reconstruction is possible, and above which reconstruction is impossible. For random DAGs with $d = 2$, where the choice of good processing functions is unclear, we show that applying alternating layers of AND and OR processing functions yields a similar phase transition phenomenon with a different critical threshold for δ. Moreover, we establish a partial converse for odd $d \geq 3$ illustrating that the identified thresholds are impossible to improve by selecting different processing functions if the decoder is restricted to using a single vertex's value. Finally, for any $\delta \in \left(0, \frac{1}{2}\right)$, we construct explicit deterministic DAGs using regular bipartite lossless expander graphs, with sufficiently large bounded degree and layers of size $L_k = \Theta(\log(k))$, such that reconstruction is possible. Specifically, the constituent expander graphs of such DAGs can be generated in either deterministic quasi-polynomial time or randomized polylogarithmic time in the number of layers. These results demonstrate a doubly-exponential advantage for storing a bit in bounded degree DAGs compared to trees (where $d = 1$ and L_k must be exponential for reconstruction to be possible [83]).

On the negative side, inspired by the literature surrounding the "positive rates conjecture" for 1D probabilistic cellular automata, cf. [109, Section 1], we conjecture

that it is impossible to propagate information in a 2D regular grid regardless of the noise level δ and of the choice of processing function (which is the same for every vertex). We take some first steps towards establishing this conjecture in chapter 5, and prove that reconstruction is impossible for any $\delta \in \left(0, \frac{1}{2}\right)$ provided all vertices use either AND or XOR for their processing functions. Lastly, we note that several proofs of the results in chapter 5 are contained in appendix D.

■ 1.1.4 General Remarks

Before delving into our study of information contraction, we make some pertinent remarks. Firstly, the expositions in chapters 2, 3, 4, and 5 are all fairly self-contained. So, each chapter possesses its own conclusion and future directions section that appears at the end of the chapter. In addition, we also conclude the broader discussion of this thesis in chapter 6. Secondly, each chapter also ends with bibliographical notes that indicate the publications or manuscripts upon which the chapter is based. Finally, we note that this thesis assumes knowledge of several rudimentary topics in mathematics. In particular, we refer readers to [17, 27, 128, 129, 266] for the relevant linear algebra and matrix theory, [239] for introductory real analysis, [199, 262] for introductory functional analysis, [34] for basic convex analysis, [41, 89] for the relevant measure theoretic probability theory, and [170] for the theory of Markov chains. This thesis also assumes familiarity with the basic concepts of information theory and statistics. So, we also refer readers to [53, 58, 230] for the relevant information theory, [81] for the pertinent snippets of network information theory, and [150, 290] for the necessary concepts of theoretical statistics and detection and estimation theory.

Chapter 2

Contraction Coefficients and Strong Data Processing Inequalities

STRONG data processing inequalities for KL divergence and mutual information [5,10,11,82,147], and more generally f-divergences [229,231,234], have been studied extensively in various contexts in information theory. They are obtained by tightening traditional data processing inequalities using distribution dependent constants known as "contraction coefficients." Contraction coefficients for f-divergences come in two flavors: those pertaining to source-channel pairs, and those pertaining only to channels. The broad goal of this chapter is to study various inequalities and equalities that elucidate the relationship between contraction coefficients of source-channel pairs. We will primarily establish general bounds on contraction coefficients for certain classes of f-divergences, as well as specific bounds on the contraction coefficient for KL divergence, in terms of the contraction coefficient for χ^2-divergence (or squared Hirschfeld-Gebelein-Rényi maximal correlation). On the other hand, we will consider contraction coefficients of channels in chapter 3, and among other things, prove an appropriate generalization of the well-known result that the contraction coefficient for KL divergence is equal to the contraction coefficient for any f-divergence with non-linear operator convex f [46].

■ 2.1 Chapter Outline

We briefly delineate the discussion in the remainder of this chapter. We will first provide an overview of the burgeoning literature on contraction coefficients in section 2.2. This section will compile formal definitions and key properties of both the aforementioned variants of contraction coefficients, and briefly outline their genesis in the study of ergodicity. Then, we will state and explain our main results, and discuss related literature in section 2.3. In section 2.4, we will present some useful bounds between f-divergences and χ^2-divergence, and use them to prove linear upper bounds on contraction coefficients of source-channel pairs for a certain class of f-divergences and KL divergence. Following this, we will prove the equivalence between certain contraction coefficients of Gaussian sources with additive Gaussian noise channels in section 2.5. Finally, we will conclude our discussion and propose future research directions in section 2.6.

■ 2.2 Overview of Contraction Coefficients

In this section, we will define contraction coefficients for f-divergences and present some well-known facts about them. We begin by introducing some preliminary definitions and notation pertaining to f-divergences in subsection 2.2.1, and then give a brief prelude on contraction coefficients and strong data processing inequalities in the ensuing subsections.

■ 2.2.1 f-Divergence

Consider a discrete sample space $\mathcal{X} \triangleq \{1, \ldots, |\mathcal{X}|\}$ with cardinality $2 \leq |\mathcal{X}| < +\infty$, where we let singletons in $\mathcal{X}$ be natural numbers without loss of generality. Let $\mathcal{P}_\mathcal{X} \subseteq (\mathbb{R}^{|\mathcal{X}|})^*$ denote the probability simplex in $(\mathbb{R}^{|\mathcal{X}|})^*$ of all probability mass functions (pmfs) on $\mathcal{X}$, where $(\mathbb{R}^{|\mathcal{X}|})^*$ is the dual vector space of $\mathbb{R}^{|\mathcal{X}|}$ consisting of all row vectors of length $|\mathcal{X}|$. We perceive $\mathcal{P}_\mathcal{X}$ as the set of all possible probability distributions of a random variable X with range $\mathcal{X}$, and construe each pmf $P_X \in \mathcal{P}_\mathcal{X}$ as a row vector $P_X = (P_X(1), \ldots, P_X(|\mathcal{X}|)) \in (\mathbb{R}^{|\mathcal{X}|})^*$. We also let $\mathcal{P}_\mathcal{X}^\circ$ denote the relative interior of $\mathcal{P}_\mathcal{X}$, which contains all strictly positive pmfs on $\mathcal{X}$. A popular notion of "distance" between pmfs in information theory is the (Csiszár) f-divergence, which was independently introduced by Csiszár in [54, 55] and by Ali and Silvey in [8].[1]

Definition 2.1 (f-Divergence [8, 54, 55]). *Given a convex function $f : (0, \infty) \to \mathbb{R}$ that satisfies $f(1) = 0$, we define the* f-divergence *of a pmf $P_X \in \mathcal{P}_\mathcal{X}$ from a pmf $R_X \in \mathcal{P}_\mathcal{X}$ as:*

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right) \tag{2.1}$$

$$= \mathbb{E}_{P_X}\left[f\left(\frac{R_X(X)}{P_X(X)}\right)\right] \tag{2.2}$$

where we assume that $f(0) = \lim_{t \to 0^+} f(t)$ (which is possibly infinity), $0f(0/0) = 0$, and for all $r > 0$, $0f(r/0) = \lim_{p \to 0^+} pf(r/p) = r \lim_{p \to 0^+} pf(1/p)$ (which could also be infinity), based on continuity and other considerations (see [174, Section 3] for details).

The f-divergences generalize several well-known divergence measures that are used in information theory, statistics, and probability theory. We present some examples below:

1. **Total variation (TV) distance:** When $f(t) = \frac{1}{2}|t - 1|$, the corresponding f-divergence is the TV distance:

$$\|R_X - P_X\|_{\mathsf{TV}} \triangleq \max_{A \subseteq \mathcal{X}} |R_X(A) - P_X(A)| \tag{2.3}$$

[1] Although f-divergences are usually credited to these references, it is worth mentioning that they were independently discovered by Morimoto in [207], Akaike in [7], and Ziv and Zakai in [295, 296].

$$= \frac{1}{2} \|R_X - P_X\|_1 \tag{2.4}$$

$$= 1 - \sum_{x \in \mathcal{X}} \min\{R_X(x), P_X(x)\} \tag{2.5}$$

$$= \min_{\substack{\mathbb{P}_{X,X'}: \\ \mathbb{P}_X = P_X, \mathbb{P}_{X'} = R_X}} \mathbb{P}(X \neq X') \tag{2.6}$$

$$= \max_{\substack{g: \mathcal{X} \to \mathbb{R}: \\ \max_{x \in \mathcal{X}} |g(x)| \leq \frac{1}{2}}} \mathbb{E}_{R_X}[g(X)] - \mathbb{E}_{P_X}[g(X)] \tag{2.7}$$

where $P_X(A) = \sum_{x \in A} P_X(x)$ for any $A \subseteq \mathcal{X}$, the maximum in (2.3) is achieved by the set $A^* = \{x \in \mathcal{X} : R_X(x) \geq P_X(x)\}$ [170, Remark 4.3],[2] (2.4) is the ℓ^1-norm characterization of TV distance [170, Proposition 4.2], (2.5) is the *affinity* characterization of TV distance [170, Equation (4.13)],[3] (2.6) is Dobrushin's *maximal coupling* representation of TV distance which maximizes $\mathbb{P}(X = X')$ over all couplings (i.e. joint pmfs) $\mathbb{P}_{X,X'}$ of the random variables $X, X' \in \mathcal{X}$ such that the marginal distributions of X and X' are $\mathbb{P}_X = P_X$ and $\mathbb{P}_{X'} = R_X$, respectively [170, Proposition 4.7], and (2.7) is the *Kantorovich-Rubinstein dual* characterization of (2.6) (where the latter can be construed as a *Wasserstein distance* of order 1 with respect to the Hamming metric) [170, Proposition 4.5].[4]

2. **Kullback-Leibler (KL) divergence [163, Section 2]:** When $f(t) = t\log(t)$, the corresponding f-divergence is the KL divergence (or *relative entropy*):

$$D(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} R_X(x) \log\left(\frac{R_X(x)}{P_X(x)}\right). \tag{2.8}$$

3. **χ^2-divergence [218]:**[5] When $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$, the corresponding f-divergence is the (Neyman) χ^2-divergence:

$$\chi^2(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} \frac{(R_X(x) - P_X(x))^2}{P_X(x)}. \tag{2.9}$$

4. **Hellinger divergence of order $\alpha \in (0,\infty)\backslash\{1\}$ [173, Definition 2.10]:** When $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$, the corresponding f-divergence is the Hellinger divergence (or *Tsallis divergence*) of order α:

$$\mathcal{H}_\alpha(R_X||P_X) \triangleq \frac{1}{\alpha - 1}\left(\sum_{x \in \mathcal{X}} R_X(x)^\alpha P_X(x)^{1-\alpha} - 1\right) \tag{2.10}$$

[2] We can perceive A^* as the maximum likelihood decoding region for a binary hypothesis test between R_X and P_X.

[3] Note that $\sum_{x \in \mathcal{X}} \min\{R_X(x), P_X(x)\}$ is known as the affinity between R_X and P_X, cf. [293].

[4] We refer readers to [284] for a comprehensive treatment of the *Monge-Kantorovich problem* and transportation theory.

[5] See e.g. [220] for other variants of χ^2-divergence due to Pearson and Vajda and their relation to f-divergences.

where $\frac{1}{2}\mathcal{H}_{1/2}(R_X||P_X)$ is known as the *squared Hellinger distance*, $\mathcal{H}_2(R_X||P_X) = \chi^2(R_X||P_X)$ is the χ^2-divergence (above), and $\alpha = 1$ corresponds to KL divergence (above), $\mathcal{H}_1(R_X||P_X) = D(R_X||P_X)$, by analytic extension, cf. [245, Section II].

5. **Vincze-Le Cam divergence of order $\lambda \in (0,1)$ [114, 167, 285]:** When $f(t) = (\lambda(1-\lambda)(t-1)^2)/(\lambda t + (1-\lambda))$, the corresponding f-divergence is the Vincze-Le Cam divergence of order λ:

$$\mathsf{LC}_\lambda(R_X||P_X) \triangleq \lambda(1-\lambda)\sum_{x\in\mathcal{X}} \frac{(R_X(x)-P_X(x))^2}{\lambda R_X(x) + (1-\lambda)P_X(x)} \tag{2.11}$$

$$= \left(\frac{\lambda}{1-\lambda}\right)\chi^2(R_X||\lambda R_X + (1-\lambda)P_X) \tag{2.12}$$

where the special case of $\lambda = \frac{1}{2}$ is called the *Vincze-Le Cam distance* or *triangular discrimination*.

Although f-divergences are not valid metrics in general,[6] they satisfy several useful properties. To present some of these properties, we let $\mathcal{Y} \triangleq \{1,\ldots,|\mathcal{Y}|\}$ denote another discrete alphabet with $2 \le |\mathcal{Y}| < +\infty$, and corresponding probability simplex $\mathcal{P}_\mathcal{Y}$ of possible pmfs of a random variable Y with range $\mathcal{Y}$. Furthermore, we let $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denote the set of $|\mathcal{X}| \times |\mathcal{Y}|$ row stochastic matrices in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$. Throughout our discussion in this chapter, the discrete channel of conditional pmfs $\{P_{Y|X=x} \in \mathcal{P}_\mathcal{Y} : x \in \mathcal{X}\}$ will correspond to a transition probability matrix $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ (where the xth row of W is $P_{Y|X=x}$). We interpret $W : \mathcal{P}_\mathcal{X} \to \mathcal{P}_\mathcal{Y}$ as a map that takes input pmfs $P_X \in \mathcal{P}_\mathcal{X}$ to output pmfs $P_Y = P_X W \in \mathcal{P}_\mathcal{Y}$. Some well-known properties of f-divergences are presented next, cf. [54, 55]:

1. **Non-negativity and reflexivity:** For every $R_X, P_X \in \mathcal{P}_\mathcal{X}$, Jensen's inequality yields:

$$D_f(R_X||P_X) \ge 0 \tag{2.13}$$

where equality holds if $R_X = P_X$. Furthermore, if f is strictly convex at unity, then equality holds if and only if $R_X = P_X$.[7]

2. **Affine invariance:** Consider any affine function $\alpha(t) = a(t-1)$ with $a \in \mathbb{R}$. Then, for every $R_X, P_X \in \mathcal{P}_\mathcal{X}$:

$$D_\alpha(R_X||P_X) = 0\,. \tag{2.14}$$

Hence, f and $f+\alpha$ define the same f-divergence,[8] i.e. for every $R_X, P_X \in \mathcal{P}_\mathcal{X}$:

$$D_{f+\alpha}(R_X||P_X) = D_f(R_X||P_X)\,. \tag{2.15}$$

[6] We often distinguish f-divergences that are metrics by dubbing them as "distances" (e.g. TV distance, Hellinger distance), while the term "divergence" is reserved for f-divergences that are not metrics (e.g. KL divergence, χ^2-divergence).

[7] Strict convexity of $f : (0,\infty) \to \mathbb{R}$ at unity implies that for every $x, y \in (0,\infty)$ and $\lambda \in (0,1)$ such that $\lambda x + (1-\lambda)y = 1$, $\lambda f(x) + (1-\lambda)f(y) > f(1)$. The aforementioned examples of f-divergences have this property. We also note that (2.13) is known as *Gibbs' inequality* in the KL divergence case.

[8] Note that $f + \alpha : (0,\infty) \to \mathbb{R}$ is the function $(f+\alpha)(t) = f(t) + a(t-1)$.

3. **Csiszár duality:** Let the *Csiszár conjugate* function of f be $f^* : (0,\infty) \to \mathbb{R}$, $f^*(t) = tf(1/t)$, which is also convex and satisfies $f^{**} = f$.[9] Then, for every $R_X, P_X \in \mathcal{P}_\mathcal{X}$:
$$D_{f^*}(P_X||R_X) = D_f(R_X||P_X)\,. \tag{2.16}$$

4. **Joint convexity:** The map $(R_X, P_X) \mapsto D_f(R_X||P_X)$ is convex in the pair of input pmfs.

5. **Data processing inequality (DPI):** For every $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, and every $R_X, P_X \in \mathcal{P}_\mathcal{X}$, we have (by the convexity of perspective functions):
$$D_f(R_X W||P_X W) \le D_f(R_X||P_X) \tag{2.17}$$
where equality holds if Y is a sufficient statistic of X for performing inference about the pair (R_X, P_X), cf. [174, Definition 5]. Furthermore, if f is strictly convex and $D_f(R_X||P_X) < \infty$, then equality holds if and only if Y is a sufficient statistic of X for performing inference about the pair (R_X, P_X) (see e.g. [174, Theorem 14], [230, Section 3.1]).

While [55] and [56, Section 2] contain the original presentation of these properties, we also refer readers to [230, Section 6] for a more didactic presentation. Note that due to property 2, we only consider f-divergences with non-linear f.

We next define a notion of "information" between random variables corresponding to any f-divergence that also exhibits a DPI. For random variables X and Y with joint pmf $P_{X,Y}$ (consisting of (P_X, W)), the *mutual f-information* between X and Y is defined as [50]:[10]
$$I_f(X;Y) \triangleq D_f(P_{X,Y}||P_X P_Y) \tag{2.18}$$
$$= \sum_{x\in\mathcal{X}} P_X(x) D_f(P_{Y|X=x}||P_Y) \tag{2.19}$$
where $P_X P_Y$ denotes the product distribution specified by the marginal pmfs P_X and P_Y (also see [234, Equation (V.8)], [75, Equation (11)]). When $f(t) = t\log(t)$, mutual f-information $I_f(X;Y)$ corresponds to standard mutual information $I(X;Y)$ (as defined by Fano, cf. [230, Section 2.3]). Moreover, mutual f-information possesses certain natural properties of information measures. For example, if X and Y are independent, then $I_f(X;Y) = 0$, and the converse holds when f is strictly convex at unity.

Now suppose U is another random variable with discrete alphabet $\mathcal{U} \triangleq \{1,\dots,|\mathcal{U}|\}$ such that $2 \le |\mathcal{U}| < +\infty$. If (U,X,Y) are jointly distributed and form a Markov chain $U \to X \to Y$, then they satisfy the DPI [56]:[11]
$$I_f(U;Y) \le I_f(U;X) \tag{2.20}$$

[9] Note also that f is strictly convex at unity if and only if f^* is strictly convex at unity.

[10] In (2.19), we use the convention that $P_X(x)D_f(P_{Y|X=x}||P_Y) = 0$ if $P_X(x) = 0$.

[11] Although Csiszár studies a different notion known as *f-informativity* in [56], (2.20) can be distilled from the proof of [56, Proposition 2.1].

where equality holds if Y is a sufficient statistic of X for performing inference about U (i.e. $U \to Y \to X$ also forms a Markov chain). Moreover, if f is strictly convex and $I_f(U;X) < \infty$, then equality holds if and only if Y is a sufficient statistic of X for performing inference about U. Needless to say, the DPIs (2.17) and (2.20) are generalizations of the better known DPIs for KL divergence and mutual information that can be found in standard texts on information theory, e.g. [53] (also see [163, Theorem 4.1]). Finally, note that although we cite [54,55] and [56] for the DPIs (2.17) and (2.20) respectively, (2.17) was also established in [207], and both DPIs were independently proved in [295, 296].[12]

We end this subsection with a brief exposition of the "local quadratic behavior" of f-divergences. Local approximations of f-divergences are geometrically appealing because they transform neighborhoods of stochastic manifolds, with certain f-divergences as the distance measures, into inner product spaces with the Fisher-Rao information metric, cf. [9,33,139]. Consider any reference pmf $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ (which forms the "center of the local neighborhood" of pmfs that we will be concerned with), and any other pmf $R_X \in \mathcal{P}_{\mathcal{X}}$. Let us define the *spherical perturbation* vector of R_X from P_X as:

$$K_X \triangleq (R_X - P_X)\,\mathsf{diag}\Big(\sqrt{P_X}\Big)^{-1} \tag{2.21}$$

where $\sqrt{\cdot}$ denotes the entry-wise square root when applied to a vector. Using K_X, we can construct a trajectory of spherically perturbed pmfs:

$$R_X^{(\epsilon)} = P_X + \epsilon\, K_X\, \mathsf{diag}\Big(\sqrt{P_X}\Big) \tag{2.22}$$

$$= (1-\epsilon)P_X + \epsilon R_X \tag{2.23}$$

which is parametrized by $\epsilon \in (0,1)$, and corresponds to the convex combinations of R_X and P_X. Note that K_X provides the direction of the trajectory (2.22), and ϵ controls how close $R_X^{(\epsilon)}$ and P_X are. Furthermore, (2.22) clarifies why K_X is called a "spherical perturbation" vector; K_X is proportional to the first order perturbation term as $\epsilon \to 0$ of $\sqrt{R_X^{(\epsilon)}}$ from $\sqrt{P_X}$, which are embeddings of the pmfs $R_X^{(\epsilon)}$ and P_X as vectors on the unit sphere in $\big(\mathbb{R}^{|\mathcal{X}|}\big)^*$.

Now suppose the function $f : (0,\infty) \to \mathbb{R}$ that defines our f-divergence is twice differentiable at unity with $f''(1) > 0$.[13] Then, Taylor's theorem can be used to show that this f-divergence is locally proportional to χ^2-divergence, cf. [59, Section 4] (or [53], [230, Section 4.2], Proposition 4.3 in chapter 4 for the KL divergence case):[14]

$$D_f(R_X^{(\epsilon)}||P_X) = \frac{f''(1)}{2}\epsilon^2\chi^2(R_X||P_X) + o(\epsilon^2) \tag{2.24}$$

[12] In particular, Ziv and Zakai studied *generalized information functionals* in [295], and a specialization of [295, Theorem 5.1] yields $D_f(P_U P_Y||P_{U,Y}) \le D_f(P_U P_X||P_{U,X})$ for any Markov chain $U \to X \to Y$. By the Csiszár duality property of f-divergences, this implies (2.20).

[13] Note that since $f : (0,\infty) \to \mathbb{R}$ is convex, it has a second derivative almost everywhere by *Alexandrov's theorem*. (This is closely related to *Lebesgue's theorem* for differentiability of monotone functions, and *Rademacher's theorem* for differentiability of Lipschitz continuous functions.)

[14] In particular, we apply a lesser known version of Taylor's theorem to $f(t)$ around $t = 1$ where the

$$= \frac{f''(1)}{2}\epsilon^2 \left\|K_X\right\|_2^2 + o(\epsilon^2)\,. \tag{2.25}$$

The local approximation in (2.25) is somewhat more flexible than the version in (2.24). Indeed, we can construct a trajectory (2.22) using a spherical perturbation vector $K_X \in \left(\mathbb{R}^{|\mathcal{X}|}\right)^*$ that satisfies the orthogonality constraint $\sqrt{P_X}K_X^T = 0$, but is not of the form (2.21). For sufficiently small $\epsilon \neq 0$ (depending on P_X and K_X), the vectors $R_X^{(\epsilon)}$ defined by (2.22) are in fact valid pmfs in $\mathcal{P}_\mathcal{X}$.[15] So, the approximation in (2.25) continues to hold because it is concerned with the regime where $\epsilon \to 0$.

It is also straightforward to verify that f-divergences with $f''(1) > 0$ are locally symmetric:

$$D_f(R_X^{(\epsilon)}||P_X) = D_f(P_X||R_X^{(\epsilon)}) + o(\epsilon^2)\,. \tag{2.26}$$

Hence, they resemble the standard Euclidean metric within a "neighborhood" of pmfs around a reference pmf in $\mathcal{P}_\mathcal{X}^\circ$. Note that the advantage of using spherical perturbations $\{K_X \in \left(\mathbb{R}^{|\mathcal{X}|}\right)^* : \sqrt{P_X}K_X^T = 0\}$ over additive perturbations (e.g. $R_X - P_X$) is that they form an inner product space equipped with the standard Euclidean inner product. This allows us to recast (2.24) using the ℓ^2-norm of K_X instead of a weighted ℓ^2-norm of the additive perturbation $K_X\mathsf{diag}(\sqrt{P_X})$. Consequently, we benefit from more polished notation and simpler algebra later on—see our proof of Theorem 2.1. Finally, we remark that perturbation ideas like (2.22) have also been exploited in various other contexts in information theory, and we refer readers to [3, 33, 104, 139] for a few examples.

■ 2.2.2 Contraction Coefficients of Source-Channel Pairs

The DPIs, (2.17) and (2.20), can be maximally tightened into so called *strong data processing inequalities* (SDPIs) by inserting in pertinent constants known as contraction coefficients. There are two variants of contraction coefficients: the first depends on a source-channel pair, and the second depends solely on a channel. We introduce the former kind of coefficient in this subsection, and defer a discussion of the latter kind to ensuing subsections.

Definition 2.2 (Contraction Coefficient of Source-Channel Pair). *For any input pmf $P_X \in \mathcal{P}_\mathcal{X}$ and any discrete channel $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ corresponding to a conditional distribution $P_{Y|X}$, the* contraction coefficient *for a particular f-divergence is:*

$$\eta_f(P_X, P_{Y|X}) \triangleq \sup_{\substack{R_X \in \mathcal{P}_\mathcal{X}: \\ 0<D_f(R_X||P_X)<+\infty}} \frac{D_f(R_X W||P_X W)}{D_f(R_X||P_X)}$$

(weak differentiability) assumption that $f(t)$ is twice differentiable at $t = 1$ yields the Peano form of the remainder [14, 120]. Standard versions of Taylor's theorem use stronger differentiability assumptions, cf. [120, 239], and admit more elegant representations of the remainder term such as the Lagrange and Cauchy forms.

[15] For larger values of ϵ (in magnitude), although $R_X^{(\epsilon)}$ always sums to 1 since $\sqrt{P_X}K_X^T = 0$, it may not be entry-wise non-negative.

where the supremum is taken over all pmfs $R_X \in \mathcal{P}_\mathcal{X}$ such that $0 < D_f(R_X||P_X) < +\infty$. Furthermore, if X or Y is a constant a.s., we define $\eta_f(P_X, P_{Y|X}) = 0$.

Using Definition 2.2, we may write the ensuing SDPI from the DPI for f-divergences in (2.17):

$$D_f(R_X W||P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X||P_X) \tag{2.27}$$

which holds for every $R_X \in \mathcal{P}_\mathcal{X}$, with fixed $P_X \in \mathcal{P}_\mathcal{X}$ and $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. The next proposition illustrates that the DPI for mutual f-information can be improved in a similar fashion.

Proposition 2.1 (Mutual f-Information Contraction Coefficient [234, Theorem V.2]). *For any input pmf $P_X \in \mathcal{P}_\mathcal{X}$, any discrete channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, and any convex function $f : (0, \infty) \to \mathbb{R}$ that is differentiable, has uniformly bounded derivative in some neighborhood of unity, and satisfies $f(1) = 0$, we have:*

$$\eta_f(P_X, P_{Y|X}) = \sup_{\substack{P_{U|X}: U \to X \to Y \\ 0 < I_f(U;X) < +\infty}} \frac{I_f(U;Y)}{I_f(U;X)}$$

where the supremum is taken over all conditional distributions $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$ and finite alphabets $\mathcal{U}$ of U such that $U \to X \to Y$ form a Markov chain.[16]

Proposition 2.1 is proved in [234, Theorem V.2]. The special case of this result for KL divergence was proved in [11] (which tackled the finite alphabet case) and [229] (which derived the general alphabet case). Intuitively, the variational problem in Proposition 2.1 determines the probability model that makes Y as close to a sufficient statistic of X for U as possible (see the comment after (2.20)). Furthermore, the result illustrates that under regularity conditions, the contraction coefficient for any f-divergence gracefully unifies the DPIs for the f-divergence and the corresponding mutual f-information as the tightest factor that can be inserted into either one of them. Indeed, when the random variables $U \to X \to Y$ form a Markov chain, we can write the SDPI version of (2.20):

$$I_f(U;Y) \leq \eta_f(P_X, P_{Y|X}) I_f(U;X) \tag{2.28}$$

which holds for every conditional distribution $P_{U|X}$, with fixed $P_X \in \mathcal{P}_\mathcal{X}$ and $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Note that even if the conditions of Proposition 2.1 do not hold, (2.28) is still true (but $\eta_f(P_X, P_{Y|X})$ may not be the tightest possible constant that can be inserted into (2.20)).

There are two contraction coefficients that will be particularly important to our study. The first is the contraction coefficient for KL divergence:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{P}_\mathcal{X}: \\ 0 < D(R_X||P_X) < +\infty}} \frac{D(R_X W||P_X W)}{D(R_X||P_X)}. \tag{2.29}$$

[16] It suffices to let $|\mathcal{U}| = 2$ in the extremization [234, Theorem V.2].

This quantity is related to the fundamental notion of *hypercontractivity* in statistics [5].[17] In fact, the authors of [5] and [10] illustrate how $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ can be defined as the chordal slope of the lower boundary of the hypercontractivity ribbon at infinity in the discrete and finite setting.

The contraction coefficient for KL divergence elucidates a striking dichotomy between the extremizations in Definition 2.2 and Proposition 2.1. To delineate this contrast, we first specialize Proposition 2.1 for KL divergence [11, 229]:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \sup_{\substack{P_U, P_{X|U}: U \to X \to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} \tag{2.30}$$

where the optimization is (equivalently) over all $P_U \in \mathcal{P}_{\mathcal{U}}$ with $\mathcal{U} = \{0, 1\}$ (without loss of generality, cf. [229, Appendix B]) and all $P_{X|U} \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$ such that marginalizing yields P_X. We next recall an example from [11] where $\mathcal{U} = \{0, 1\}$, $X \sim \mathsf{Bernoulli}(\frac{1}{2})$, and $P_{Y|X}$ is an asymmetric erasure channel. In this numerical example, the supremum in (2.30) is achieved by the sequences of pmfs:

- $\{P^{(k)}_{X|U=0} \in \mathcal{P}_{\mathcal{X}} : k \in \mathbb{N}\}$,
- $\{P^{(k)}_{X|U=1} \in \mathcal{P}_{\mathcal{X}} : k \in \mathbb{N}\}$,
- $\{P^{(k)}_{U} \in \mathcal{P}_{\mathcal{U}} : k \in \mathbb{N}\}$,

satisfying the following conditions:

$$\lim_{k\to\infty} P^{(k)}_U(1) = 0\,, \tag{2.31}$$

$$\lim_{k\to\infty} D(P^{(k)}_{X|U=0}||P_X) = 0\,, \tag{2.32}$$

$$\liminf_{k\to\infty} D(P^{(k)}_{X|U=1}||P_X) > 0\,, \tag{2.33}$$

$$\limsup_{k\to\infty} \frac{D(P^{(k)}_{Y|U=0}||P_Y)}{D(P^{(k)}_{X|U=0}||P_X)} < \eta_{\mathsf{KL}}(P_X, P_{Y|X}), \tag{2.34}$$

$$\lim_{k\to\infty} \frac{D(P^{(k)}_{Y|U=1}||P_Y)}{D(P^{(k)}_{X|U=1}||P_X)} = \eta_{\mathsf{KL}}(P_X, P_{Y|X}). \tag{2.35}$$

This example conveys that in general, although (2.30) is maximized when $I(U;X) \to 0$ [82], (2.29) is often achieved by a sequence of pmfs $\{R^{(k)}_X \in \mathcal{P}_{\mathcal{X}} \backslash \{P_X\} : k \in \mathbb{N}\}$ that

[17] Hypercontractivity refers to the phenomenon that some conditional expectation operators are contractive even when their input functional space has a (probabilistic) $\mathcal{L}^q$-norm while their output functional space has a (probabilistic) $\mathcal{L}^p$-norm with $1 \leq q < p$ (see e.g. [10]). This notion has found applications in information theory because hypercontractive quantities are often imparted with tensorization properties which permit single letterization.

does not tend to P_X (due to the non-concave nature of this extremal problem). At first glance, this is counter-intuitive because the DPI (2.17) is tight when $R_X = P_X$. However, Theorem 2.1 (presented in subsection 2.3.1) will portray that maximizing the ratio of KL divergences with the constraint that $D(R_X||P_X) \to 0$ actually achieves $\eta_{\chi^2}(P_X, P_{Y|X})$, which is often strictly less than $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ [11]. Therefore, there is a stark contrast between the behaviors of the optimization problems in (2.29) and (2.30).

The second important contraction coefficient is the contraction coefficient for χ^2-divergence:

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}}: \\ 0<\chi^2(R_X||P_X)<+\infty}} \frac{\chi^2(R_X W||P_X W)}{\chi^2(R_X||P_X)} \tag{2.36}$$

which is closely related to a generalization of the Pearson correlation coefficient between X and Y known as the *Hirschfeld-Gebelein-Rényi maximal correlation*, or simply maximal correlation [96, 125, 236, 242]. We next define maximal correlation, which was proven to be a measure of statistical dependence satisfying seven natural axioms (some of which will be given in Proposition 2.3 later) that such measures should exhibit [236].

Definition 2.3 (Maximal Correlation [96, 125, 236, 242]). *For two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, the maximal correlation between X and Y is given by:*

$$\rho_{\mathsf{max}}(X;Y) \triangleq \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\, g:\mathcal{Y}\to\mathbb{R}: \\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0 \\ \mathbb{E}\left[f(X)^2\right]=\mathbb{E}\left[g(Y)^2\right]=1}} \mathbb{E}\left[f(X)g(Y)\right]$$

where the supremum is taken over all Borel measurable functions f and g satisfying the constraints. Furthermore, if X or Y is a constant a.s., there exist no functions f and g that satisfy the constraints, and we define $\rho_{\mathsf{max}}(X;Y) = 0$.

It can be shown that the contraction coefficient for χ^2-divergence is precisely the squared maximal correlation [242]:

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\mathsf{max}}(X;Y)^2 \,. \tag{2.37}$$

Furthermore, the next proposition portrays that maximal correlation can be represented as a singular value; this was originally observed in [125, 236] in slightly different forms (also see [11, 75, 289] and [180, Theorem 3.2.4]).

Proposition 2.2 (Singular Value Characterization of Maximal Correlation [125, 236]). *Given the random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with joint pmf $P_{X,Y}$ (consisting of (P_X, W)), we may define a* divergence transition matrix *(DTM):*[18]

$$B \triangleq \mathsf{diag}\left(\sqrt{P_X}\right) W \mathsf{diag}\left(\sqrt{P_Y}\right)^{\dagger} . \tag{2.38}$$

[18] Note that for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ with $P_X(x) > 0$ and $P_Y(y) > 0$, the entry $[B]_{x,y} = P_{X,Y}(x,y)/\sqrt{P_X(x)P_Y(y)}$, and $[B]_{x,y} = 0$ otherwise.

Then, the maximal correlation $\rho_{\max}(X;Y)$ is the second largest singular value of B.

Proof. See appendix A.1. ■

From Proposition 2.2 and (2.37), we see that the contraction coefficient for χ^2-divergence is in fact the squared second largest singular value of the DTM B. We can write this using the Courant-Fischer-Weyl variational characterization of eigenvalues or singular values (cf. Theorem C.1 in appendix C.1 or [129, Theorems 4.2.6 and 7.3.8]) as:

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \max_{\substack{x\in\mathbb{R}^{|\mathcal{X}|}\setminus\{\mathbf{0}\}:\\ \sqrt{P_X}x=0}} \frac{\left\|B^T x\right\|_2^2}{\|x\|_2^2} \tag{2.39}$$

where $\sqrt{P_X}^T$ is the right singular vector of B^T corresponding to its maximum singular value of unity.

Singular value decompositions (SVDs) of DTMs and their relation to χ^2-divergence have been well-studied in statistics. For example, the field of *correspondence analysis*, which was initiated by Hirschfeld in 1935 [125], deals with understanding the dependence between categorical random variables. In particular, *simple* correspondence analysis views a bivariate pmf $P_{X,Y}$ as a contingency table, and decomposes the dependence between X and Y into so called *principal inertia components* using the SVD of B, cf. [111], [110, Section 2], and the references therein. In [125], Hirschfeld used this observation to produce a modal decomposition of mutual χ^2-information (or Pearson's mean square contingency $\chi^2(P_{X,Y}||P_XP_Y)$). Although correspondence analysis was merely used as a data visualization technique in the past, it has become part of the broader toolkit of geometric data analysis now. Recently, the authors of [75] have studied principal inertia components (which are eigenvalues of the Gramian matrix B^TB) in the context of information and estimation theory. They generalize the first principal inertia component (i.e. squared maximal correlation) into a quantity known as k-correlation for $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$ (which is the Ky Fan $(k+1)$-norm of B^TB minus 1), prove some properties of k-correlation such as convexity and DPI [75, Section II], and demonstrate several applications.

While correspondence analysis concerns categorical random variables, the analysis and identification of so called *Lancaster distributions* is a related line of inquiry due to Lancaster that studies the dependence between general (non-categorical) random variables [165,166]. In particular, given a joint distribution $\mathbb{P}_{X,Y}$ over a product measurable space $\mathcal{X}\times\mathcal{Y}$ such that $\chi^2(\mathbb{P}_{X,Y}||\mathbb{P}_X\mathbb{P}_Y) < \infty$,[19] Lancaster proved in [165] that there exist orthonormal bases, $\{f_j \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : 0 \le j < |\mathcal{X}|\}$ and $\{g_k \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) : 0 \le k < |\mathcal{Y}|\}$, and some sequence $\{\sigma_k \ge 0 : 0 \le k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ of non-negative correlations, such

[19] Note that $\mathbb{P}_X$ and $\mathbb{P}_Y$ are marginal distributions of $\mathbb{P}_{X,Y}$, and $\mathbb{P}_X\mathbb{P}_Y$ denotes their product distribution. Furthermore, the condition $\chi^2(\mathbb{P}_{X,Y}||\mathbb{P}_X\mathbb{P}_Y) < \infty$, which can be perceived as a *Hilbert-Schmidt condition* that ensures compactness of the conditional expectation operators associated with $\mathbb{P}_{X,Y}$, cf. [191, Equation (43)], implies that $\mathbb{P}_{X,Y}$ is absolutely continuous with respect to $\mathbb{P}_X\mathbb{P}_Y$.

that $\mathbb{P}_{X,Y}$ is a Lancaster distribution exhibiting the decomposition:

$$\frac{d\,\mathbb{P}_{X,Y}}{d\,\mathbb{P}_X\mathbb{P}_Y}(x,y) = \sum_{k=0}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}-1} \sigma_k f_k(x) g_k(y) \tag{2.40}$$

where $d\,\mathbb{P}_{X,Y}/d\,\mathbb{P}_X\mathbb{P}_Y$ is the Radon-Nikodym derivative of $\mathbb{P}_{X,Y}$ with respect to $\mathbb{P}_X\mathbb{P}_Y$. When $\mathcal{X}$ and $\mathcal{Y}$ are finite, the decomposition in (2.40) precisely captures the SVD structure of B corresponding to $P_{X,Y}$. It is worth mentioning that explicit expansions of the form (2.40) in terms of orthogonal polynomials have been derived for various bivariate distributions. We refer readers to [68, 78, 112, 191] and the references therein for further details on such classical work. More contemporary results on Lancaster distributions are presented in [157, 158] and the references therein. As explained in [158], one direction of research is to find the extremal sequences of non-negative correlations corresponding to the extremal points of the compact, convex set of Lancaster distributions associated with certain marginal distributions and their orthogonal polynomial sequences. We refer readers to [191, Section II-D] for further references on this general area.

Yet another direction of research has focused on the computational aspects of decomposing DTMs. A well-known method of computing SVDs of DTMs is the alternating conditional expectations algorithm—see [35] for the original algorithm in the context of non-linear regression, and [182] for a variant of the algorithm in the context of feature extraction. At its heart, the alternating conditional expectations algorithm employs a power iteration method to estimate singular vectors of the DTM. It turns out that such singular vectors corresponding to larger singular values can be identified as "more informative" score functions. This insight has been exploited to perform inference on hidden Markov models in an image processing setting in [132], and has been framed as a means of performing *universal* feature extraction in [135].

Having introduced the pertinent contraction coefficients, we now present several properties of contraction coefficients for f-divergences; many of these properties are well-known or straightforward to prove, but a few have not appeared in the literature to our knowledge. (Furthermore, many of these properties can be extended to hold for general random variables X and Y.)

Proposition 2.3 (Properties of Contraction Coefficients of Source-Channel Pairs). *The contraction coefficient for an f-divergence satisfies the following properties:*

1. (Normalization) *For any joint pmf $P_{X,Y}$, we have that $0 \leq \eta_f(P_X, P_{Y|X}) \leq 1$.*

2. (Independence) *Given random variables X and Y with joint pmf $P_{X,Y}$, if X and Y are independent, then $\eta_f(P_X, P_{Y|X}) = 0$. Conversely, if f is strictly convex at unity and $\eta_f(P_X, P_{Y|X}) = 0$, then X and Y are independent.*

3. (Decomposability) *Suppose we have a joint pmf $P_{X,Y}$ such that the marginal pmfs satisfy $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. We say that $P_{X,Y}$ is* decomposable *when there exist functions $h : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ such that $h(X) = g(Y)$ a.s. and*

$\mathbb{VAR}(h(X)) > 0$,[20] *or equivalently, when the undirected bipartite graph with disjoint vertex sets* $\mathcal{X}$ *and* $\mathcal{Y}$ *and edge set* $\{(x,y) \in \mathcal{X} \times \mathcal{Y} : P_{Y|X}(y|x) > 0\}$ *has two or more connected components, cf. [5, Section 1].*[21] *If* f *is strictly convex, twice differentiable at unity with* $f''(1) > 0$*, and* $f(0) < \infty$*, then* $P_{X,Y}$ *is decomposable if and only if* $\eta_f(P_X, P_{Y|X}) = 1$*.*

4. (Convexity [234, Proposition III.3]) *For any fixed* $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$*, the function* $\mathcal{P}_{\mathcal{Y}|\mathcal{X}} \ni P_{Y|X} \mapsto \eta_f(P_X, P_{Y|X})$ *is convex in the channel* $P_{Y|X}$*.*

5. (Tensorization [234, Theorem III.9]) *If* f *induces a sub-additive and homogeneous* f*-entropy,*[22] *and* $\{P_{X_i,Y_i} : P_{X_i} \in \mathcal{P}_{\mathcal{X}_i}^{\circ}$ *and* $P_{Y_i} \in \mathcal{P}_{\mathcal{Y}_i}^{\circ}$ *for* $i \in \{1,\ldots,n\}\}$ *are independent joint pmfs, then we have:*

$$\eta_f(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \max_{1 \le i \le n} \eta_f(P_{X_i}, P_{Y_i|X_i})$$

where $X_1^n = (X_1,\ldots,X_n)$ *and* $Y_1^n = (Y_1,\ldots,Y_n)$*.*

6. (Sub-multiplicativity) *If* $U \to X \to Y$ *are discrete random variables with finite ranges that form a Markov chain, then we have:*

$$\eta_f(P_U, P_{Y|U}) \le \eta_f(P_U, P_{X|U}) \eta_f(P_X, P_{Y|X})\,.$$

Furthermore, for any fixed joint pmf $P_{X,Y}$ *such that* X *is not a constant a.s., we have:*

$$\eta_f(P_X, P_{Y|X}) = \sup_{\substack{P_{U|X}: U \to X \to Y \\ \eta_f(P_U, P_{X|U}) > 0}} \frac{\eta_f(P_U, P_{Y|U})}{\eta_f(P_U, P_{X|U})}$$

where the supremum is over all arbitrary finite ranges $\mathcal{U}$ *of* U*, and over all conditional distributions* $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$ *such that* $U \to X \to Y$ *form a Markov chain.*

7. (Maximal Correlation Lower Bound [234, Theorem III.3], [231, Theorem 2]) *Suppose we have a joint pmf* $P_{X,Y}$ *such that the marginal pmfs satisfy* $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ *and* $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$*. If* f *is twice differentiable at unity with* $f''(1) > 0$*, then we have:*

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\mathsf{max}}(X;Y)^2 \le \eta_f(P_X, P_{Y|X})\,.$$

Proof. See appendix A.2 for certain proofs, as well as relevant references for specializations of the results. ■

[20] For general alphabets $\mathcal{X}$ and $\mathcal{Y}$, the functions h and g must be Borel measurable.

[21] In this thesis, for any two vertices u, v of a graph, we let (u,v) denote an undirected edge between u and v if the graph is undirected, and a directed edge from u to v if the graph is directed.

[22] For a convex function $f : (0,\infty) \to \mathbb{R}$, the f*-entropy* of a non-negative random variable Z is defined as $\mathsf{Ent}_f(Z) \triangleq \mathbb{E}[f(Z)] - f(\mathbb{E}[Z])$, where it is assumed that $\mathbb{E}[f(Z)] < \infty$ (see [234, Section II] and the references therein).

We now make some relevant remarks. Firstly, to our knowledge, part 3 has not appeared in the literature before to this level of generality; only the η_{χ^2} and η_{KL} cases were known, cf. [5, 289]. Furthermore, the equivalence between the definition of decomposability and its combinatorial characterization holds because h and g exist if and only if the row stochastic transition probability matrix $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ corresponding to $P_{Y|X}$ has block diagonal structure after appropriately permuting its rows and columns (where the blocks are determined by the preimage sets of h and g), and these blocks correspond to connected components in the associated bipartite graph. We also note that when $\eta_f(P_X, P_{Y|X}) = 1$, it can be verified that the *zero error capacity* of the channel $P_{Y|X}$ is strictly positive [250].

Secondly, since parts 1, 2, and 3 of Proposition 2.3 illustrate that contraction coefficients are normalized measures of statistical dependence between random variables, we can perceive the sub-multiplicativity property in part 6 as a *meta-SDPI* for contraction coefficients in analogy with (2.28). In fact, part 6 also portrays that the contraction coefficient of the meta-SDPI for η_f is given by η_f itself. This latter aspect of part 6, while quite simple, has also not explicitly appeared in the literature to this level of generality to our knowledge; only the η_{χ^2} case is presented in [16, Lemma 6].

Thirdly, the version of the DPI for η_{KL} presented in [5] (also see [10, Section II-A]) holds for general η_f. Indeed, if $U \to X \to Y \to V$ are discrete random variables with finite ranges that form a Markov chain, then a straightforward consequence of parts 1 and 6 of Proposition 2.3 is the following monotonicity property:

$$\eta_f(P_U, P_{V|U}) \leq \eta_f(P_X, P_{Y|X}). \tag{2.41}$$

Fourthly, the maximal correlation lower bound in part 7 of Proposition 2.3 can be achieved with equality. For instance, let $f(t) = t\log(t)$ and consider a doubly symmetric binary source (DSBS) with parameter $\alpha \in [0,1]$, denoted $\mathsf{DSBS}(\alpha)$. A DSBS describes a joint distribution of two uniform Bernoulli random variables (X, Y), where X is passed through a binary symmetric channel (BSC) with crossover probability α, denoted $\mathsf{BSC}(\alpha)$, to produce Y. Recall that a $\mathsf{BSC}(\alpha)$ is a channel from $\mathcal{X} = \{0,1\}$ to $\mathcal{Y} = \{0,1\}$ such that:

$$\forall x, y \in \{0,1\}, \;\; P_{Y|X}(y|x) = \begin{cases} 1-\alpha & , \quad y = x \\ \alpha & , \quad y \neq x \end{cases}. \tag{2.42}$$

It is proven in [5] that for $(X, Y) \sim \mathsf{DSBS}(\alpha)$, the maximal correlation lower bound holds with equality:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = (1-2\alpha)^2 \tag{2.43}$$

where $\eta_{\chi^2}(P_X, P_{Y|X}) = (1-2\alpha)^2$ can be readily computed using the singular value characterization of maximal correlation presented in Proposition 2.2. As another example, consider $P_{Y|X}$ defined by an $|\mathcal{X}|$-ary erasure channel $E_\beta \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ with erasure probability $\beta \in [0,1]$, which has input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y} = \mathcal{X} \cup \{\mathsf{e}\}$, where e is the erasure symbol. Recall that given an input $x \in \mathcal{X}$, E_β erases x and outputs e

with probability β, and copies its input x with probability $1-\beta$. It is straightforward to verify that $D_f(R_X E_\beta || P_X E_\beta) = (1-\beta) D_f(R_X || P_X)$ for every $R_X, P_X \in \mathcal{P}_\mathcal{X}$. Therefore, for every input pmf $P_X \in \mathcal{P}_\mathcal{X}$ and every f-divergence, $\eta_f(P_X, P_{Y|X}) = 1 - \beta$.

Finally, we note that although we independently proved part 7 of Proposition 2.3 using the local approximation of f-divergence idea from [189, Theorem 5], the same idea is used by [234, Theorem III.3] and [231, Theorem 2] to prove this result. In fact, this idea turns out to stem from the proof of [49, Theorem 5.4] (which is presented later in part 6 of Proposition 2.5).

■ 2.2.3 Coefficients of Ergodicity

Before discussing contraction coefficients that depend solely on channels, we briefly introduce the broader notion of coefficients of ergodicity. Coefficients of ergodicity were first studied in the context of understanding ergodicity and convergence rates of finite state-space (time) inhomogeneous Markov chains, cf. [248, Section 1]. We present their definition below.

Definition 2.4 (Coefficient of Ergodicity [249, Definition 4.6]). *A* coefficient of ergodicity *is a continuous scalar function $\eta : \mathcal{P}_{\mathcal{Y}|\mathcal{X}} \to [0,1]$ from $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ (with fixed dimension) to $[0,1]$.*[23] *Such a coefficient is* proper *if for any $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, $\eta(W) = 0$ if and only if $W = \mathbf{1}P_Y$ for some pmf $P_Y \in \mathcal{P}_\mathcal{Y}$ (i.e. W is unit rank).*

One useful property of proper coefficients of ergodicity is their connection to weak ergodicity. Consider a sequence of row stochastic matrices $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$ that define an inhomogeneous Markov chain on the state space $\mathcal{X}$. Let the forward product of $r \geq 1$ of these consecutive matrices starting at index $p \in \mathbb{N}$ be:

$$T_{(p,r)} \triangleq \prod_{i=0}^{r-1} W_{p+i} \,. \tag{2.44}$$

The Markov chain $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$ is said to be *weakly ergodic* (in the Kolmogorov sense) if for all $x_1, x_2, x_3 \in \mathcal{X}$ and all $p \in \mathbb{N}$ [249, Definition 4.4]:

$$\lim_{r\to\infty} \left[T_{(p,r)}\right]_{x_1,x_3} - \left[T_{(p,r)}\right]_{x_2,x_3} = 0 \,. \tag{2.45}$$

This definition captures the intuition that the rows of a forward product should equalize when $r \to \infty$ for an ergodic Markov chain.[24] The next proposition conveys that weak ergodicity can be equivalently defined using proper coefficients of ergodicity.

[23] The set $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ is endowed with the standard topology induced by the Frobenius norm. Furthermore, $\mathcal{X}$ is typically the same as $\mathcal{Y}$ in Markov chain settings.

[24] Note that if the limiting row stochastic matrix $\lim_{r\to\infty} T_{(p,r)}$ exists for all $p \in \mathbb{N}$, then the Markov chain is *strongly ergodic* [249, Definition 4.5].

Proposition 2.4 (Weak Ergodicity [249, Lemma 4.1]). *Let $\eta : \mathcal{P}_{\mathcal{X}|\mathcal{X}} \to [0,1]$ be a proper coefficient of ergodicity. Then, the inhomogeneous Markov chain $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$ is weakly ergodic if and only if:*

$$\forall p \in \mathbb{N}, \quad \lim_{r\to\infty} \eta(T_{(p,r)}) = 0\,.$$

To intuitively understand this result, notice that $T_{(p,r)}$ becomes (approximately) unit rank as $r \to \infty$ for a weakly ergodic Markov chain. So, we also expect $\lim_{r\to\infty} \eta(T_{(p,r)}) = 0$, since a proper coefficient of ergodicity is continuous, and equals zero when its input is unit rank. We refer readers to [249, Lemma 4.1] for a formal proof of Proposition 2.4. We also suggest [248], [249, Chapters 3 and 4], [145], [247, Chapter 3], and the references therein for further expositions of such ideas.

One of the earliest and most notable examples of proper coefficients of ergodicity is the Dobrushin contraction coefficient. Given a row stochastic matrix $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ corresponding to a channel $P_{Y|X}$, its *Dobrushin contraction coefficient* is defined as the Lipschitz constant of the map $\mathcal{P}_{\mathcal{X}} \ni P_X \mapsto P_X W$ with respect to the ℓ^1-norm (or TV distance) [72]:[25]

$$\eta_{\mathsf{TV}}(W) \triangleq \sup_{\substack{R_X, P_X \in \mathcal{P}_{\mathcal{X}}: \\ R_X \neq P_X}} \frac{\|R_X W - P_X W\|_{\mathsf{TV}}}{\|R_X - P_X\|_{\mathsf{TV}}} \tag{2.46}$$

$$= \max_{\substack{v \in (\mathbb{R}^{|\mathcal{X}|})^*: \\ \|v\|_1 = 1,\, v\mathbf{1} = 0}} \|vW\|_1 \tag{2.47}$$

$$= \max_{R_X, P_X \in \mathcal{P}_{\mathcal{X}}} \|R_X W - P_X W\|_{\mathsf{TV}} \tag{2.48}$$

$$= \max_{x, x' \in \mathcal{X}} \left\| P_{Y|X=x} - P_{Y|X=x'} \right\|_{\mathsf{TV}} \tag{2.49}$$

$$= 1 - \min_{x, x' \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \min\left\{ P_{Y|X}(y|x), P_{Y|X}(y|x') \right\} \tag{2.50}$$

where the various equivalent characterizations of (2.46) in (2.47), (2.48), (2.49) (Dobrushin's two-point characterization [72]), and (2.50) (affinity characterization, cf. (2.5)) can be found in (or easily deduced from) [249, Chapter 4.3]. The characterization in (2.50) illustrates that $\eta_{\mathsf{TV}}(W) < 1$ if and only if W is a *scrambling matrix* (which means that no two rows of W are orthogonal) [249, p.82].[26]

In addition to the properties of proper coefficients of ergodicity, η_{TV} also exhibits the following properties:

1. **Lipschitz continuity [145, Theorem 3.4, Remark 3.5]:** For all $V, W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$:

$$|\eta_{\mathsf{TV}}(V) - \eta_{\mathsf{TV}}(W)| \leq \|V - W\|_\infty \tag{2.51}$$

[25] Based on the bibliographic discussion in [249, pp.144-147], the Dobrushin contraction coefficient (or equivalently, the Dobrushin ergodicity coefficient) may also be attributed (at least partly) to both Doeblin and Markov. In fact, the coefficient has been called the *Doeblin contraction coefficient* or presented as the *Markov contraction lemma* in the literature (see e.g. [155, p.619]).

[26] Thus, $\eta_{\mathsf{TV}}(W) < 1$ if and only if the zero error capacity of W is 0 [250].

where $\|\cdot\|_\infty$ denotes the maximum absolute row sum of a matrix.

2. **Sub-multiplicativity [249, Lemma 4.3]:** For every $V \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$ and $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$:
$$\eta_{\mathsf{TV}}(VW) \le \eta_{\mathsf{TV}}(V)\eta_{\mathsf{TV}}(W)\,. \tag{2.52}$$

3. **Sub-dominant eigenvalue bound [248, p.584, Equation (9)]:** For every $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ and every sub-dominant eigenvalue $\lambda \neq 1$ of W:
$$\eta_{\mathsf{TV}}(W) \ge |\lambda|\,. \tag{2.53}$$

The last two properties make η_{TV} a convenient tool for analyzing inhomogeneous Markov chains. As explained in [145, Section 1], for a homogeneous Markov chain $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ with stationary pmf $\pi \in \mathcal{P}_{\mathcal{X}}$, it is well-known that the *second largest eigenvalue modulus* (SLEM) of W, denoted $\mu(W)$, controls the rate of convergence to stationarity. Indeed, if $\mu(W) < 1$, then $\mu(W^n) = \mu(W)^n$, and $\lim_{n\to\infty} W^n = \mathbf{1}\pi$ with rate determined by $\mu(W)$. However, for an inhomogeneous Markov chain $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$, $\mu(T_{(1,n)}) \neq \prod_{i=1}^n \mu(W_i)$ in general because SLEMs are not multiplicative. The last two properties of η_{TV} illustrate that it is a viable replacement for SLEMs in the study of inhomogeneous Markov chains.

■ 2.2.4 Contraction Coefficients of Channels

Contraction coefficients of channels form a broad class of coefficients of ergodicity. They are defined similarly to (2.46), but using f-divergences in place of TV distance.

Definition 2.5 (Contraction Coefficient of Channel). *For any discrete channel $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ corresponding to a conditional distribution $P_{Y|X}$, the* contraction coefficient *for a particular f-divergence is:*

$$\begin{aligned}\eta_f(P_{Y|X}) &\triangleq \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} \eta_f(P_X, P_{Y|X}) \\ &= \sup_{\substack{R_X, P_X \in \mathcal{P}_{\mathcal{X}}: \\ 0<D_f(R_X||P_X)<+\infty}} \frac{D_f(R_X W||P_X W)}{D_f(R_X||P_X)}\end{aligned}$$

where the supremum is taken over all pmfs R_X and P_X such that $0 < D_f(R_X||P_X) < +\infty$. Furthermore, if Y is a constant a.s., we define $\eta_f(P_{Y|X}) = 0$.

This definition transparently yields SDPIs analogous to (2.27) and (2.28) for contraction coefficients of channels. Furthermore, a version of Proposition 2.1 also holds for contraction coefficients of channels. Indeed, using Definition 2.5 and Proposition 2.1, we observe that for any discrete channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, and any convex function $f : (0, \infty) \to \mathbb{R}$ that is differentiable, has uniformly bounded derivative in some neighborhood of unity, and satisfies $f(1) = 0$, we have:

$$\eta_f(P_{Y|X}) = \sup_{\substack{P_{U,X}: U \to X \to Y \\ 0<I_f(U;X)<+\infty}} \frac{I_f(U;Y)}{I_f(U;X)} \tag{2.54}$$

where the supremum is taken over all joint pmfs $P_{U,X}$ and finite alphabets $\mathcal{U}$ of U such that $U \to X \to Y$ form a Markov chain. The specialization of this result for KL divergence can be found in [58, p.345, Problem 15.12] (finite alphabet case) and [231] (general alphabet case).

There are two important examples of contraction coefficients of channels: the Dobrushin contraction coefficient for TV distance (defined in (2.46)), and the contraction coefficient for KL divergence. As seen earlier, given a channel $P_{Y|X}$, we use the notation $\eta_{\mathsf{TV}}(P_{Y|X})$, $\eta_{\mathsf{KL}}(P_{Y|X})$, and $\eta_{\chi^2}(P_{Y|X})$ to represent the contraction coefficient of $P_{Y|X}$ for TV distance, KL divergence, and χ^2-divergence, respectively. It is proved in [5] that for any channel $P_{Y|X}$, we have:

$$\eta_{\mathsf{KL}}(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X})\,. \tag{2.55}$$

Therefore, we do not need to consider η_{KL} and η_{χ^2} separately when studying contraction coefficients of channels. We remark that an alternative proof of (2.55) (which holds for general measurable spaces) is given in [231, Theorem 3]. Furthermore, a perhaps lesser known observation is that the proof technique of [85, Lemma 1, Theorem 1] (which analytically computes $\eta_{\mathsf{KL}}(P_{Y|X})$ for any binary channel $P_{Y|X}$ with $|\mathcal{X}| = |\mathcal{Y}| = 2$), when appropriately generalized for arbitrary finite alphabet sizes, also yields a proof of (2.55). It is worth mentioning that the main contribution of Evans and Schulman in [85] is an inductive approach to upper bound η_{KL} in Bayesian networks (or directed acyclic graphs). We refer readers to [231] for an insightful distillation of this approach, as well as for proofs of its generalization to TV distance (via Goldstein's simultaneously maximal coupling representation of the TV distance between two joint distributions [105]) and its connection to site percolation.

We next present some well-known properties of contraction coefficients of channels.

Proposition 2.5 (Properties of Contraction Coefficients of Channels). *The contraction coefficient for an f-divergence satisfies the following properties:*

1. (Normalization) *For any channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, we have that $0 \le \eta_f(P_{Y|X}) \le 1$.*
2. (Independence [49, Section 4]) *Given a channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, if X and Y are independent, then $\eta_f(P_{Y|X}) = 0$. Conversely, if f is strictly convex at unity and $\eta_f(P_{Y|X}) = 0$, then X and Y are independent.*
3. (Scrambling [49, Theorem 4.2]) *Given a channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, $P_{Y|X}$ is a scrambling matrix if and only if $\eta_f(P_{Y|X}) < 1$.*
4. (Convexity [49, Section 4], [234, Proposition III.3]) *The function $\mathcal{P}_{\mathcal{Y}|\mathcal{X}} \ni P_{Y|X} \mapsto \eta_f(P_{Y|X})$ is convex.*
5. (Sub-multiplicativity [49, Section 4]) *If $U \to X \to Y$ are discrete random variables with finite ranges that form a Markov chain, then we have:*

$$\eta_f(P_{Y|U}) \le \eta_f(P_{X|U})\eta_f(P_{Y|X})\,.$$

6. (η_{χ^2} Lower Bound [49, Theorem 5.4], [50, Proposition II.6.15]) *Given a channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, if f is twice differentiable at unity with $f''(1) > 0$, then we have:*

$$\eta_{\chi^2}(P_{Y|X}) \leq \eta_f(P_{Y|X})\,.$$

7. (η_{TV} Upper Bound [49, Theorem 4.1], [50, Proposition II.4.10]) *For any channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, we have:*

$$\eta_f(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X})\,.$$

We omit proofs of these results, because the proofs are either analogous to the corresponding proofs in Proposition 2.3, or are given in the associated references. Parts 1, 2, and 4 of Proposition 2.5 portray that contraction coefficients of channels are often valid proper coefficients of ergodicity.[27] We note that part 3 illustrates that $\eta_f(P_{Y|X}) = 1$ if and only if $\eta_{\mathsf{TV}}(P_{Y|X}) = 1$ [49, Theorem 4.2], and it is straightforward to verify that $\eta_f(P_{Y|X}) = 1$ if and only if the zero error capacity of $P_{Y|X}$ is strictly positive [250]. We also remark that an extremization result analogous to part 6 of Proposition 2.3, albeit less meaningful, can be derived in part 5 of Proposition 2.5.

While (2.55) shows that part 6 of Proposition 2.5 can be easily achieved with equality, the inequality in part 7 is often strict. For example, when $P_{Y|X}$ is a binary channel with parameters $a, b \in [0,1]$ and row stochastic transition probability matrix:

$$W = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} \tag{2.56}$$

it is straightforward to verify that $\eta_{\mathsf{KL}}(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X})$, with the inequality usually strict, since we have:

$$\eta_{\mathsf{KL}}(P_{Y|X}) = 1 - \left(\sqrt{a(1-b)} + \sqrt{b(1-a)}\right)^2 \tag{2.57}$$

$$\eta_{\mathsf{TV}}(P_{Y|X}) = |1-a-b| \tag{2.58}$$

where (2.57) is proved in [85, Theorem 1], and (2.58) is easily computed via (2.49). Moreover, in the special case where $P_{Y|X}$ is a $\mathsf{BSC}(\alpha)$ with $\alpha \in [0,1]$, we get [5]:

$$\eta_{\mathsf{KL}}(P_{Y|X}) = (1-2\alpha)^2 \leq |1-2\alpha| = \eta_{\mathsf{TV}}(P_{Y|X})\,. \tag{2.59}$$

On the other hand, as shown towards the end of subsection 2.2.2, $\eta_f(P_{Y|X}) = 1-\beta$ for every f-divergence when $P_{Y|X}$ is an $|\mathcal{X}|$-ary erasure channel with erasure probability $\beta \in [0,1]$.

In view of part 6 and (2.55), it is natural to wonder whether there are other f-divergences whose contraction coefficients (for channels) also collapse to η_{χ^2}. The following result from [46, Theorem 1] generalizes (2.55) and addresses this question.

[27] The convexity of $P_{Y|X} \mapsto \eta_f(P_{Y|X})$ in part 4 of Proposition 2.5 implies that this map is continuous on the interior of $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. So, only η_f that are also continuous on the boundary of $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ are valid coefficients of ergodicity.

Proposition 2.6 (Contraction Coefficients for Non-Linear Operator Convex f-Divergences [46, Theorem 1], [50]). *For every non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$, and every channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, we have:*

$$\eta_f(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X}) \,.$$

The proof of [46, Theorem 1] relies on an elegant integral representation of operator convex functions. Such representations are powerful tools for proving inequalities between contraction coefficients, and we will use them to generalize Proposition 2.6 in chapter 3. In fact, part 7 of Proposition 2.5 can also be proved using an integral representation argument, cf. [234, Theorem III.1]. In closing this overview, we also refer readers to [231, Section 2] for a complementary and comprehensive survey of contraction coefficients, and for references to various applications of these ideas in the literature.

■ 2.3 Main Results and Discussion

We will primarily derive bounds between various contraction coefficients in this chapter. In particular, we will address the following leading questions:

1. *Can we achieve the maximal correlation lower bound in Proposition 2.3 by adding constraints to the extremal problem that defines contraction coefficients of source-channel pairs?*
   Yes, we can constrain the input f-divergence to be small as shown in Theorem 2.1 in subsection 2.3.1.

2. While we typically lower bound $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ using $\eta_{\chi^2}(P_X, P_{Y|X})$ (Proposition 2.3 part 7), we typically upper bound it using $\eta_{\mathsf{TV}}(P_{Y|X})$ (Proposition 2.5 part 7). *Is there a simple upper bound on $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ in terms of $\eta_{\chi^2}(P_X, P_{Y|X})$?*
   Yes, two such bounds are given in Corollary 2.1 and Theorem 2.3 in subsection 2.3.2.

3. *Can we extend this upper bound for KL divergence to other f-divergences?*
   Yes, a more general bound is presented in Theorem 2.2 in subsection 2.3.2.

4. When X and Y are jointly Gaussian, the mutual information characterization in (2.30) can be used to establish that $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$, cf. [82, Theorem 7]. *Is there a simple proof of this result that directly uses the definition of η_{KL}? Does this equality hold when we add a power constraint to the extremization in η_{KL}?*
   Yes, we discuss the Gaussian case in subsection 2.3.3, and prove this equality for η_{KL} with a power constraint in Theorem 2.4. Our proof also establishes the known equality using the KL divergence definition of η_{KL}.

The bounds we will derive in response to questions 2, 3, and 4 have the form of the upper bound in:

$$\eta_{\chi^2}(P_X, P_{Y|X}) \le \eta_f(P_X, P_{Y|X}) \le C\, \eta_{\chi^2}(P_X, P_{Y|X}) \tag{2.60}$$

where the first inequality is simply the maximal correlation lower bound from Proposition 2.3, and the constant C depends on $P_{X,Y}$ and f; note that $C = 1$ in the setting of question 4. We refer to such bounds as *linear bounds* between contraction coefficients of source-channel pairs. We state our main results in the next few subsections.

■ 2.3.1 Local Approximation of Contraction Coefficients

We assume in this subsection and in subsection 2.3.2 that we are given the random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with joint pmf $P_{X,Y}$ such that the marginal pmfs satisfy $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Our first result portrays that forcing the input f-divergence to be small translates general contraction coefficients into the contraction coefficient for χ^2-divergence.

Theorem 2.1 (Local Approximation of Contraction Coefficients). *Suppose we are given a convex function $f : (0, \infty) \to \mathbb{R}$ that is strictly convex at unity and twice differentiable at unity with $f(1) = 0$ and $f''(1) > 0$. Then, we have:*

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \lim_{\delta \to 0^+} \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}}: \\ 0 < D_f(R_X||P_X) \leq \delta}} \frac{D_f(R_X W||P_X W)}{D_f(R_X||P_X)}$$

where $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ is the row stochastic transition probability matrix representing the channel $P_{Y|X}$.

We refer readers to appendix A.3 for the proof, and note that the specialization of Theorem 2.1 for KL divergence was presented along with a proof sketch in the conference paper [189, Theorem 3]. We now make several pertinent remarks. Firstly, notice that the proof of part 7 of Proposition 2.3 in appendix A.2 (or the independent proofs in [234, Theorem III.2] and [231, Theorem 2]) already captures the intuition that performing the optimization of $\eta_f(P_X, P_{Y|X})$ over local perturbations of P_X yields $\eta_{\chi^2}(P_X, P_{Y|X})$ due to (2.25) and (2.39). However, this proof (with minor modifications) only demonstrates that $\eta_{\chi^2}(P_X, P_{Y|X})$ is upper bounded by the right hand side of Theorem 2.1. While it is intuitively clear that this upper bound is met with equality, the formal proof requires a few technical details as shown in appendix A.3.

Secondly, Theorem 2.1 transparently portrays that the maximal correlation lower bound in part 7 of Proposition 2.3 can be achieved when the optimization that defines $\eta_f(P_X, P_{Y|X})$ imposes an additional constraint that the input f-divergence is small. (Hence, Theorem 2.1 implies the maximal correlation lower bound.) This insight has proved useful in comparing $\eta_{\chi^2}(P_X, P_{Y|X})$ and $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ in statistical contexts [152, p.5].

Thirdly, Theorem 2.1 can be construed as a minimax characterization of the contraction coefficient for χ^2-divergence, $\eta_{\chi^2}(P_X, P_{Y|X})$, since the supremum of the ratio of f-divergences is a non-increasing function of δ and the limit (as $\delta \to 0^+$) can therefore be replaced by an infimum (over all $\delta > 0$).

Fourthly, when the conditions of Proposition 2.1 and Theorem 2.1 hold, it is straightforward to verify that:

$$\eta_f(P_X, P_{Y|X}) = \lim_{\delta \to 0^+} \sup_{\substack{P_{U|X}: U \to X \to Y \\ 0 < I_f(U;X) \leq \delta}} \frac{I_f(U;Y)}{I_f(U;X)} \tag{2.61}$$

where the supremum is taken over all conditional distributions $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$ such that $\mathcal{U} = \{0,1\}$, $U \sim \mathsf{Bernoulli}(\frac{1}{2})$, and $U \to X \to Y$ form a Markov chain. Thus, the small input f-divergence constraint in the f-divergence formulation of $\eta_f(P_X, P_{Y|X})$ corresponds to the small $I_f(U;X)$ and $U \sim \mathsf{Bernoulli}(\frac{1}{2})$ constraints in (2.61).

Lastly, consider the trajectory of input pmfs $R_X^{(\epsilon)} = P_X + \epsilon\, K_X^* \,\mathsf{diag}(\sqrt{P_X})$, where $\epsilon > 0$ is sufficiently small, and $K_X^* \in \left(\mathbb{R}^{|\mathcal{X}|}\right)^*$ is the unit norm left singular vector corresponding to the second largest singular value of the DTM B (see (2.39)). As the proof in appendix A.3 illustrates, this trajectory satisfies $\lim_{\epsilon \to 0} D_f(R_X^{(\epsilon)} || P_X) = 0$ and achieves $\eta_{\chi^2}(P_X, P_{Y|X})$ in Theorem 2.1:

$$\lim_{\epsilon \to 0} \frac{D_f(R_X^{(\epsilon)} W || P_X W)}{D_f(R_X^{(\epsilon)} || P_X)} = \eta_{\chi^2}(P_X, P_{Y|X})\,. \tag{2.62}$$

The corresponding trajectory of conditional distributions for (2.61) is:

$$\left\{ P_{X|U=u}^{(\epsilon)} = P_X + (2u-1)\,\epsilon\, K_X^* \,\mathsf{diag}\left(\sqrt{P_X}\right) : u \in \{0,1\} \right\}$$

where $\epsilon > 0$ is sufficient small. It is straightforward to verify that this trajectory satisfies $\lim_{\epsilon \to 0} I_f(P_U, P_{X|U}^{(\epsilon)}) = 0$, produces P_X after $(P_U, P_{X|U}^{(\epsilon)})$ is marginalized, and achieves $\eta_{\chi^2}(P_X, P_{Y|X})$ in (2.61):

$$\lim_{\epsilon \to 0} \frac{I_f(P_U, P_{Y|U}^{(\epsilon)})}{I_f(P_U, P_{X|U}^{(\epsilon)})} = \eta_{\chi^2}(P_X, P_{Y|X}) \tag{2.63}$$

where $U \sim \mathsf{Bernoulli}(\frac{1}{2})$, and $P_{Y|U}^{(\epsilon)} = P_{X|U}^{(\epsilon)} P_{Y|X}$ as row stochastic matrices.

■ 2.3.2 Linear Bounds between Contraction Coefficients

For any joint pmf $P_{X,Y}$ with $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, our next result provides a linear upper bound on $\eta_f(P_X, P_{Y|X})$ using $\eta_{\chi^2}(P_X, P_{Y|X})$ for a certain class of f-divergences.

Theorem 2.2 (Contraction Coefficient Bound). *Suppose we are given a continuous convex function $f : [0,\infty) \to \mathbb{R}$ that is thrice differentiable at unity with $f(1) = 0$ and $f''(1) > 0$, and satisfies (2.80) for every $t \in (0,\infty)$ (see subsection 2.4.1). Suppose further that the difference quotient $g : (0,\infty) \to \mathbb{R}$, defined as $g(x) = \frac{f(x)-f(0)}{x}$, is concave. Then, we have:*

$$\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f''(1) \min_{x \in \mathcal{X}} P_X(x)}\, \eta_{\chi^2}(P_X, P_{Y|X})\,.$$

Theorem 2.2 is proved in subsection 2.4.2. The conditions on f ensure that the resulting f-divergence exhibits the properties of KL divergence required by the proof of Theorem 2.3 (see below). So, a similar proof technique also works for Theorem 2.2. A straightforward specialization of this theorem for KL divergence (which we first proved in the conference paper [189, Theorem 10]) is presented next.

Corollary 2.1 (KL Contraction Coefficient Bound).

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \leq \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{\min\limits_{x \in \mathcal{X}} P_X(x)} \, .$$

Proof. This can be recovered from Theorem 2.2 by verifying that $f(t) = t \log(t)$ satisfies the conditions of Theorem 2.2, cf. [101]. See appendix A.4 for details. ■

The constant in this upper bound on $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ can be improved, and the ensuing theorem presents this improvement.

Theorem 2.3 (Refined KL Contraction Coefficient Bound).

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \leq \frac{2\,\eta_{\chi^2}(P_X, P_{Y|X})}{\phi\Big(\max\limits_{A \subseteq \mathcal{X}} \pi(A)\Big) \min\limits_{x \in \mathcal{X}} P_X(x)}$$

where $\pi(A) \triangleq \min\{P_X(A), 1 - P_X(A)\}$ *for any* $A \subseteq \mathcal{X}$*, and the function* $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ *is defined in* (2.73) *(see subsection 2.4.1).*

Theorem 2.3 is also proved in subsection 2.4.2, and it is tighter than the bound in Corollary 2.1 due to (2.75) in subsection 2.4.1. We now make some pertinent remarks about Corollary 2.1 and Theorems 2.2 and 2.3.

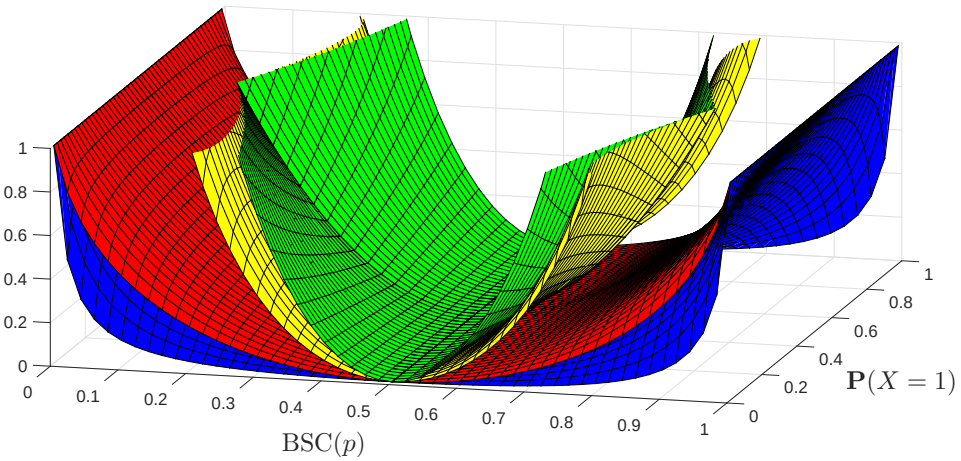
Firstly, as shown in Figure 2.1(a), the upper bounds in these results can be strictly less than the trivial bound of unity. For example, when $(X, Y) \sim \mathsf{DSBS}(p)$ for some $p \in [0, 1]$ (which is a slice along $\mathbb{P}(X = 1) = \frac{1}{2}$ in Figure 2.1(a)), the upper bounds in Corollary 2.1 and Theorem 2.3 are both equal to:

$$\frac{2\,\eta_{\chi^2}(P_X, P_{Y|X})}{\phi\Big(\max\limits_{A \subseteq \mathcal{X}} \pi(A)\Big) \min\limits_{x \in \mathcal{X}} P_X(x)} = \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{\min\limits_{x \in \mathcal{X}} P_X(x)} = 2(1-2p)^2 \tag{2.64}$$

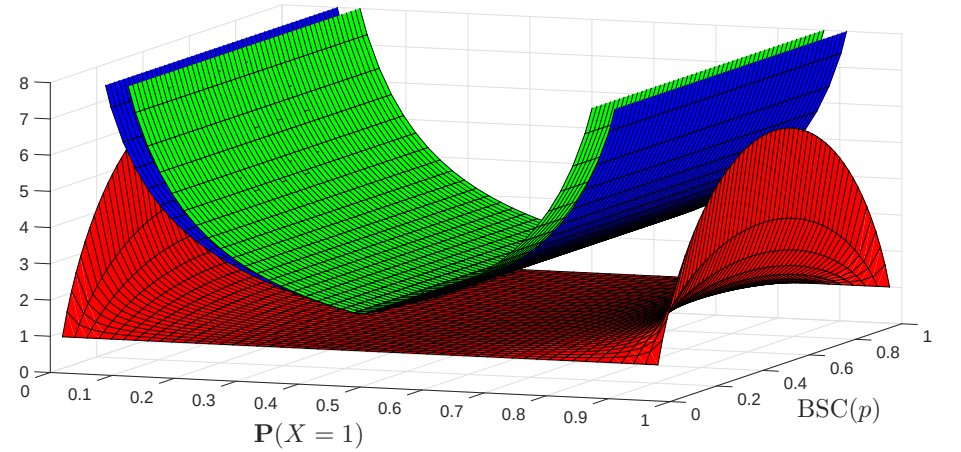
using (2.43) and the fact that $\max_{A \subseteq \mathcal{X}} \pi(A) = \frac{1}{2}$. This upper bound is tighter than the trivial bound of unity when:

$$2(1-2p)^2 < 1 \quad \Leftrightarrow \quad \frac{2-\sqrt{2}}{4} < p < \frac{2+\sqrt{2}}{4} \, . \tag{2.65}$$

We also note that this upper bound is not achieved with equality in this scenario since $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = (1-2p)^2$, as shown in (2.43).



(a) Plots of $\eta_{\chi^2}(P_X, P_{Y|X})$ (blue mesh), $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ (red mesh), and linear upper bounds on $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$. The green mesh denotes the upper bound from Corollary 2.1, and the yellow mesh denotes the tighter upper bound from Theorem 2.3.



(b) Plots of upper bounds on the ratio $\eta_{\mathsf{KL}}(P_X, P_{Y|X})/\eta_{\chi^2}(P_X, P_{Y|X})$, which is denoted by the red mesh. The upper bound $1/\min_{x\in\mathcal{X}} P_X(x)$ from Corollary 2.1 is the green mesh, and the upper bound $2/(\phi(\max_{A\subseteq\mathcal{X}} \pi(A)) \min_{x\in\mathcal{X}} P_X(x))$ from Theorem 2.3 is the blue mesh.

Figure 2.1. Plots of the contraction coefficient bounds in Corollary 2.1 and Theorem 2.3 for a BSC, $P_{Y|X}$, with crossover probability $p \in [0,1]$, and input random variable $X \sim \mathsf{Bernoulli}(\mathbb{P}(X=1))$.

Secondly, our proofs of Theorems 2.2 and 2.3 will rely on extensions of the well-known *Pinsker's inequality* (or the Csiszár-Kemperman-Kullback-Pinsker inequality, cf. [283, Section V]) which upper bound TV distance using KL and other f-divergences. So, it is natural to ask: Are these bounds tighter than the TV distance contraction bound in part 7 of Proposition 2.5? As the ensuing example illustrates, our bounds are tighter

in certain regimes. Let $(X, Y) \sim \mathsf{DSBS}(p)$ for some $p \in [0, 1]$. Then, (2.64) presents the upper bounds in Corollary 2.1 and Theorem 2.3, and the TV distance contraction bound is:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \leq \eta_{\mathsf{KL}}(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X}) = |1 - 2p| \tag{2.66}$$

using Definition 2.5, part 7 of Proposition 2.5, and (2.59). Hence, our bound in (2.64) is tighter than the η_{TV} bound when:

$$2(1-2p)^2 < |1-2p| \quad \Leftrightarrow \quad \frac{1}{4} < p < \frac{1}{2} \text{ or } \frac{1}{2} < p < \frac{3}{4}. \tag{2.67}$$

Since our upper bounds can be greater than 1 (see (2.65)), we cannot hope to beat the η_{TV} (≤ 1) bound in all regimes. On the other hand, one advantage of our upper bounds is that they "match" the η_{χ^2} lower bound in part 7 of Proposition 2.3; we will illustrate a useful application of this in subsection 2.4.3.

Thirdly, we intuitively expect a bound between contraction coefficients to depend on the cardinalities $|\mathcal{X}|$ or $|\mathcal{Y}|$. Since the minimum probability term in all our upper bounds satisfies:

$$\frac{1}{\min\limits_{x \in \mathcal{X}} P_X(x)} \geq |\mathcal{X}| \tag{2.68}$$

we can superficially construe it as "modeling" $|\mathcal{X}|$. Unfortunately, this intuition is quite misleading. Simulations for the binary case, depicted in Figure 2.1(b), illustrate that the ratio $\eta_{\mathsf{KL}}(P_X, P_{Y|X})/\eta_{\chi^2}(P_X, P_{Y|X})$ increases significantly near the boundary of $\mathcal{P}_{\mathcal{X}}$ when any of the probability masses of P_X is close to 0. This effect, while unsurprising given the skewed nature of probability simplices at their boundaries with respect to KL divergence as the distance measure, is correctly captured by the upper bounds in Corollary 2.1 and Theorem 2.3 because $1/\min_{x \in \mathcal{X}} P_X(x)$ increases when any of the input probability masses tends to 0 (see Figure 2.1(b)). Clearly, linear upper bounds on $\eta_f(P_X, P_{Y|X})$ that are purely in terms of $|\mathcal{X}|$ or $|\mathcal{Y}|$ cannot capture this effect. This gives credence to the existence of the minimum probability term in our linear bounds.

Finally, we note that the inequality (2.68) does not preclude the possibility of $1/\min_{x \in \mathcal{X}} P_X(x)$ being much larger than $|\mathcal{X}|$. So, our bounds can become loose when $|\mathcal{X}|$ is large (see the example in subsection 2.4.4). As a result, the bounds in Theorem 2.2, Corollary 2.1, and Theorem 2.3 are usually of interest in the following settings:

1. $|\mathcal{X}|$ and $|\mathcal{Y}|$ are small: Figure 2.1 portrays that our bounds can be quite tight when $|\mathcal{X}| = |\mathcal{Y}| = 2$.

2. Weak dependence, i.e. $I(X;Y)$ is small: This situation naturally arises in the analysis of ergodicity of Markov chains—see subsection 2.4.3.

3. Product Distributions: If the underlying joint pmf is a product pmf, we can exploit tensorization of contraction coefficients (Proposition 2.3 part 5)—see subsection 2.4.4.

■ 2.3.3 Contraction Coefficients of Gaussian Random Variables

In this subsection, we consider contraction coefficients for KL and χ^2-divergences corresponding to Gaussian source-channel pairs. Suppose X and Y are jointly Gaussian random variables. Their joint distribution has three possible forms:

1. X or Y are constants $a.s.$ In this case, we define the contraction coefficients to be $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = 0$.
2. $aX + bY = c$ $a.s.$ for some constants $a, b, c \in \mathbb{R}$ such that $a \neq 0$ and $b \neq 0$. In this case, it is straightforward to verify that $\rho_{\mathsf{max}}(X;Y) = 1$, which implies that $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = 1$.[28]
3. The joint probability density function (pdf) of X and Y, $P_{X,Y}$, exists with respect to the Lebesgue measure on $\mathbb{R}$.

The final non-degenerate case is our regime of interest. For simplicity, we will assume that X and Y are zero-mean, and analyze the classical *additive white Gaussian noise* (AWGN) channel model [53, Chapter 9]:

$$Y = X + W, \quad X \perp\!\!\!\perp W \tag{2.69}$$

where the input is $X \sim \mathcal{N}(0, \sigma_X^2)$ with $\sigma_X^2 > 0$ (i.e. X has a Gaussian pdf with mean 0 and variance σ_X^2), the Gaussian noise is $W \sim \mathcal{N}(0, \sigma_W^2)$ with $\sigma_W^2 > 0$, and X is independent of W. This relation also defines the channel conditional pdfs $\{P_{Y|X=x} = \mathcal{N}(x, \sigma_W^2) : x \in \mathbb{R}\}$.

For the jointly Gaussian pdf $P_{X,Y}$ define above, the contraction coefficients for KL and χ^2-divergences are given by (cf. (2.29) and (2.36)):

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \sup_{\substack{R_X: \\ 0<D(R_X||P_X)<+\infty}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \tag{2.70}$$

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \sup_{\substack{R_X: \\ 0<\chi^2(R_X||P_X)<+\infty}} \frac{\chi^2(R_Y||P_Y)}{\chi^2(R_X||P_X)} \tag{2.71}$$

where the suprema are over all pdfs R_X (which differ from P_X on a set with non-zero Lebesgue measure),[29] and R_Y denotes the marginal pdf of Y after passing R_X

[28] Note that Definition 2.3 holds for general random variables, and (2.37) and part 7 of Proposition 2.3 (which also hold generally—see [231, Equations (9) and (13)]) can be used to conclude $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = 1$.

[29] When P_X is a general probability measure and $P_{Y|X}$ is a Markov kernel between two measurable spaces, the contraction coefficients for KL and χ^2-divergences are defined exactly as in (2.29) and (2.36) using the measure theoretic definitions of KL and χ^2-divergences [231, Section 2]. In (2.70), when we optimize over all probability measures R_X on $\mathbb{R}$ (with its Borel σ-algebra), the constraint $D(R_X||P_X) < +\infty$ implies that R_X must be absolutely continuous with respect to the Gaussian distribution P_X, cf. [230, Section 1.6]. Hence, the supremum in (2.70) can be taken over all pdfs R_X such that $0 < D(R_X||P_X) < +\infty$. A similar argument applies for (2.71). (Note that KL and χ^2-divergences for pdfs are defined just as in (2.8) and (2.9) with Lebesgue integrals replacing summations.)

through the AWGN channel $P_{Y|X}$. In particular, $R_Y = R_X * \mathcal{N}(0, \sigma_W^2)$, where $*$ denotes the *convolution* operation. Furthermore, we define the *contraction coefficient for KL divergence with average power constraint* $p \geq \sigma_X^2$ as:

$$\eta_{\mathsf{KL}}^{(p)}(P_X, P_{Y|X}) \triangleq \sup_{\substack{R_X : \mathbb{E}_{R_X}[X^2] \leq p \\ 0 < D(R_X||P_X) < +\infty}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \tag{2.72}$$

where the supremum is over all pdfs R_X satisfying the average power constraint $\mathbb{E}[X^2] \leq p$. Note that setting $p = +\infty$ yields the standard contraction coefficient in (2.70).

It is well-known in the literature that $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$ for the jointly Gaussian pdf $P_{X,Y}$ defined via (2.69). For example, [82, Theorem 7] proves this result in the context of investment portfolio theory, [217, p.2] proves a generalization of it in the context of Gaussian hypercontractivity, and [152, Section 5.2, part 5] proves it in an effort to axiomatize η_{KL}. While the proofs in [82, Theorems 6 and 7] and [152, Section 5.2, part 5] use the mutual information characterization of η_{KL} in (2.30) (cf. [231, Theorem 4]), our last main result provides an alternative proof of this result in section 2.5 that directly uses the KL divergence definition of η_{KL} in (2.70). Furthermore, our proof also establishes that $\eta_{\mathsf{KL}}^{(p)}(P_X, P_{Y|X})$ equals $\eta_{\chi^2}(P_X, P_{Y|X})$ for every $p \in [\sigma_X^2, \infty]$. Although this latter result follows easily from our proof, it has not explicitly appeared in the literature to our knowledge. The ensuing theorem states these results formally.

Theorem 2.4 (Gaussian Contraction Coefficients). *Given the jointly Gaussian pdf $P_{X,Y}$, defined via (2.69) with source $P_X = \mathcal{N}(0, \sigma_X^2)$ and channel $\{P_{Y|X=x} = \mathcal{N}(x, \sigma_W^2) : x \in \mathbb{R}\}$ such that $\sigma_X^2, \sigma_W^2 > 0$, the following quantities are equivalent:*

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\mathsf{KL}}^{(p)}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}$$

where the average power constraint $p \geq \sigma_X^2$.

As mentioned earlier, we prove this in section 2.5. In contrast to Theorem 2.4, where $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ and $\eta_{\mathsf{KL}}^{(p)}(P_X, P_{Y|X})$ can both be strictly less than 1, we note that the contraction coefficients for KL divergence of channels (i.e. the setting of Definition 2.5) are equal to 1 regardless of whether we impose power constraints, cf. [229, Section 1.2] and [76, Section 1].

■ 2.4 Proofs of Linear Bounds between Contraction Coefficients

In this section, we will prove Theorems 2.2 and 2.3. The central idea to establish these results entails upper and lower bounding the f-divergences in the numerator and denominator of Definition 2.2 respectively, using χ^2-divergences. To this end, we will illustrate some simple bounds between f-divergences and χ^2-divergence in the next subsection, and prove the main results in subsection 2.4.2.

■ 2.4.1 Bounds on f-Divergences using χ^2-Divergence

We first present bounds between KL divergence and χ^2-divergence. To derive our lower bound on KL divergence, we will require the following "distribution dependent refinement of Pinsker's inequality" proved in [222].

Lemma 2.1 (Distribution Dependent Pinsker's Inequality [222, Theorem 2.1]). *For any two pmfs $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$, we have:*

$$D(R_X||P_X) \geq \phi\Big(\max_{A\subseteq\mathcal{X}} \pi(A)\Big)\, \|R_X - P_X\|_{\mathsf{TV}}^2$$

where $\pi(A) = \min\{P_X(A), 1-P_X(A)\}$ for any $A \subseteq \mathcal{X}$, and the function $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ is defined as:

$$\phi(p) \triangleq \begin{cases} \frac{1}{1-2p}\log\left(\frac{1-p}{p}\right) & , \quad p \in \left[0, \frac{1}{2}\right) \\ 2 & , \quad p = \frac{1}{2} \end{cases}. \tag{2.73}$$

Moreover, this inequality uses the optimal distribution dependent constant in the sense that for any fixed $P_X \in \mathcal{P}_{\mathcal{X}}$:

$$\inf_{R_X \in \mathcal{P}_{\mathcal{X}}\setminus\{P_X\}} \frac{D(R_X||P_X)}{\|R_X - P_X\|_{\mathsf{TV}}^2} = \phi\Big(\max_{A\subseteq\mathcal{X}} \pi(A)\Big).$$

Recall that Pinsker's inequality states that for any $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$ (see e.g. [53, Lemma 11.6.1]):

$$D(R_X||P_X) \geq 2\,\|R_X - P_X\|_{\mathsf{TV}}^2. \tag{2.74}$$

Hence, Lemma 2.1 is tighter than Pinsker's inequality, because $0 \leq \max_{A\subseteq\mathcal{X}} \pi(A) \leq \frac{1}{2}$, and hence:

$$\phi\Big(\max_{A\subseteq\mathcal{X}} \pi(A)\Big) \geq 2 \tag{2.75}$$

with equality if and only if $\max_{A\subseteq\mathcal{X}} \pi(A) = \frac{1}{2}$, cf. [222, Section III]. The ensuing lemma uses Lemma 2.1 to lower bound KL divergence using χ^2-divergence.

Lemma 2.2 (KL Divergence Lower Bound). *Given any two pmfs $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$, we have:*[30]

$$D(R_X||P_X) \geq \frac{\phi\Big(\max_{A\subseteq\mathcal{X}} \pi(A)\Big)\min_{x\in\mathcal{X}} P_X(x)}{2}\,\chi^2(R_X||P_X)$$

where $\pi(\cdot)$ and $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ are defined in Lemma 2.1.

Proof. Observe that if $R_X = P_X$ or $\min_{x\in\mathcal{X}} P_X(x) = 0$, then the inequality is trivially satisfied. So, we assume without loss of generality that $R_X \neq P_X$ and $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$.

[30] Throughout this chapter, when $\min_{x\in\mathcal{X}} P_X(x) = 0$ and $\chi^2(R_X||P_X) = +\infty$, we assume that $\min_{x\in\mathcal{X}} P_X(x)\,\chi^2(R_X||P_X) = 0$.

Since χ^2-divergence resembles a weighted ℓ^2-norm, we first use Lemma 2.1 to get the lower bound:

$$D(R_X||P_X) \geq \phi\Big(\max_{A\subseteq\mathcal{X}} \pi(A)\Big) \frac{\|R_X - P_X\|_1^2}{4} \tag{2.76}$$

where we use the ℓ^1-norm characterization of TV distance given in (2.4). We next notice using (2.9) that:

$$\begin{aligned}\chi^2(R_X||P_X) &= \sum_{x\in\mathcal{X}} |R_X(x) - P_X(x)| \left|\frac{R_X(x) - P_X(x)}{P_X(x)}\right| \\ &\leq \frac{\|R_X - P_X\|_\infty}{\min_{x\in\mathcal{X}} P_X(x)} \|R_X - P_X\|_1 .\end{aligned}$$

This implies that:

$$\begin{aligned}\frac{\|R_X - P_X\|_1^2}{\min_{x\in\mathcal{X}} P_X(x)} &\geq \chi^2(R_X||P_X) \frac{\|R_X - P_X\|_1}{\|R_X - P_X\|_\infty} \\ &\geq \chi^2(R_X||P_X) \min_{\substack{S_X, Q_X\in\mathcal{P}_\mathcal{X}: \\ S_X\neq Q_X}} \frac{\|S_X - Q_X\|_1}{\|S_X - Q_X\|_\infty} \\ &= 2\,\chi^2(R_X||P_X)\end{aligned} \tag{2.77}$$

where we use the fact that:

$$\min_{\substack{S_X, Q_X\in\mathcal{P}_\mathcal{X}: \\ S_X\neq Q_X}} \frac{\|S_X - Q_X\|_1}{\|S_X - Q_X\|_\infty} = 2 . \tag{2.78}$$

To prove (2.78), note that for every $S_X, Q_X \in \mathcal{P}_\mathcal{X}$ (see e.g. [243, Lemma 1]):

$$\|S_X - Q_X\|_\infty \leq \frac{1}{2} \|S_X - Q_X\|_1$$

because $(S_X - Q_X)\mathbf{1} = 0$, and this inequality can in fact be tight. For example, choose any pmf $Q_X \in \mathcal{P}_\mathcal{X}^\circ$ and let $x_0 = \arg\min_{x\in\mathcal{X}} Q_X(x)$. Then, select $S_X \in \mathcal{P}_\mathcal{X}$ such that $S_X(x_0) = Q_X(x_0) + \delta$ for some sufficiently small $\delta > 0$, $S_X(x_1) = Q_X(x_1) - \delta$ for some $x_1 \in \mathcal{X}\backslash\{x_0\}$, and $S_X(x) = Q_X(x)$ for every other $x \in \mathcal{X}\backslash\{x_0, x_1\}$. These choices of S_X and Q_X yield $\|S_X - Q_X\|_\infty = \delta = \frac{1}{2}\|S_X - Q_X\|_1$.

Finally, combining (2.76) and (2.77), we get:

$$D(R_X||P_X) \geq \frac{\phi\Big(\max_{A\subseteq\mathcal{X}} \pi(A)\Big) \min_{x\in\mathcal{X}} P_X(x)}{2} \chi^2(R_X||P_X)$$

which completes the proof. ■

We remark that if we apply (2.75) to Lemma 2.2, or equivalently, if we use the standard Pinsker's inequality (2.74) instead of Lemma 2.1 in the proof of Lemma 2.2, then we obtain the well-known weaker inequality (see e.g. [245, Equation (338)]):

$$D\left(R_X||P_X\right) \geq \min_{x\in\mathcal{X}} P_X(x)\, \chi^2(R_X||P_X) \tag{2.79}$$

for every $R_X, P_X \in \mathcal{P}_\mathcal{X}$.

It is worth mentioning that a systematic method of deriving optimal distribution independent bounds between any pair of f-divergences is given by the *Harremoës-Vajda joint range* [122].[31] However, we cannot use this technique to derive lower bounds on KL divergence using χ^2-divergence since no such general lower bound exists (when both input distributions vary) [230, Section 7.3]. On the other hand, distribution dependent bounds can be easily found using ad hoc techniques. Our proof of Lemma 2.2 demonstrates one such ad hoc approach based on Pinsker's inequality.

It is tempting to try and improve Lemma 2.2 by using better lower bounds on KL divergence in terms of TV distance. For example, the best possible lower bound on KL divergence via TV distance is the lower boundary of their Harremoës-Vajda joint range, cf. [122, Figure 1]. This lower boundary, known as *Vajda's tight lower bound*, gives the minimum possible KL divergence for each value of TV distance, and is completely characterized using a parametric formula in [88, Theorem 1] (also see [230, Section 7.2.2]). Although Vajda's tight lower bound yields a non-linear lower bound on KL divergence using χ^2-divergence, this lower bound is difficult to apply in conjunction with Lemma 2.3 (shown below) to obtain a non-linear upper bound on a ratio of KL divergences using a ratio of χ^2-divergences (see the proof of Theorem 2.3 in subsection 2.4.2). For this reason, we resort to using simple linear bounds between KL and χ^2-divergence, which yields a linear bound in Theorem 2.3.

Another subtler reason for proving a linear lower bound on KL divergence using χ^2-divergence is to exploit Lemma 2.1. Although Pinsker's inequality is the best lower bound on KL divergence using squared TV distance over all pairs of input pmfs (see e.g. [88, Equation (9)]), the contraction coefficients in subsection 2.3.2 have a fixed source pmf P_X. Therefore, we can use the distribution dependent improvement of Pinsker's inequality in Lemma 2.1 to obtain a tighter bound than (2.79).

We next present an upper bound on KL divergence using χ^2-divergence which trivially follows from Jensen's inequality. This bound was derived in the context of studying ergodicity of Markov chains in [267], and has been re-derived in the study of inequalities related to f-divergences, cf. [74, 244] (also see [100, Theorem 5]).

Lemma 2.3 (KL Divergence Upper Bound [267]). *For any two pmfs $P_X, R_X \in \mathcal{P}_\mathcal{X}$, we have:*

$$D(R_X||P_X) \leq \log\Big(1+\chi^2(R_X||P_X)\Big) \leq \chi^2(R_X||P_X)\,.$$

[31] A "distribution independent" bound between two f-divergences is a bound that only depends on the input distributions through the corresponding f-divergences.

Proof. We provide a proof for completeness, cf. [74]. Assume without loss of generality that there does not exist $x \in \mathcal{X}$ such that $R_X(x) > P_X(x) = 0$. (If this is not the case, then $\chi^2(R_X||P_X) = +\infty$ and the inequalities are trivially true.) So, restricting $\mathcal{X}$ to be the support of P_X, we assume that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ (which ensures that none of the ensuing quantities are infinity). Since $x \mapsto \log(x)$ is a concave function, using Jensen's inequality, we have:

$$\begin{aligned} D(R_X||P_X) &= \sum_{x\in\mathcal{X}} R_X(x) \log\left(\frac{R_X(x)}{P_X(x)}\right) \\ &\leq \log\left(\sum_{x\in\mathcal{X}} \frac{R_X(x)^2}{P_X(x)}\right) \\ &= \log\Big(1 + \chi^2(R_X||P_X)\Big) \\ &\leq \chi^2(R_X||P_X) \end{aligned}$$

where the third equality follows from (2.9) after some algebra, and the final inequality follows from the well-known inequality: $\log(1+x) \leq x$ for all $x > -1$. ■

We remark that the first non-linear bound in Lemma 2.3 turns out to capture the Harremoës-Vajda joint range [230, Section 7.3]. Although it is tighter than the second linear bound, we will use the latter to prove Theorem 2.3 (as explained earlier). The latter bound has also been derived in [60, Lemma 6.3].

We now present bounds between general f-divergences and χ^2-divergence. To derive our lower bound on f-divergences, we first state a generalization of Pinsker's inequality for f-divergences that is proved in [101].

Lemma 2.4 (Generalized Pinsker's Inequality for f-Divergence [101, Theorem 3]). *Suppose we are given a convex function $f : (0,\infty) \to \mathbb{R}$ that is thrice differentiable at unity with $f(1) = 0$ and $f''(1) > 0$, and satisfies:*

$$(f(t) - f'(1)(t-1))\left(1 - \frac{f'''(1)}{3f''(1)}(t-1)\right) \geq \frac{f''(1)}{2}(t-1)^2 \tag{2.80}$$

for every $t \in (0,\infty)$. Then, we have for every $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$:

$$D_f(R_X||P_X) \geq 2\, f''(1)\, \|R_X - P_X\|_{\mathsf{TV}}^2 \,.$$

Moreover, this inequality uses the optimal constant in the sense that:

$$\inf_{\substack{R_X, P_X \in \mathcal{P}_{\mathcal{X}}: \\ R_X \neq P_X}} \frac{D_f(R_X||P_X)}{\|R_X - P_X\|_{\mathsf{TV}}^2} = 2\, f''(1)\,.$$

We remark that $f(t) = t\log(t)$ satisfies the conditions of Lemma 2.4 with $f''(1) = 1$ as shown in appendix A.4; this yields the standard Pinsker's inequality presented in

(2.74). Since (2.80) can be difficult to check for other f-divergences, the author of [101] provides sufficient conditions for (2.80) in [101, Corollary 4]. (These conditions can be verified to yield a variant of Pinsker's inequality for *Rényi divergences* of order $\alpha \in (0,1)$ [101, Corollary 6].) The ensuing lemma uses Lemma 2.4 to establish a lower bound on certain f-divergences using χ^2-divergence which parallels Lemma 2.2 (or more precisely, (2.79), since it follows from the standard Pinsker's inequality).

Lemma 2.5 (f-Divergence Lower Bound). *Suppose we are given a convex function $f : (0,\infty) \to \mathbb{R}$ that is thrice differentiable at unity with $f(1) = 0$ and $f''(1) > 0$, and satisfies (2.80) for every $t \in (0,\infty)$. Then, for any two pmfs $R_X \in \mathcal{P}_{\mathcal{X}}$ and $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, we have:*

$$D_f(R_X||P_X) \geq f''(1) \min_{x \in \mathcal{X}} P_X(x)\, \chi^2(R_X||P_X)\,.$$

Proof. We follow the proof of Lemma 2.2 mutatis mutandis. Assume without loss of generality that $R_X \neq P_X$. The generalized Pinsker's inequality for f-divergences in Lemma 2.4 yields:

$$D_f(R_X||P_X) \geq \frac{f''(1)}{2}\, \|R_X - P_X\|_1^2$$

using the ℓ^1-norm characterization of TV distance in (2.4). Applying (2.77) to this inequality produces the desired result. ■

Note that setting $f(t) = t\log(t)$ in Lemma 2.5 gives (2.79).

Finally, we present an upper bound on certain f-divergences using χ^2-divergence which is analogous to Lemma 2.3. This upper bound was proved in [234, Lemma A.2] with the assumption that f is differentiable, but we only need to check differentiability at unity as seen below. (It is instructive for readers to revisit the proof of Lemma 2.3 to see how the ensuing proof generalizes it for f-divergences.)

Lemma 2.6 (f-Divergence Upper Bound [234, Lemma A.2]). *Suppose we are given a continuous convex function $f : [0,\infty) \to \mathbb{R}$ that is differentiable at unity with $f(1) = 0$ such that the difference quotient $g : (0,\infty) \to \mathbb{R}$, defined as $g(x) = \frac{f(x)-f(0)}{x}$, is concave.[32] Then, for any two pmfs $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$, we have:*

$$D_f(R_X||P_X) \leq \big(f'(1) + f(0)\big)\, \chi^2(R_X||P_X)\,.$$

Proof. We provide the proof in [234] for completeness. As in the proof of Lemma 2.3, we may assume without loss of generality that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ so that none of the ensuing quantities are infinity. We then have the following sequence of equalities and inequalities:

$$D_f(R_X||P_X) = \sum_{x \in \mathcal{X}} P_X(x)\, f\left(\frac{R_X(x)}{P_X(x)}\right)$$

[32] Since f is convex, it is clearly continuous on $(0,\infty)$. So, the continuity assumption on f only asserts that $f(0) = \lim_{t \to 0+} f(t)$ (see Definition 2.1).

$$
\begin{aligned}
&= f(0) + \sum_{x \in \mathcal{X}} R_X(x)\, g\left(\frac{R_X(x)}{P_X(x)}\right) \\
&\leq f(0) + g\left(\sum_{x \in \mathcal{X}} \frac{R_X(x)^2}{P_X(x)}\right) \\
&= f(0) + g\Big(1 + \chi^2(R_X||P_X)\Big) \qquad (2.81) \\
&\leq f(0) + g(1) + g'(1)\chi^2(R_X||P_X) \\
&= \big(f'(1) + f(0)\big)\, \chi^2(R_X||P_X)
\end{aligned}
$$

where the second equality uses the convention $0\, g(0) = 0$, the third inequality follows from Jensen's inequality since $g : (0, \infty) \to \mathbb{R}$ is concave, the fifth inequality is also a consequence of the concavity of $g : (0, \infty) \to \mathbb{R}$ as shown in [34, Section 3.1.3], and the final equality holds because $g(1) = -f(0)$ (as $f(1) = 0$) and:

$$
\begin{aligned}
g'(1) &= \lim_{\delta \to 0} \frac{g(1+\delta) + f(0)}{\delta} \\
&= \lim_{\delta \to 0} \frac{f(1+\delta) + \delta f(0)}{\delta(1+\delta)} \\
&= \left(\lim_{\delta \to 0} \frac{1}{1+\delta}\right)\left(f(0) + \lim_{\delta \to 0} \frac{f(1+\delta)}{\delta}\right) \\
&= f'(1) + f(0)\,.
\end{aligned}
$$

This completes the proof. ■

We note that (2.81) is the analogue of the tighter (non-linear) bound in Lemma 2.3. Furthermore, we remark that $g(x) = \frac{f(x)}{x}$ (when it is assumed to be concave) is a valid definition for the function in Lemma 2.6 instead of the difference quotient. The proof carries through with a constant of $f'(1)$ instead of $f'(1) + f(0)$. However, we choose the difference quotient to prove Lemma 2.6 in view of the affine invariance property of f-divergences (cf. subsection 2.2.1). It is easy to verify that the quantity $f'(1) + f(0)$ is invariant to appropriate affine shifts, but the quantity $f'(1)$ is not. We also remark that the constant $f''(1)$ in Lemma 2.5 is invariant to appropriate affine shifts.

■ 2.4.2 Proofs of Theorems 2.2 and 2.3

Recall from the outset of subsection 2.3.1 that we are given a joint pmf $P_{X,Y}$ such that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Moreover, we let $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denote the row stochastic transition probability matrix of the channel $P_{Y|X}$. Using Lemmata 2.2 and 2.3 from subsection 2.4.1, we can now prove Theorem 2.3.

Proof of Theorem 2.3. For every pmf $R_X \in \mathcal{P}_{\mathcal{X}}$ such that $R_X \neq P_X$, we have:

$$
\frac{D(R_X W||P_Y)}{D(R_X||P_X)} \leq \frac{2\,\chi^2(R_X W||P_Y)}{\phi\Big(\max\limits_{A \subseteq \mathcal{X}} \pi(A)\Big) \min\limits_{x \in \mathcal{X}} P_X(x)\, \chi^2(R_X||P_X)}
$$

using Lemmata 2.2 and 2.3, where $P_Y = P_X W$. Taking the supremum over all $R_X \neq P_X$ on both sides produces:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \leq \frac{2\,\eta_{\chi^2}(P_X, P_{Y|X})}{\phi\Big(\max\limits_{A \subseteq \mathcal{X}} \pi(A)\Big) \min\limits_{x \in \mathcal{X}} P_X(x)}$$

using (2.29) and (2.36). This completes the proof. ■

We now make a few pertinent remarks. Firstly, applying (2.79) instead of Lemma 2.2 in the preceding proof yields Corollary 2.1.

Secondly, while our conference paper proves Corollary 2.1 (see [189, Theorem 10]), it also proves the following weaker upper bound on $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ [189, Theorem 9]:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \leq \frac{2}{\min\limits_{x \in \mathcal{X}} P_X(x)}\, \eta_{\chi^2}(P_X, P_{Y|X}) \tag{2.82}$$

which is independently derived in [234, Equation III.19]. Our proof of (2.82) in [189, Theorem 9] uses the ensuing variant of (2.79) that is looser by a factor of 2, cf. [189, Lemma 6]:

$$D(S_X||Q_X) \geq \frac{\min\limits_{x \in \mathcal{X}} Q_X(x)}{2}\, \chi^2(S_X||Q_X) \tag{2.83}$$

for all $S_X, Q_X \in \mathcal{P}_{\mathcal{X}}$. This follows from executing the proof of Lemma 2.2 using the bound $\|S_X - Q_X\|_\infty \leq \|S_X - Q_X\|_1$ (which neglects the information that $(S_X - Q_X)\mathbf{1} = 0$) instead of (2.78), and then applying (2.75) to the resulting lower bound on KL divergence. Alternatively, we provide a proof of (2.83) via *Bregman divergences* in appendix A.5 for completeness. The improvement by a factor of 2 from (2.83) to (2.79) is also observed in [245, Remark 33], where the authors mention that our result [189, Theorem 9] (see (2.82)) in our conference paper can be improved by a factor of 2 by using (2.79) instead (2.83). We believe the authors of [245] may have missed our result [189, Theorem 10] (see Corollary 2.1) in the conference paper, which presents precisely this improvement by a factor of 2.

Lastly, we remark that [234, Section III-D] also presents upper bounds on the contraction coefficient $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ that use the function $\phi : [0, \frac{1}{2}] \to \mathbb{R}$, which stems from the refined Pinsker's inequality in [222]. However, these upper bounds are not in terms of $\eta_{\chi^2}(P_X, P_{Y|X})$.

We next prove Theorem 2.2 by combining Lemmata 2.5 and 2.6 from subsection 2.4.1.

Proof of Theorem 2.2. The conditions of Theorem 2.2 encapsulate all the conditions of Lemmata 2.5 and 2.6. Hence, using Lemmata 2.5 and 2.6, for every pmf $R_X \in \mathcal{P}_{\mathcal{X}}$ such that $R_X \neq P_X$, we have:

$$\frac{D_f(R_X W||P_Y)}{D_f(R_X||P_X)} \leq \frac{(f'(1) + f(0))\, \chi^2(R_X W||P_Y)}{f''(1) \min\limits_{x \in \mathcal{X}} P_X(x)\, \chi^2(R_X||P_X)}$$

where $P_Y = P_X W$. Taking the supremum over all $R_X \neq P_X$ on both sides produces:

$$\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f''(1) \min_{x \in \mathcal{X}} P_X(x)} \eta_{\chi^2}(P_X, P_{Y|X})$$

where we use Definition 2.2 and (2.36). This completes the proof. ■

We remark that [234, Theorem III.4] presents an alternative linear upper bound on $\eta_f(P_X, P_{Y|X})$ using $\eta_{\chi^2}(P_X, P_{Y|X})$. Suppose $f : [0, \infty) \to \mathbb{R}$ is a twice differentiable convex function that has $f(1) = 0$, is strictly convex at unity, and has non-increasing second derivative. If we further assume that the difference quotient $x \mapsto \frac{f(x)-f(0)}{x}$ is concave, then the following bound holds [234, Theorem III.4]:

$$\eta_f(P_X, P_{Y|X}) \leq \frac{2\,(f'(1) + f(0))}{f''(1/p_\star)} \eta_{\chi^2}(P_X, P_{Y|X}) \tag{2.84}$$

where $p_\star = \min_{x \in \mathcal{X}} P_X(x)$. Hence, when f is additionally thrice differentiable at unity, has $f''(1) > 0$, and satisfies (2.80) for every $t \in (0, \infty)$, we can improve the upper bound in Theorem 2.2 to:

$$\eta_f(P_X, P_{Y|X}) \leq \min\left\{ \frac{f'(1) + f(0)}{f''(1)\, p_\star}, \frac{2\,(f'(1) + f(0))}{f''(1/p_\star)} \right\} \eta_{\chi^2}(P_X, P_{Y|X})\,. \tag{2.85}$$

Observe that our bound in Theorem 2.2 is tighter than that in (2.84) if and only if:

$$\frac{2\,(f'(1) + f(0))}{f''(1/p_\star)} \geq \frac{f'(1) + f(0)}{f''(1)\, p_\star} \tag{2.86}$$

$$\Leftrightarrow \quad 2\, f''(1)\, p_\star \geq f''(1/p_\star)\,. \tag{2.87}$$

One function that satisfies the conditions of Theorem 2.2 and (2.84) as well as (2.87) is $f(t) = t \log(t)$. This engenders the improvement that Corollary 2.1 (which can be recovered from Theorem 2.2) offers over (2.82) (which can be recovered from [234, Theorem III.4]).

As another example, consider the function $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$ for $\alpha \in (0, 2] \backslash \{1\}$, which defines the Hellinger divergence of order α (see subsection 2.2.1). It is straightforward to verify that this function satisfies the conditions of Theorem 2.2 and (2.84), cf. [101, Corollary 6], [234, Section III-B, p.3362]. In this case, our bound in Theorem 2.2 is tighter than (2.84) for all Hellinger divergences of order α satisfying (2.87), i.e. $2f''(1)\, p_\star = 2\alpha p_\star \geq \alpha (1/p_\star)^{\alpha-2} = f''(1/p_\star) \Leftrightarrow p_\star^{\alpha-1} \geq \frac{1}{2}$, or equivalently, $0 < \alpha \leq 1 + (\log(2)/\log(1/p_\star))$ (where $\alpha = 1$ corresponds to KL divergence—see subsection 2.2.1).

■ 2.4.3 Ergodicity of Markov Chains

In this subsection, we derive a corollary of Corollary 2.1 that illustrates one use of upper bounds on contraction coefficients of source-channel pairs via $\eta_{\chi^2}(P_X, P_{Y|X})$.

Consider a Markov kernel $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ on a state space $\mathcal{X}$ that defines an *irreducible* and *aperiodic* (time homogeneous) discrete-time Markov chain with unique stationary pmf (or invariant measure) $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ such that $P_X W = P_X$, cf. [170, Section 1.3].[33] For simplicity, suppose further that W is *reversible* (i.e. the detailed balance equations, $P_X(x)[W]_{x,y} = P_X(y)[W]_{y,x}$ for all $x, y \in \mathcal{X}$, hold [170, Section 1.6]). This means that W is self-adjoint with respect to the weighted inner product defined by P_X, and has all real eigenvalues $1 = \lambda_1(W) > \lambda_2(W) \geq \cdots \geq \lambda_{|\mathcal{X}|}(W) > -1$. Let $\mu(W) \triangleq \max\{|\lambda_2(W)|, |\lambda_{|\mathcal{X}|}(W)|\} \in [0, 1)$ denote the SLEM of W (see subsection 2.2.3).

Since this Markov chain is *ergodic*, $\lim_{n\to\infty} R_X W^n = P_X$ for all $R_X \in \mathcal{P}_{\mathcal{X}}$ [170, Theorem 4.9]. This implies that $\lim_{n\to\infty} D(R_X W^n || P_X) = 0$ by the continuity of KL divergence [230, Proposition 3.1]. Let us estimate the rate at which this "distance to stationarity" (measured by KL divergence) vanishes. A naïve approach is to apply the SDPI (2.27) for KL divergence recursively to get:

$$D(R_X W^n || P_X) \leq \eta_{\mathsf{KL}}(P_X, W)^n D(R_X || P_X) \tag{2.88}$$

for all $R_X \in \mathcal{P}_{\mathcal{X}}$ and all epochs $n \in \mathbb{N}$. Using (2.29), this implies that:

$$\limsup_{n\to\infty} \eta_{\mathsf{KL}}(P_X, W^n)^{\frac{1}{n}} \leq \eta_{\mathsf{KL}}(P_X, W) \tag{2.89}$$

which turns out to be a loose bound on the rate in general.

When n is large, since $R_X W^n$ is close to P_X, we intuitively expect $D(R_X W^n || P_X)$ to resemble a χ^2-divergence (see (2.24) in subsection 2.2.1), which suggests that the contraction coefficient $\eta_{\mathsf{KL}}(P_X, W^n)$ should scale like $\eta_{\chi^2}(P_X, W)^n$. This intuition is rigorously executed in [49, Section 6]. Indeed, when $\mu(W)$ is strictly greater than the third largest eigenvalue modulus of W, a consequence of [49, Corollary 6.2] is:

$$\lim_{n\to\infty} \frac{D(R_X W^n || P_X)}{D(R_X W^{n-1} || P_X)} \leq \mu(W)^2 \tag{2.90}$$

for all $R_X \in \mathcal{P}_{\mathcal{X}}$ such that the denominator is always positive. (This limit is either 0 or $\mu(W)^2$.) Hence, after employing a Cesàro convergence argument and telescoping, we get:

$$\lim_{n\to\infty} \frac{D(R_X W^n || P_X)}{D(R_X W^{n-1} || P_X)} = \lim_{n\to\infty} \left(\frac{D(R_X W^n || P_X)}{D(R_X || P_X)} \right)^{\frac{1}{n}} \leq \mu(W)^2 \tag{2.91}$$

which suggests that $\limsup_{n\to\infty} \eta_{\mathsf{KL}}(P_X, W^n)^{\frac{1}{n}} \leq \mu(W)^2$. The next result proves that this inequality is in fact tight.

Corollary 2.2 (Rate of Convergence). *For every irreducible, aperiodic, and reversible Markov chain with transition kernel $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ and stationary pmf $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, we have:*

$$\lim_{n\to\infty} \eta_{\mathsf{KL}}(P_X, W^n)^{\frac{1}{n}} = \eta_{\chi^2}(P_X, W) = \mu(W)^2 .$$

[33] The matrix W is *primitive* since the chain is irreducible and aperiodic.

Proof. Since W is reversible and P_X is its stationary pmf, the corresponding DTM $B = \mathsf{diag}(\sqrt{P_X})\, W \mathsf{diag}(\sqrt{P_X})^{-1}$ is symmetric and similar to W (see definition (2.38)). Hence, W and B share the same eigenvalues, and $\mu(W)$ is the second largest singular value of B. Using Proposition 2.2 and (2.37), we have $\eta_{\chi^2}(P_X, W) = \mu(W)^2$, which proves the second equality.

Likewise, $\eta_{\chi^2}(P_X, W^n) = \mu(W^n)^2$ since W^n is reversible for any $n \in \mathbb{N}$. This yields:

$$\eta_{\chi^2}(P_X, W^n) = \mu(W^n)^2 = \mu(W)^{2n} = \eta_{\chi^2}(P_X, W)^n \tag{2.92}$$

where the second equality holds because eigenvalues of W^n are nth powers of eigenvalues of W. Using (2.92), part 7 of Proposition 2.3, and Corollary 2.1, we get:

$$\eta_{\chi^2}(P_X, W)^n \le \eta_{\mathsf{KL}}(P_X, W^n) \le \frac{\eta_{\chi^2}(P_X, W)^n}{\min\limits_{x \in \mathcal{X}} P_X(x)}\,.$$

Taking nth roots and letting $n \to \infty$ yields the desired result. ■

Corollary 2.2 portrays the well-understood phenomenon that $D(R_X W^n || P_X)$ vanishes with rate determined by $\mu(W)^2 = \eta_{\chi^2}(P_X, W)$. More generally, it illustrates that the bounds in Corollary 2.1 and Theorems 2.2 and 2.3 are useful in the regime where the random variables X and Y are weakly dependent (e.g. X is the initial state of an ergodic reversible Markov chain, and Y is the state after a large number of time steps). In this regime, these bounds are quite tight, and beat the η_{TV} bound in part 7 of Proposition 2.5.

■ 2.4.4 Tensorization of Bounds between Contraction Coefficients

In the absence of weak dependence, the upper bounds in Corollary 2.1 and Theorems 2.2 and 2.3 can be loose. In fact, they can be rendered arbitrarily loose because the constants in these bounds do not tensorize, while contraction coefficients do (as shown in part 5 of Proposition 2.3). For instance, if we are given $P_{X,Y}$ with $X \sim \mathsf{Bernoulli}(\frac{1}{2})$, then the constant in the upper bound of Corollary 2.1 is $1/\min_{x \in \{0,1\}} P_X(x) = 2$. If we instead consider a sequence of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are i.i.d. according to $P_{X,Y}$, then the constant in the upper bound of Corollary 2.1 is $1/\min_{x_1^n \in \{0,1\}^n} P_{X_1^n}(x_1^n) = 2^n$. However, since $\eta_{\mathsf{KL}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \eta_{\mathsf{KL}}(P_X, P_{Y|X})$ and $\eta_{\chi^2}(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \eta_{\chi^2}(P_X, P_{Y|X})$ by the tensorization property in part 5 of Proposition 2.3, the constant 2^n becomes arbitrarily loose as n grows. The next corollary presents a partial remedy for this i.i.d. slackening attack for Corollary 2.1.

Corollary 2.3 (Tensorized KL Contraction Coefficient Bound). *If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with joint pmf $P_{X,Y}$ such that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, then:*

$$\eta_{\mathsf{KL}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) \le \frac{\eta_{\chi^2}(P_{X_1^n}, P_{Y_1^n|X_1^n})}{\min\limits_{x \in \mathcal{X}} P_X(x)}\,.$$

Proof. This follows trivially from Corollary 2.1 and the tensorization property in part 5 of Proposition 2.3. ■

In the product distribution context, this corollary permits us to use the tighter factor $1/\min_{x\in\mathcal{X}} P_X(x)$ in the upper bound of Corollary 2.1 instead of $1/\min_{x_1^n\in\mathcal{X}^n} P_{X_1^n}(x_1^n) = (1/\min_{x\in\mathcal{X}} P_X(x))^n$. As shown in the ensuing corollaries, similar adjustments can be made for the constants in Theorems 2.2 and 2.3 in this context as well.

Corollary 2.4 (Tensorized Contraction Coefficient Bound). *If $(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. with joint pmf $P_{X,Y}$ such that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, and the conditions of Theorem 2.2 are satisfied, then:*

$$\eta_f(P_{X_1^n}, P_{Y_1^n|X_1^n}) \leq \frac{f'(1)+f(0)}{f''(1)\min\limits_{x\in\mathcal{X}} P_X(x)}\,\eta_{\chi^2}(P_{X_1^n}, P_{Y_1^n|X_1^n})\,.$$

Proof. This follows trivially from Theorem 2.2 and the tensorization property in part 5 of Proposition 2.3. ■

Corollary 2.5 (Tensorized Refined KL Contraction Coefficient Bound). *If $(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. with joint pmf $P_{X,Y}$ such that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, then:*

$$\eta_{\mathsf{KL}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) \leq \frac{2\,\eta_{\chi^2}(P_{X_1^n}, P_{Y_1^n|X_1^n})}{\phi\Big(\max\limits_{A\subseteq\mathcal{X}} \pi(A)\Big)\min\limits_{x\in\mathcal{X}} P_X(x)}$$

where $\pi(\cdot)$ and $\phi:\left[0,\frac{1}{2}\right]\to\mathbb{R}$ are defined in Lemma 2.1.

Proof. This follows trivially from Theorem 2.3 and the tensorization property in part 5 of Proposition 2.3. ■

Thus, tensorization can improve the upper bounds in Corollary 2.1 and Theorems 2.2 and 2.3.

■ 2.5 Proof of Equivalence between Gaussian Contraction Coefficients

Finally, we prove Theorem 2.4 in this section. To this end, recall from subsection 2.3.3 that we are given the jointly Gaussian pdf $P_{X,Y}$ defined via (2.69), with source pdf $P_X = \mathcal{N}(0,\sigma_X^2)$ and channel conditional pdfs $\{P_{Y|X=x} = \mathcal{N}(x,\sigma_W^2) : x\in\mathbb{R}\}$ such that $\sigma_X^2,\sigma_W^2 > 0$. Let $\mathcal{T}$ be the set of all pdfs with bounded support. Thus, a pdf $R_X \in \mathcal{T}$ if and only if there exists (finite) $C>0$ such that:

$$R_X(x) = R_X(x)\,\mathbb{1}\{x\in[-C,C]\} \tag{2.93}$$

almost everywhere with respect to the Lebesgue measure on $\mathbb{R}$. We first derive the following useful lemma.

Lemma 2.7 (Bounded Support Characterization of η_{KL}). *The supremum in (2.70) can be restricted to pdfs in $\mathcal{T}$:*

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{T}: \\ D(R_X||P_X)<+\infty}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where $R_Y = R_X * P_W$ for each R_X, and $P_W = \mathcal{N}(0, \sigma_W^2)$.

Proof. Consider any pdf R_X such that $0 < D(R_X||P_X) < +\infty$, and define a corresponding the sequence of pdfs $R_X^{(n)} \in \mathcal{T}$ via:

$$\forall x \in \mathbb{R}, \;\; R_X^{(n)}(x) = \frac{1}{C_n} R_X(x)\, \mathbb{1}\{x \in [-n, n]\}$$

where $C_n = \mathbb{E}_{R_X}[\mathbb{1}\{X \in [-n, n]\}]$, the indices $n \in \mathbb{N}$ are sufficiently large so that $C_n > 0$, and $\lim_{n\to\infty} C_n = 1$. Observe that:

$$D(R_X^{(n)}||P_X) = \frac{1}{C_n}\mathbb{E}_{R_X}\left[\mathbb{1}\{X \in [-n, n]\} \log\left(\frac{R_X(X)}{P_X(X)}\right)\right] - \log(C_n)\,.$$

Clearly, we have:

1. $\lim\limits_{n\to\infty} \mathbb{1}\{x \in [-n, n]\} \log\left(\frac{R_X(x)}{P_X(x)}\right) = \log\left(\frac{R_X(x)}{P_X(x)}\right)$ pointwise R_X-$a.s.$
2. $\left|\mathbb{1}\{x \in [-n, n]\} \log\left(\frac{R_X(x)}{P_X(x)}\right)\right| \le \left|\log\left(\frac{R_X(x)}{P_X(x)}\right)\right|$ pointwise R_X-$a.s.$
3. $\mathbb{E}_{R_X}\left[\left|\log\left(\frac{R_X(X)}{P_X(X)}\right)\right|\right] < +\infty$

where the finiteness follows from $D(R_X||P_X) < +\infty$. Hence, the dominated convergence theorem (DCT) yields:

$$\lim_{n\to\infty} D(R_X^{(n)}||P_X) = D(R_X||P_X)\,. \tag{2.94}$$

Furthermore, let $R_Y^{(n)} = R_X^{(n)} * P_W$ so that for every $y \in \mathbb{R}$:

$$R_Y(y) - C_n R_Y^{(n)}(y) = \mathbb{E}_{R_X}[\mathbb{1}\{X \in \mathbb{R}\backslash[-n, n]\} P_W(y - X)]\,.$$

Since for all $x, y \in \mathbb{R}$, we have:

1. $\lim\limits_{n\to\infty} \mathbb{1}\{x \in \mathbb{R}\backslash[-n, n]\} P_W(y - x) = 0$
2. $0 \le \mathbb{1}\{x \in \mathbb{R}\backslash[-n, n]\} P_W(y - x) \le P_W(y - x)$
3. $\mathbb{E}_{R_X}[P_W(y - X)] = R_Y(y) < +\infty$

applying the DCT shows the pointwise convergence of the pdfs $\{R_Y^{(n)}\}$:

$$\forall y \in \mathbb{R}, \quad \lim_{n\to\infty} C_n R_Y^{(n)}(y) = \lim_{n\to\infty} R_Y^{(n)}(y) = R_Y(y)\,.$$

This implies that $R_Y^{(n)}$ converges weakly to R_Y as $n \to \infty$ by Scheffé's lemma. Hence, by the weak lower semi-continuity of KL divergence [230, Theorem 3.6, Section 3.5]:

$$\liminf_{n\to\infty} D(R_Y^{(n)}||P_Y) \geq D(R_Y||P_Y)\,. \tag{2.95}$$

Combining (2.94) and (2.95), we get:

$$\liminf_{n\to\infty} \frac{D(R_Y^{(n)}||P_Y)}{D(R_X^{(n)}||P_X)} \geq \frac{D(R_Y||P_Y)}{D(R_X||P_X)}\,. \tag{2.96}$$

To complete the proof, we use a "diagonalization argument." Suppose $\{R_{X,m} : m \in \mathbb{N}\}$ is a sequence of pdfs that satisfies $0 < D(R_{X,m}||P_X) < +\infty$ for all $m \in \mathbb{N}$ and achieves the supremum in (2.70):

$$\lim_{m\to\infty} \frac{D(R_{Y,m}||P_Y)}{D(R_{X,m}||P_X)} = \eta_{\mathsf{KL}}(P_X, P_{Y|X})$$

where $R_{Y,m} = R_{X,m} * P_W$. Then, since (2.96) is true, we can construct a sequence $\{R_{X,m}^{(n(m))} \in \mathcal{T} : m \in \mathbb{N}\}$, where each $n(m)$ is chosen such that for every $m \in \mathbb{N}$:

$$\frac{D(R_{Y,m}^{(n(m))}||P_Y)}{D(R_{X,m}^{(n(m))}||P_X)} \geq \frac{D(R_{Y,m}||P_Y)}{D(R_{X,m}||P_X)} - \frac{1}{2^m}$$

where $R_{Y,m}^{(n(m))} = R_{X,m}^{(n(m))} * P_W$. Letting $m \to \infty$, we have:

$$\liminf_{m\to\infty} \frac{D(R_{Y,m}^{(n(m))}||P_Y)}{D(R_{X,m}^{(n(m))}||P_X)} \geq \eta_{\mathsf{KL}}(P_X, P_{Y|X})\,.$$

Since the supremum in (2.70) is over all pdfs (which certainly includes all pdfs in $\mathcal{T}$), this inequality is actually an equality. This completes the proof. (Also note that for any $R_X \in \mathcal{T}$, the constraint $D(R_X||P_X) > 0$ is automatically true since $P_X = \mathcal{N}(0, \sigma_X^2)$. So, the supremum in the lemma statement does not include this constraint.) ■

We next prove Theorem 2.4 using Lemma 2.7, which ensures that all differential entropy terms in the ensuing argument are well-defined and finite.

Proof of Theorem 2.4. First note that:

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\mathsf{max}}(X;Y)^2 = \frac{\mathbb{COV}(X,Y)^2}{\mathbb{VAR}(X)\mathbb{VAR}(Y)} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}$$

where the first equality is precisely (2.37) (which holds for general random variables [242]), the second equality follows from to Rényi's seventh axiom that $\rho_{\mathsf{max}}(X;Y)$ is the absolute value of the Pearson correlation coefficient of jointly Gaussian X and Y [236], and the final equality follows from direct computation.

We next prove that for any $p \geq \sigma_X^2$:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \geq \eta_{\mathsf{KL}}^{(p)}(P_X, P_{Y|X}) \geq \eta_{\chi^2}(P_X, P_{Y|X})\,.$$

The first inequality is obvious from (2.70) and (2.72). For the second inequality, let $R_X = \mathcal{N}(\sqrt{\delta}, \sigma_X^2 - \delta)$ and $R_Y = R_X * P_W = \mathcal{N}(\sqrt{\delta}, \sigma_X^2 + \sigma_W^2 - \delta)$ for any $\delta > 0$. Then, we get:

$$\lim_{\delta \to 0^+} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \lim_{\delta \to 0^+} \frac{\log\left(\frac{\sigma_X^2+\sigma_W^2}{\sigma_X^2+\sigma_W^2-\delta}\right)}{\log\left(\frac{\sigma_X^2}{\sigma_X^2-\delta}\right)} = \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}$$

where the second equality follows from l'Hôpital's rule. Since $\mathbb{E}_{R_X}[X^2] = \sigma_X^2$ for every $\delta > 0$, we have:

$$\eta_{\mathsf{KL}}^{(p)}(P_X, P_{Y|X}) \geq \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}$$

for any $p \geq \sigma_X^2$.

Therefore, it suffices to prove that:

$$\eta_{\mathsf{KL}}(P_X, P_{Y|X}) \leq \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}\,.$$

Using Lemma 2.7, we can equivalently show that:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2} \tag{2.97}$$

for every pdf $R_X \in \mathcal{T}$ with $D(R_X||P_X) < +\infty$.

For any pdf R_X, we define the *differential entropy* of R_X as:

$$h(R_X) \triangleq -\mathbb{E}_{R_X}[\log(R_X(X))]\,. \tag{2.98}$$

To check that such differential entropy terms are well-defined and finite for $R_X \in \mathcal{T}$, we employ the argument in [15, Lemma 8.3.1, Theorem 8.3.3]. Observe that for all $x \in \operatorname{ess\,supp}(R_X)$:

$$\log(R_X(x)) = \log\left(\frac{R_X(x)}{P_X(x)}\right) - \frac{1}{2}\log\left(2\pi\sigma_X^2\right) - \frac{x^2}{2\sigma_X^2}\,.$$

Since $D(R_X||P_X)$ must be finite in (2.70) and $X^2 \geq 0$, we can take expectations with respect to R_X to get:

$$-h(R_X) = D(R_X||P_X) - \frac{1}{2}\log\left(2\pi\sigma_X^2\right) - \frac{\mathbb{E}_{R_X}[X^2]}{2\sigma_X^2} \tag{2.99}$$

which shows that $h(R_X)$ always exists, $h(R_X)$ is finite when $\mathbb{E}_{R_X}[X^2] < +\infty$, and $h(R_X) = +\infty$ when $\mathbb{E}_{R_X}[X^2] = +\infty$. Furthermore, if the pdf $R_X \in \mathcal{T}$ has bounded support, $\mathbb{E}_{R_X}[X^2] < +\infty$ and $h(R_X)$ is well-defined and finite.

Let $R_X \in \mathcal{T}$ and $R_Y = R_X * P_W$ have second moments $\mathbb{E}_{R_X}[X^2] = \sigma_X^2 + q > 0$ and $\mathbb{E}_{R_Y}[Y^2] = \sigma_X^2 + \sigma_W^2 + q > 0$ for some $q > -\sigma_X^2$. Using (2.99), we have:

$$\begin{aligned} D(R_X||P_X) &= \frac{1}{2}\log\left(2\pi\sigma_X^2\right) + \frac{\sigma_X^2 + q}{2\sigma_X^2} - h(R_X) \\ &= h(P_X) - h(R_X) + \frac{q}{2\sigma_X^2}\,, \\ D(R_Y||P_Y) &= h(P_Y) - h(R_Y) + \frac{q}{2\left(\sigma_X^2 + \sigma_W^2\right)} \end{aligned}$$

where $h(R_Y)$ exists and is finite because $\mathbb{E}_{R_Y}[Y^2]$ is finite (as argued earlier using (2.99)). Hence, it suffices to prove that:

$$h(P_Y) - h(R_Y) \le \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(h(P_X) - h(R_X)\right) \tag{2.100}$$

which is equivalent to (2.97). We can recast (2.100) as:

$$\begin{aligned} \exp(2h(P_Y) - 2h(R_Y))^{\sigma_X^2+\sigma_W^2} &\le \exp(2h(P_X) - 2h(R_X))^{\sigma_X^2} \\ \left(\frac{\frac{1}{2\pi e}\exp(2h(P_Y))}{\frac{1}{2\pi e}\exp(2h(R_Y))}\right)^{\sigma_X^2+\sigma_W^2} &\le \left(\frac{\frac{1}{2\pi e}\exp(2h(P_X))}{\frac{1}{2\pi e}\exp(2h(R_X))}\right)^{\sigma_X^2} \\ \left(\frac{N(P_Y)}{N(R_Y)}\right)^{\sigma_X^2+\sigma_W^2} &\le \left(\frac{N(P_X)}{N(R_X)}\right)^{\sigma_X^2} \end{aligned}$$

where for any pdf Q_X such that $h(Q_X)$ exists, we define the *entropy power* of Q_X as, cf. [64, Section III-A]:

$$N(Q_X) \triangleq \frac{\exp(2h(Q_X))}{2\pi e}\,. \tag{2.101}$$

For $P_X = \mathcal{N}(0, \sigma_X^2)$, $P_W = \mathcal{N}(0, \sigma_W^2)$, and $P_Y = P_X * P_W = \mathcal{N}(0, \sigma_X^2 + \sigma_W^2)$, the entropy powers are $N(P_X) = \sigma_X^2$, $N(P_W) = \sigma_W^2$, and $N(P_Y) = \sigma_X^2 + \sigma_W^2$, respectively. Applying the *entropy power inequality* to R_X, P_W, and $R_Y = R_X * P_W$ [64, Theorem 4], we have:

$$N(R_Y) \ge N(R_X) + N(P_W) = N(R_X) + \sigma_W^2\,. \tag{2.102}$$

Hence, it is sufficient to prove that:

$$\left(\frac{\sigma_X^2 + \sigma_W^2}{N(R_X) + \sigma_W^2}\right)^{\sigma_X^2+\sigma_W^2} \le \left(\frac{\sigma_X^2}{N(R_X)}\right)^{\sigma_X^2}\,.$$

Let $a = \sigma_X^2 + \sigma_W^2$, $b = \sigma_X^2 - N(R_X)$, and $c = \sigma_X^2$. Then, we have $a > c > 0$ and $c > b$ (which follows from the finiteness of $h(R_X)$), and it is sufficient to prove that:

$$\left(\frac{a}{a-b}\right)^a \leq \left(\frac{c}{c-b}\right)^c$$

which is equivalent to proving:

$$a > c > 0 \text{ and } c > b \quad \Rightarrow \quad \left(1 - \frac{b}{c}\right)^c \leq \left(1 - \frac{b}{a}\right)^a.$$

This statement is a variant of Bernoulli's inequality proved in [172, Theorem 3.1, parts (r_7') and (r_7'')]. This completes the proof. ■

■ 2.6 Conclusion and Future Directions

In closing this chapter, we briefly recapitulate our main contributions and then propose some directions for future research. We first illustrated in Theorem 2.1 that if the optimization problem defining $\eta_f(P_X, P_{Y|X})$ is subjected to an additional "local approximation" constraint that forces the input f-divergence to vanish, then the resulting optimum value is $\eta_{\chi^2}(P_X, P_{Y|X})$. This transparently captures the intuition behind the maximal correlation lower bound in part 7 of Proposition 2.3. We then derived a linear upper bound on $\eta_f(P_X, P_{Y|X})$ in terms of $\eta_{\chi^2}(P_X, P_{Y|X})$ for a class of f-divergences in Theorem 2.2, and improved this bound for the salient special case of $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ in Theorem 2.3. Such bounds are useful in weak dependence regimes such as in the analysis of ergodicity of Markov chains (as shown in Corollary 2.2). Finally, in the spirit of comparing contraction coefficients of source-channel pairs, we also gave an alternative proof of the equivalence, $\eta_{\mathsf{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$, for jointly Gaussian distributions $P_{X,Y}$ defined by AWGN channels in Theorem 2.4 and section 2.5. This proof showed that adding a large enough power constraint to the extremization in $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ does not change its value.

As discussed in subsection 2.4.4, the constants in the linear bounds in Theorems 2.2 and 2.3 vary "blindly" with the dimension of a product distribution. While results like Corollary 2.3, 2.4, and 2.5 partially remedy this tensorization issue, one compelling direction of future work is to discover linear bounds whose constants gracefully tensorize. Another, perhaps more concrete, avenue of future work is to derive the optimal distribution dependent refinement of Lemma 2.4 (as suggested in [101, Remark, p.5380]). Such a refinement could be used to tighten Theorem 2.2 so that it specializes to Theorem 2.3 instead of Corollary 2.1. However, such a refinement cannot circumvent the more critical tensorization issue that ails these bounds.

■ 2.7 Bibliographical Notes

The results in chapter 2 are refinements, generalizations, and more rigorous variants of the results in the master's thesis [180, Chapter 3]. In particular, both chapter 2 and

appendix A are based on the manuscript [192]. This manuscript was published in part at the Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing 2015 [189].

Chapter 3

Extension using Comparison of Channels

FOR any Markov chain $U \to X \to Y$, it is well-known that the data processing inequality holds:

$$I(U;Y) \leq I(U;X)\,. \tag{3.1}$$

As discussed in chapter 2, this result can be strengthened to the strong data processing inequality [5]:

$$I(U;Y) \leq \eta_{\mathsf{KL}}(P_{Y|X})\, I(U;X) \tag{3.2}$$

where the contraction coefficient $\eta_{\mathsf{KL}}(P_{Y|X}) \in [0,1]$ only depends on the channel $P_{Y|X}$, and (3.2) holds for all joint distributions $P_{U,X}$. Frequently, one obtains $\eta_{\mathsf{KL}}(P_{Y|X}) < 1$ so that the resulting inequality is a strict improvement over the DPI (3.1). SDPIs have been recently simultaneously rediscovered and applied in several disciplines; see section 2.2 and [231, Section 2] for short surveys. In [231, Section 6], it was noticed that the validity of (3.2) for all $P_{U,X}$ is equivalent to the statement that an erasure channel with erasure probability $1 - \eta_{\mathsf{KL}}(P_{Y|X})$ is *less noisy* than the given channel $P_{Y|X}$. In this way, the entire field of SDPIs is equivalent to determining whether a given channel is dominated by an erasure channel. (Note that throughout this chapter, we only consider SDPIs and contraction coefficients of channels, not source-channel pairs.)

This chapter initiates the study of a natural extension of the concept of SDPI by replacing the distinguished role played by erasure channels with q-ary symmetric channels. We give simple criteria for testing this type of domination and explain how the latter can be used to prove logarithmic Sobolev inequalities. In the process, we also prove equivalent characterizations of the less noisy preorder over channels using non-linear operator convex f-divergences by generalizing the main result of [46] (see Proposition 2.6 in chapter 2). In the next section, we introduce some basic definitions and background. We state and motivate our main question in section 3.2, and then present our main results and delineate the remainder of our discussion in section 3.3.

3.1 Background

We will require background from two aspects of information theory in this chapter: channel comparison, and symmetric and additive noise channels. The ensuing subsection surveys the former topic, and the subsequent subsection presents the latter.

3.1.1 Channel Preorders in Information Theory

Since we will study preorders over discrete channels that capture various notions of relative "noisiness" between channels, we provide an overview of some well-known channel preorders in the literature. Consider an input random variable $X \in \mathcal{X}$ and an output random variable $Y \in \mathcal{Y}$, where the alphabets are $\mathcal{X} = [q] \triangleq \{0, \ldots, q-1\}$ and $\mathcal{Y} = [r]$ for $q, r \in \mathbb{N}$ without loss of generality. We let $\mathcal{P}_q = \mathcal{P}_{\mathcal{X}}$ be the set of all pmfs of X, where every pmf $P_X = (P_X(0), \ldots, P_X(q-1)) \in \mathcal{P}_q$ and is perceived as a row vector. Likewise, we let $\mathcal{P}_r = \mathcal{P}_{\mathcal{Y}}$ be the set of all pmfs of Y. A *channel* is the set of conditional distributions $W_{Y|X}$ that associates each $x \in \mathcal{X}$ with a conditional pmf $W_{Y|X=x} \in \mathcal{P}_r$. So, we represent each channel with a stochastic matrix $W \in \mathbb{R}_{\mathsf{sto}}^{q \times r} = \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ that is defined entry-wise as:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; [W]_{x+1,y+1} \triangleq W_{Y|X}(y|x) \tag{3.3}$$

where the $(x+1)$th row of W corresponds to the conditional pmf $W_{Y|X=x} \in \mathcal{P}_r$, and each column of W has at least one non-zero entry so that no output alphabet letters are redundant. Moreover, we think of such a channel as a (linear) map $W : \mathcal{P}_q \to \mathcal{P}_r$ that takes any row probability vector $P_X \in \mathcal{P}_q$ to the row probability vector $P_Y = P_X W \in \mathcal{P}_r$. Note that in this chapter, we use the notation $\mathcal{P}_q$, $\mathcal{P}_r$, and $\mathbb{R}_{\mathsf{sto}}^{q \times r}$ instead of the notation $\mathcal{P}_{\mathcal{X}}$, $\mathcal{P}_{\mathcal{Y}}$, and $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, respectively (which was introduced in chapter 2), because the material in this chapter benefits from a matrix theoretic perspective, and our notation makes the dimensions of various quantities easily readable.

One of the earliest preorders over channels was the notion of *channel inclusion* proposed by Shannon in [251].[34] Given two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q \times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{s \times t}$ for some $q, r, s, t \in \mathbb{N}$, he stated that W includes V, denoted $W \succeq_{\mathsf{inc}} V$, if there exist a pmf $g \in \mathcal{P}_m$ for some $m \in \mathbb{N}$, and two sets of channels $\{A_k \in \mathbb{R}_{\mathsf{sto}}^{r \times t} : k = 1, \ldots, m\}$ and $\{B_k \in \mathbb{R}_{\mathsf{sto}}^{s \times q} : k = 1, \ldots, m\}$, such that:

$$V = \sum_{k=1}^{m} g_k B_k W A_k \,. \tag{3.4}$$

Channel inclusion is preserved under channel addition and multiplication (which are defined in [250]), and the existence of a code for V implies the existence of as good

[34] Throughout this thesis, we will refer to various information theoretic orders over channels as *preorders* rather than *partial orders* (although the latter is more standard terminology in the literature). This is because we will think of channels as individual stochastic matrices rather than equivalence classes of stochastic matrices (e.g. identifying all stochastic matrices with permuted columns), and as a result, the anti-symmetric property will not hold.

a code for W in a probability of error sense [251]. The channel inclusion preorder includes the *input-output degradation* preorder, which can be found in [50], as a special case. Indeed, V is an input-output degraded version of W, denoted $W \succeq_{\mathsf{iod}} V$, if there exist channels $A \in \mathbb{R}_{\mathsf{sto}}^{r\times t}$ and $B \in \mathbb{R}_{\mathsf{sto}}^{s\times q}$ such that $V = BWA$. We will study an even more specialized case of Shannon's channel inclusion known as *degradation*, which first appeared in the information theory literature in the study of broadcast channels [25,52].

Definition 3.1 (Degradation Preorder). *A channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ is said to be a* degraded *version of a channel $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ with the same input alphabet, denoted $W \succeq_{\mathsf{deg}} V$, if $V = WA$ for some channel $A \in \mathbb{R}_{\mathsf{sto}}^{r\times s}$.*

The degradation preorder has a long history. Its study actually originated in the statistics literature [30, 252, 260], where it is also known as the *Blackwell order*. In a statistical decision theoretic context, the channels W and V can be construed as observation models (or statistical experiments) of the parameter space $\mathcal{X} = [q]$. For any model $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$, any prior distribution $P_X \in \mathcal{P}_q$, and any loss function $L : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (which can be encoded as a matrix in $\mathbb{R}^{q\times q}$), we define the *Bayes risk* as:

$$R(W, P_X, L) \triangleq \inf_{d(\cdot)} \mathbb{E}[L(X, d(Y))] \tag{3.5}$$

where the infimum is over all randomized decision rules $d : [r] \to [q]$ for X based on Y (which could be viewed as Markov kernels), and the expectation is over the joint pmf of (X, Y) defined by (P_X, W) and the randomness of d. According to the Blackwell order, the model W is said to be *more informative* than the model V (with the same input alphabet) if for every prior pmf $P_X \in \mathcal{P}_q$, and every loss function $L : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $R(W, P_X, L) \leq R(V, P_X, L)$. This definition is quite natural from a decision theoretic perspective. Somewhat surprisingly, the celebrated *Blackwell-Sherman-Stein theorem* states that W is more informative than V if and only if V is a degraded version of W [30, 252, 260], i.e. the Blackwell order coincides with the degradation preorder. We refer readers to [169] for an elegant and simple proof of this result using the *separating hyperplane theorem*.

Furthermore, degradation has beautiful ties with non-Bayesian binary hypothesis testing as well. When $q = 2$, W and V can be construed as *dichotomies*, which refer to pairs of distributions corresponding to two hypotheses $\{X = 0\}$ and $\{X = 1\}$. For any dichotomy $W \in \mathbb{R}_{\mathsf{sto}}^{2\times r}$, we can define a concave non-decreasing *Neyman-Pearson function* (or *receiver operating characteristic curve*) $\beta_W : [0, 1] \to [0, 1]$, cf. [290, Chapters 3 and 4]:

$$\forall \alpha \in [0, 1], \quad \beta_W(\alpha) \triangleq \sup_{d(\cdot):\, \mathbb{P}(d(Y)=1|X=0)\leq\alpha} \mathbb{P}(d(Y) = 1|X = 1) \tag{3.6}$$

which maximizes the detection probability over all randomized decision rules $d : [r] \to \{0, 1\}$ subject to a constraint on the false-alarm probability, where the probabilities are calculated with respect to the conditional distribution W and the randomness of d. This function captures all the statistical information required to distinguish between

the two hypotheses under the model W—see [275, Example 1.4.3] for various fascinating properties Neyman-Pearson functions. Intuitively, the dichotomy W is statistically more informative than the dichotomy V if $\beta_W \geq \beta_V$ pointwise. It turns out that this notion also coincides with degradation, i.e. $\beta_W \geq \beta_V$ pointwise if and only if V is a degraded version of W [276, Theorem 5.3], [275, Section 9.3]. Furthermore, suppose each row of W belongs to $\mathcal{P}_r^\circ$ and each row of V belongs to $\mathcal{P}_s^\circ$. Then, it can be shown that V is a degraded version of W if and only if for all convex functions $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the following f-divergence inequality holds [276, Theorem 5.3], [275, Section 9.3]:[35]

$$D_f(W_{Y|X=1}||W_{Y|X=0}) \geq D_f(V_{Y|X=1}||V_{Y|X=0})\,. \tag{3.7}$$

Equivalently, when W and V are entry-wise strictly positive, V is a degraded version of W if and only if for all convex functions $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, we have:

$$\forall P_X, Q_X \in \mathcal{P}_2,\;\; D_f(P_X W||Q_X W) \geq D_f(P_X V||Q_X V)\,. \tag{3.8}$$

We refer readers to the comprehensive treatise [275] for more details on comparison of statistical experiments.

Finally, we note that when Definition 3.1 of degradation is applied to general matrices (rather than stochastic matrices), it is equivalent to Definition C.8 of *matrix majorization* in [195, Chapter 15] (which has been studied by Dahl in [62] and [61]). Many other generalizations of the majorization preorder over vectors (briefly introduced in appendix B.1) that apply to matrices are also presented in [195, Chapter 15].

Körner and Marton defined two other preorders over channels in [156] known as the *more capable* and *less noisy* preorders. While the original definitions of these preorders explicitly reflect their significance in channel coding, we will define them using equivalent mutual information characterizations proved in [156]. (See [58, Problems 6.16-6.18] for more on the relationship between channel coding and some of the aforementioned preorders.) We say a channel $W \in \mathbb{R}_{\mathsf{sto}}^{q \times r}$ is *more capable* than a channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times s}$ with the same input alphabet, denoted $W \succeq_{\mathsf{mc}} V$, if $I(P_X, W_{Y|X}) \geq I(P_X, V_{Y|X})$ for every input pmf $P_X \in \mathcal{P}_q$, where $I(P_X, W_{Y|X})$ denotes the mutual information of the joint pmf defined by P_X and $W_{Y|X}$. The next definition presents the less noisy preorder, which will be a key player in our study.

Definition 3.2 (Less Noisy Preorder). *Given two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q \times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q \times s}$ with the same input alphabet, let Y_W and Y_V denote the output random variables of W and V, respectively. Then, W is* less noisy *than V, denoted $W \succeq_{\mathsf{ln}} V$, if $I(U; Y_W) \geq I(U; Y_V)$ for every joint distribution $P_{U,X}$, where the random variable $U \in \mathcal{U}$ has some arbitrary range $\mathcal{U}$, and $U \to X \to (Y_W, Y_V)$ forms a Markov chain.*

An analogous characterization of the less noisy preorder using KL divergence is given in the next proposition.

[35] For convenience, we will sometimes abuse notation and refer to the output random variables of both W and V as Y although they are different (see e.g. Proposition 3.2). To be precise, one should distinguish between these random variables as shown in Definition 3.2.

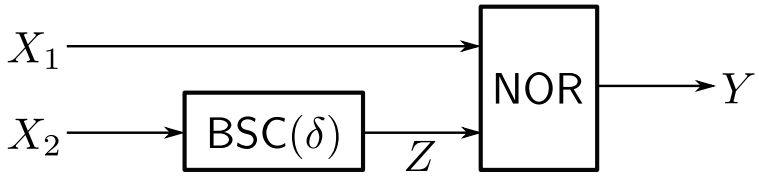


Figure 3.1. Illustration of a Bayesian network where $X_1, X_2, Z, Y \in \{0, 1\}$ are binary random variables, $P_{Z|X_2}$ is a $\mathsf{BSC}(\delta)$ with $\delta \in (0, 1)$, and $P_{Y|X_1,Z}$ is defined by a deterministic NOR gate.

Proposition 3.1 (KL Divergence Characterization of Less Noisy [156]). *Given two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ with the same input alphabet, $W \succeq_{\mathsf{ln}} V$ if and only if $D(P_X W||Q_X W) \geq D(P_X V||Q_X V)$ for every pair of input pmfs $P_X, Q_X \in \mathcal{P}_q$.*[36]

We will primarily use this KL divergence characterization of $\succeq_{\mathsf{ln}}$ in our discourse because of its simplicity. This characterization conveys that the pair of pmfs $P_X W$ and $Q_X W$ is always "more distinguishable" than the pair $P_X V$ and $Q_X V$, which indeed intuitively corresponds to W being "less noisy" than V. Another well-known equivalent characterization of $\succeq_{\mathsf{ln}}$ due to van Dijk is presented below, cf. [282, Theorem 2]. We will derive some useful corollaries from it later in subsection 3.6.3.

Proposition 3.2 (van Dijk Characterization of Less Noisy [282]). *Given two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ with the same input alphabet, consider the functional $F : \mathcal{P}_q \to \mathbb{R}$:*

$$\forall P_X \in \mathcal{P}_q, \;\; F(P_X) \triangleq I(P_X, W_{Y|X}) - I(P_X, V_{Y|X}) \, .$$

Then, $W \succeq_{\mathsf{ln}} V$ if and only if F is concave.

The more capable and less noisy preorders have both been used to study the capacity regions of broadcast channels. We refer readers to [80,97,216], and the references therein for further details. We also remark that the more capable and less noisy preorders tensorize, as shown in [58, Problem 6.18] and [231, Proposition 16], [268, Proposition 5] (also see Lemma B.3 in appendix B.4), respectively.

On the other hand, these preorders exhibit rather counter-intuitive behavior in the context of *Bayesian networks* (or directed graphical models). Consider a Bayesian network with "source" nodes X (with no inbound edges) and "sink" nodes Y (with no outbound edges). If we select a node Z in this network and replace the channel from the parents of Z to Z with a less noisy channel, then we may reasonably conjecture that the channel from X to Y also becomes less noisy (motivated by the results in [231]). However, this conjecture is false. To see this, consider the Bayesian network in Figure 3.1 (inspired by the results in [281]), where the source nodes are $X_1 \sim \mathsf{Bernoulli}(\frac{1}{2})$ and $X_2 = 1$ *a.s.*, the node Z is the output of a BSC with crossover probability $\delta \in (0, 1)$, denoted

[36] Throughout this thesis, we adhere to the convention that $\infty \geq \infty$ is true. So, $D(P_X W||Q_X W) \geq D(P_X V||Q_X V)$ is not violated when both KL divergences are infinity.

$\mathsf{BSC}(\delta)$, and the sink node Y is the output of a NOR gate. Let $I(\delta) = I(X_1, X_2; Y)$ be the end-to-end mutual information. Then, although $\mathsf{BSC}(0) \succeq_{\mathsf{ln}} \mathsf{BSC}(\delta)$ for $\delta \in (0, 1)$, it is easy to verify that $I(\delta) > I(0) = 0$. So, when we replace the $\mathsf{BSC}(\delta)$ with a less noisy $\mathsf{BSC}(0)$, the end-to-end channel does *not* become less noisy (or more capable).

The next proposition illustrates certain well-known relationships between the various preorders discussed in this subsection.

Proposition 3.3 (Relations between Channel Preorders). *Given two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q \times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q \times s}$ with the same input alphabet, we have:*

1. $W \succeq_{\mathsf{deg}} V \;\Rightarrow\; W \succeq_{\mathsf{iod}} V \;\Rightarrow\; W \succeq_{\mathsf{inc}} V$,
2. $W \succeq_{\mathsf{deg}} V \;\Rightarrow\; W \succeq_{\mathsf{ln}} V \;\Rightarrow\; W \succeq_{\mathsf{mc}} V$.

These observations follow in a straightforward manner from the definitions of the various preorders. Perhaps the only nontrivial implication is $W \succeq_{\mathsf{deg}} V \Rightarrow W \succeq_{\mathsf{ln}} V$, which can be proven using Proposition 3.1 and the data processing inequality.

■ 3.1.2 Symmetric Channels and Their Properties

We next formally define q-ary symmetric channels and convey some of their properties. To this end, we first introduce some properties of Abelian groups and define additive noise channels. Let us fix some $q \in \mathbb{N}$ with $q \geq 2$ and consider an *Abelian group* $(\mathcal{X}, \oplus)$ of order q equipped with a binary "addition" operation denoted by $\oplus$. Without loss of generality, we let $\mathcal{X} = [q]$, and let 0 denote the identity element. This endows an ordering to the elements of $\mathcal{X}$. Each element $x \in \mathcal{X}$ permutes the entries of the row vector $(0, \ldots, q-1)$ to $(\sigma_x(0), \ldots, \sigma_x(q-1))$ by (left) addition in the Cayley table of the group, where $\sigma_x : [q] \to [q]$ denotes a permutation of $[q]$, and $\sigma_x(y) = x \oplus y$ for every $y \in \mathcal{X}$. So, corresponding to each $x \in \mathcal{X}$, we can define a permutation matrix:

$$P_x \triangleq \left[e_{\sigma_x(0)+1} \; \cdots \; e_{\sigma_x(q-1)+1} \right] \in \mathbb{R}^{q \times q} \tag{3.9}$$

such that:

$$[v_0 \; \cdots \; v_{q-1}] \, P_x = \left[v_{\sigma_x(0)} \; \cdots \; v_{\sigma_x(q-1)} \right] \tag{3.10}$$

for any $v_0, \ldots, v_{q-1} \in \mathbb{R}$, where $e_i \in \mathbb{R}^q$ is the ith standard basis vector for each $i \in \{1, \ldots, q\}$. The permutation matrices $\{P_x \in \mathbb{R}^{q \times q} : x \in \mathcal{X}\}$ (with the matrix multiplication operation) form a group that is isomorphic to $(\mathcal{X}, \oplus)$ (see *Cayley's theorem*, and permutation and regular representations of groups in [13, Sections 6.11, 7.1, 10.6]). In particular, these matrices commute as $(\mathcal{X}, \oplus)$ is Abelian, and are jointly unitarily diagonalizable by a *Fourier matrix of characters* (using [129, Theorem 2.5.5]).[37] We

[37] We refer readers who have less familiarity with abstract algebra to [261, Chapter 7] for a concise introduction to Fourier analysis on finite Abelian groups.

now recall that given a row vector $x = (x_0, \ldots, x_{q-1}) \in (\mathbb{R}^q)^*$, we may define a corresponding $\mathcal{X}$-*circulant matrix*, $\mathsf{circ}_{\mathcal{X}}(x) \in \mathbb{R}^{q \times q}$, that is given entry-wise by [67, Chapter 3E, Section 4]:

$$\forall a, b \in [q], \;\; \left[\mathsf{circ}_{\mathcal{X}}(x)\right]_{a+1,b+1} \triangleq x_{(-a) \oplus b} \, . \tag{3.11}$$

where $-a \in \mathcal{X}$ denotes the inverse of $a \in \mathcal{X}$. Moreover, we can decompose this $\mathcal{X}$-circulant matrix as:

$$\mathsf{circ}_{\mathcal{X}}(x) = \sum_{i=0}^{q-1} x_i P_i^T \tag{3.12}$$

since we have:

$$\sum_{i=0}^{q-1} x_i \left[P_i^T\right]_{a+1,b+1} = \sum_{i=0}^{q-1} x_i \left[e_{\sigma_i(a)+1}\right]_{b+1} = x_{(-a) \oplus b} \tag{3.13}$$

for every $a, b \in [q]$. Using similar reasoning, we can write:

$$\mathsf{circ}_{\mathcal{X}}(x) = \left[P_0 \, y \, \cdots \, P_{q-1} \, y\right] = \left[P_0 \, x^T \, \cdots \, P_{q-1} \, x^T\right]^T \tag{3.14}$$

where $y = \left[x_0 \;\; x_{-1} \cdots x_{-(q-1)}\right]^T \in \mathbb{R}^q$, and $P_0 = I_q \in \mathbb{R}^{q \times q}$ is the $q \times q$ identity matrix. Using (3.12), we see that $\mathcal{X}$-circulant matrices are normal, form a commutative algebra, and are jointly unitarily diagonalizable by a Fourier matrix. Furthermore, given two row vectors $x, y \in (\mathbb{R}^q)^*$, we can define $x \, \mathsf{circ}_{\mathcal{X}}(y) = y \, \mathsf{circ}_{\mathcal{X}}(x)$ as the $\mathcal{X}$-*circular convolution* of x and y, where the commutativity of $\mathcal{X}$-circular convolution follows from the commutativity of $\mathcal{X}$-circulant matrices.

A salient specialization of this discussion is the case where $\oplus$ is addition modulo q, and $(\mathcal{X} = [q], \oplus)$ is the cyclic Abelian group $\mathbb{Z}/q\mathbb{Z}$. In this scenario, $\mathcal{X}$-circulant matrices correspond to the standard circulant matrices which are jointly unitarily diagonalized by the *discrete Fourier transform* (DFT) matrix.[38] Furthermore, for each $x \in [q]$, the permutation matrix $P_x^T = P_q^x$, where $P_q \in \mathbb{R}^{q \times q}$ is the *generator cyclic permutation matrix* as presented in [129, Section 0.9.6]:

$$\forall a, b \in [q], \;\; \left[P_q\right]_{a+1,b+1} \triangleq \mathbb{1}\{b - a \equiv 1 \pmod{q}\} \tag{3.15}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. The matrix P_q cyclically shifts any input row vector to the right once, i.e. $(x_1, x_2, \ldots, x_q) \, P_q = (x_q, x_1, \ldots, x_{q-1})$.

Let us now consider a channel with common input and output alphabet $\mathcal{X} = \mathcal{Y} = [q]$, where $(\mathcal{X}, \oplus)$ is an Abelian group. Such a channel operating on an Abelian group is called an *additive noise channel* when it is defined as:

$$Y = X \oplus Z \tag{3.16}$$

[38] We refer readers to [221, Chapters 8-10] for a discussion of the DFT from a signal processing perspective.

where $X \in \mathcal{X}$ is the input random variable, $Y \in \mathcal{X}$ is the output random variable, and $Z \in \mathcal{X}$ is the additive noise random variable that is independent of X with pmf $P_Z = (P_Z(0), \ldots, P_Z(q-1)) \in \mathcal{P}_q$. The channel transition probability matrix corresponding to (3.16) is the $\mathcal{X}$-circulant stochastic matrix $\mathsf{circ}_{\mathcal{X}}(P_Z) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$, which is also doubly stochastic (i.e. both $\mathsf{circ}_{\mathcal{X}}(P_Z)$ and $\mathsf{circ}_{\mathcal{X}}(P_Z)^T$ belong to $\mathbb{R}_{\mathsf{sto}}^{q\times q}$). Indeed, for an additive noise channel, it is well-known that the pmf of Y, $P_Y \in \mathcal{P}_q$, can be obtained from the pmf of X, $P_X \in \mathcal{P}_q$, by $\mathcal{X}$-circular convolution:

$$P_Y = P_X \, \mathsf{circ}_{\mathcal{X}}(P_Z) \, . \tag{3.17}$$

We remark that in the context of various channel symmetries in the literature (see [227, Section VI.B] for a discussion), additive noise channels correspond to "group-noise" channels, and are input symmetric, output symmetric, Dobrushin symmetric, and Gallager symmetric.

The *q-ary symmetric channel* is an additive noise channel on the Abelian group $(\mathcal{X}, \oplus)$ with noise pmf:

$$P_Z = w_\delta \triangleq \left(1-\delta, \frac{\delta}{q-1}, \ldots, \frac{\delta}{q-1}\right) \in \mathcal{P}_q \tag{3.18}$$

where $\delta \in [0,1]$. Its channel transition probability matrix is denoted $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$:

$$W_\delta \triangleq \mathsf{circ}_{\mathcal{X}}(w_\delta) = \left[{w_\delta}^T \;\; P_q^T {w_\delta}^T \cdots \left(P_q^T\right)^{q-1} {w_\delta}^T \right]^T \tag{3.19}$$

which has $1-\delta$ in the principal diagonal entries and $\delta/(q-1)$ in all other entries regardless of the choice of group $(\mathcal{X}, \oplus)$. We may interpret δ as the total crossover probability of the symmetric channel. Indeed, when $q = 2$, W_δ represents a BSC with crossover probability $\delta \in [0,1]$. Although W_δ is only stochastic when $\delta \in [0,1]$, we will refer to the parametrized convex set of matrices $\{W_\delta \in \mathbb{R}_{\mathsf{sym}}^{q\times q} : \delta \in \mathbb{R}\}$ with parameter δ as the "symmetric channel matrices," where each W_δ has the form (3.19) such that every row and column sums to unity. We conclude this subsection with a list of properties of symmetric channel matrices.

Proposition 3.4 (Properties of Symmetric Channel Matrices). *The symmetric channel matrices, $\{W_\delta \in \mathbb{R}_{\mathsf{sym}}^{q\times q} : \delta \in \mathbb{R}\}$, satisfy the following properties:*

1. *For every $\delta \in \mathbb{R}$, W_δ is a symmetric circulant matrix.*
2. *The DFT matrix $F_q \in \mathcal{V}_q(\mathbb{C}^q)$, which is defined entry-wise as:*

$$\forall j,k \in \{1,\ldots,q\}, \;\; [F_q]_{j,k} = \frac{1}{\sqrt{q}} \exp\left(\frac{2\pi(j-1)(k-1)i}{q}\right) \tag{3.20}$$

where $i = \sqrt{-1}$, jointly diagonalizes W_δ for every $\delta \in \mathbb{R}$. Moreover, the corresponding eigenvalues or Fourier coefficients, $\{\lambda_j(W_\delta) = [F_q^H W_\delta F_q]_{j,j} : j \in \{1,\ldots,q\}\}$, are real:

$$\lambda_j(W_\delta) = \begin{cases} 1 & , \quad j = 1 \\ 1-\delta-\frac{\delta}{q-1} & , \quad j \in \{2,\ldots,q\} \end{cases} .$$

3. *For every* $\delta \in [0,1]$, W_δ *is a doubly stochastic matrix that has the uniform pmf* $\boldsymbol{u} \triangleq (1/q, \ldots, 1/q)$ *as its stationary distribution:*

$$\boldsymbol{u} W_\delta = \boldsymbol{u}\,.$$

4. *For every* $\delta \in \mathbb{R}$ *such that* $\delta \neq \frac{q-1}{q}$, $W_\delta^{-1} = W_\tau$, *where:*

$$\tau = -\frac{\delta}{1-\delta-\frac{\delta}{q-1}}\,,$$

and for $\delta = \frac{q-1}{q}$, $W_\delta = \frac{1}{q}\boldsymbol{1}\boldsymbol{1}^T$ *is unit rank and singular.*

5. *The set* $\left\{W_\delta \in \mathbb{R}_{\mathsf{sym}}^{q\times q} : \delta \in \mathbb{R} \text{ and } \delta \neq \frac{q-1}{q}\right\}$ *with the operation of matrix multiplication is an Abelian group.*

Proof. See appendix B.3. ■

We remark that the (complex) unitary diagonalization of W_δ in part 2 of Proposition 3.4 holds because circulant matrices are normal, and normal matrices admit a (complex) orthonormal eigenbasis by the complex spectral theorem, cf. [17, Theorem 7.9]. However, since W_δ is also symmetric, it admits a (real) orthonormal eigenbasis by the real spectral theorem, cf. [17, Theorem 7.13]. Indeed, from a Fourier analysis perspective, the vector w_δ is circularly symmetric, and hence its Fourier series only has cosine terms composed of conjugate DFT bases. These cosines can be completed to a basis as shown in [129, Problem 2.2.P10], where the (real) orthogonal matrix that diagonalizes W_δ (or any other real symmetric circulant matrix for that matter) is the *discrete Hartley transform* matrix $H_q \in \mathcal{V}_q(\mathbb{R}^q)$, which can be constructed from the DFT matrix:

$$H_q = \mathrm{Re}\{F_q\} + \mathrm{Im}\{F_q\}\,. \tag{3.21}$$

The eigenvalues of W_δ are of course real as W_δ is symmetric.

Furthermore, intuition from signal processing can be used to understand part 4 of Proposition 3.4 as well. Note that when $\delta < \frac{q-1}{q}$, the diagonal of W_δ is greater than all other entries, and when $\delta > \frac{q-1}{q}$, the diagonal of W_δ is less than all other entries. So, for $\delta \in [0, \frac{q-1}{q})$, w_δ can be construed as an impulse response with non-negative entries that behaves like a *low-pass filter*. Its inverse filter must be a *high-pass filter* whose impulse response has both positive and negative entries, and indeed, $W_\delta^{-1} = W_\tau$ has $\tau \leq 0$ according to part 4 of Proposition 3.4. In particular, $W_0^{-1} = W_0$ is the identity *all-pass filter* case.

■ 3.2 Motivation: Criteria for Domination by a Symmetric Channel

As we mentioned at the outset, our work is partly motivated by [231, Section 6], where the authors demonstrate an intriguing relation between less noisy domination by an

erasure channel and the contraction coefficient of the SDPI (3.2). For a common input alphabet $\mathcal{X} = [q]$, consider a channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ and a *q-ary erasure channel* $E_\epsilon \in \mathbb{R}_{\mathsf{sto}}^{q\times(q+1)}$ with erasure probability $\epsilon \in [0,1]$. It is proved in [231, Proposition 15] that $E_\epsilon \succeq_{\mathsf{ln}} V$ if and only if $\eta_{\mathsf{KL}}(V) \le 1-\epsilon$, where $\eta_{\mathsf{KL}}(V) \in [0,1]$ is the contraction coefficient of the channel $V = P_{Y|X}$ for KL divergence (cf. Definition 2.5 in chapter 2), and it is also the best possible constant that can be inserted into the SDPI (3.2) (see e.g. (2.54) in chapter 2 or [231, Theorem 4]). This result illustrates that the q-ary erasure channel E_ϵ with the largest erasure probability $\epsilon \in [0,1]$ (or the smallest channel capacity) that is less noisy than V has $\epsilon = 1 - \eta_{\mathsf{KL}}(V)$:[39]

$$\eta_{\mathsf{KL}}(V) = \min\{\beta \in [0,1] : E_{1-\beta} \succeq_{\mathsf{ln}} V\}\,. \tag{3.22}$$

Furthermore, there are several simple upper bounds on η_{KL} that provide sufficient conditions for such less noisy domination. For example, if the ℓ^1-distances between the rows of V are all bounded by 2α for some $\alpha \in [0,1]$, then $\eta_{\mathsf{KL}}(V) \le \alpha$, cf. [49] or part 7 of Proposition 2.5. Another criterion follows from *Doeblin minorization* [234, Remark III.2]: if for some pmf $p \in \mathcal{P}_s$ and some $\alpha \in (0,1)$, $V \ge \alpha\,\mathbf{1}p$ entry-wise, then $E_\alpha \succeq_{\mathsf{deg}} V$, which implies that $E_\alpha \succeq_{\mathsf{ln}} V$ (using Proposition 3.3), and hence, $\eta_{\mathsf{KL}}(V) \le 1-\alpha$.[40]

To extend these ideas, we consider the following question: *What is the q-ary symmetric channel W_δ with the largest value of $\delta \in \left[0, \frac{q-1}{q}\right]$ (or the smallest channel capacity) such that $W_\delta \succeq_{\mathsf{ln}} V$?*[41] Much like the bounds on η_{KL} in the erasure channel context, the goal of this chapter is to address this question by establishing simple criteria for testing $\succeq_{\mathsf{ln}}$ domination by a q-ary symmetric channel. We next provide several other reasons why determining whether a q-ary symmetric channel dominates a given channel V is interesting.

Firstly, since $\succeq_{\mathsf{ln}}$ tensorizes (see Lemma B.3 in appendix B.4), if $W \succeq_{\mathsf{ln}} V$, then $W^{\otimes n} \succeq_{\mathsf{ln}} V^{\otimes n}$, where $W^{\otimes n}$ is the n-fold tensor product of W (or equivalently, the n-fold Kronecker product of the matrix W with itself). This in turn implies that $I(U;Y_{W}{}_1^n) \ge I(U;Y_{V}{}_1^n)$ for every Markov chain $U \to X_1^n \to (Y_{W}{}_1^n, Y_{V}{}_1^n)$ (see Definition 3.2), where the conditional distributions of $Y_{W}{}_1^n$ and $Y_{V}{}_1^n$ given X_1^n are determined by $W^{\otimes n}$ and $V^{\otimes n}$, respectively. Thus, many impossibility results (in statistical decision theory for example) that are proven by exhibiting bounds on quantities such as $I(U;Y_{W}{}_1^n)$ transparently carry over to statistical experiments with observations on the basis of $Y_{V}{}_1^n$. Since it is common to study the q-ary symmetric observation model

[39] A q-ary erasure channel E_ϵ with erasure probability $\epsilon \in [0,1]$ has channel capacity $C(\epsilon) = \log(q)(1-\epsilon)$, which is linear and decreasing.

[40] In fact, the stronger condition that $E_\alpha \succeq_{\mathsf{deg}} V$ allows one to prove that contraction coefficients for general f-divergences are upper bounded by $1-\alpha$ as shown in [234]. We refer readers to [29, Section 3] for a classical treatment of how Doeblin minorization was applied to Markov processes to derive uniform geometric rates of convergence in TV distance to their stationary distributions.

[41] A q-ary symmetric channel W_δ with total crossover probability $\delta \in \left[0, \frac{q-1}{q}\right]$ has channel capacity $C(\delta) = \log(q) - H(w_\delta)$, which is convex and decreasing. Here, $H(w_\delta)$ denotes the Shannon entropy of the pmf w_δ.

(especially with $q = 2$), we can leverage its sample complexity lower bounds for other V.

Secondly, we present a self-contained information theoretic motivation. $W \succeq_{\mathsf{ln}} V$ if and only if $C_S = 0$, where C_S is the secrecy capacity of the *Wyner wiretap channel* with V as the main (legal receiver) channel and W as the eavesdropper channel [57, Corollary 3], [58, Corollary 17.11]. Therefore, finding the maximally noisy q-ary symmetric channel that dominates V establishes the minimal noise required on the eavesdropper link so that secret communication is feasible.

Thirdly, $\succeq_{\mathsf{ln}}$ domination turns out to entail a comparison of Dirichlet forms (see subsection 3.3.4), and consequently, allows us to prove *Poincaré and logarithmic Sobolev inequalities* for V from well-known results on q-ary symmetric channels. These functional inequalities are cornerstones of the modern approach to Markov chains and concentration of measure [69, 206].

■ 3.3 Main Results

In this section, we first delineate some guiding sub-questions of our study, indicate the main results that address them, and then present these results in the ensuing subsections. We will delve into the following four leading questions:

1. *Can we test the less noisy preorder $\succeq_{\mathsf{ln}}$ without using KL divergence?*
   Yes, we can use any non-linear operator convex f-divergence as shown in Theorem 3.1. This theorem generalizes the result in Proposition 2.6 in chapter 2, and we provide an application of the theorem in Theorem 3.2. Furthermore, we generalize Theorem 3.1 to arbitrary measurable spaces for the special case of χ^2-divergence in Theorem 3.3. This χ^2-divergence case turns out to be most useful in our study.

2. *Given a channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$, is there a simple sufficient condition for less noisy domination by a q-ary symmetric channel $W_\delta \succeq_{\mathsf{ln}} V$?*
   Yes, a condition using degradation (which implies less noisy domination) is presented in Theorem 3.4.

3. *Can we say anything stronger about less noisy domination by a q-ary symmetric channel when $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ is an additive noise channel?*
   Yes, Theorem 3.5 outlines the structure of additive noise channels in this context (and Figure 3.2 depicts it).

4. *Why are we interested in less noisy domination by q-ary symmetric channels?*
   Because this permits us to compare Dirichlet forms as portrayed in Theorem 3.6.

We next elaborate on these aforementioned theorems.

■ 3.3.1 Characterization of Less Noisy Preorder using Operator Convexity

Our most general result illustrates that although less noisy domination is a preorder defined using KL divergence, one can equivalently define it using any non-linear op-

erator convex f-divergence. We refer readers to appendix B.2 for a brief primer on operator monotonicity and operator convexity (if desired). The next theorem presents our equivalent characterization of $\succeq_{\mathsf{In}}$.

Theorem 3.1 (Operator Convex f-Divergence Characterization of $\succeq_{\mathsf{In}}$). *Consider any non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$. Then, for any two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q \times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q \times s}$ with the same input alphabet, $W \succeq_{\mathsf{In}} V$ if and only if:*

$$D_f(P_X W || Q_X W) \geq D_f(P_X V || Q_X V)$$

for every pair of input pmfs $P_X, Q_X \in \mathcal{P}_q$.

Theorem 3.1 is proved in subsection 3.6.1 using techniques from [46]. As conveyed in parts 2 and 3 of Theorem B.1 in appendix B.2, it is well-known that $f(t) = t \log(t)$ and $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$ for any $\alpha \in (0, 1) \cup (1, 2]$ are operator convex functions. Hence, one class of f-divergences that satisfy the conditions of the theorem are the Hellinger divergences of order $\alpha \in (0, 2]$, where the cases $\alpha = 1$ and $\alpha = 2$ correspond to KL and χ^2-divergences, respectively (see subsection 2.2.1 in chapter 2).

It is worth comparing Theorem 3.1 with the equivalent characterization of degradation over dichotomies in (3.8). In particular, these results transparently unify the definitions of degradation and less noisy in the $q = 2$ setting, because both preorders are defined by (3.8) holding for different classes of functions $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$. Indeed, the class of convex functions defines degradation, and the smaller class of operator convex functions defines less noisy.

We next demonstrate an application of Theorem 3.1 by proving a generalization of the so called *Samorodnitsky's SDPI*. Following the exposition in [231, Section 6.2], consider the discrete random variables $U, X_1, \ldots, X_n, Y_1, \ldots, Y_n$ with finite alphabets, where $n \in \mathbb{N}$. Suppose we are given a *memoryless* channel $P_{Y_1^n | X_1^n}$:

$$P_{Y_1^n | X_1^n} = \prod_{j=1}^{n} P_{Y_j | X_j} \tag{3.23}$$

which means that the stochastic matrix corresponding to $P_{Y_1^n | X_1^n}$ is a tensor product of the stochastic matrices corresponding to $P_{Y_j | X_j}$ over all $j \in \{1, \ldots, n\}$. Define the contraction coefficient of the channel $P_{Y_j | X_j}$ for any f-divergence as, cf. Definition 2.5 in chapter 2:

$$\eta_j \triangleq \eta_f(P_{Y_j | X_j}) \tag{3.24}$$

for every $j \in \{1, \ldots, n\}$. While the SDPI for $P_{Y_1^n | X_1^n}$ is characterized by the contraction coefficient $\eta_f(P_{Y_1^n | X_1^n})$, it is desirable to obtain a loosening of this SDPI in terms of the single-letter contraction coefficients $\{\eta_j : j \in \{1, \ldots, n\}\}$. To illustrate one such tensorized SDPI, suppose $\eta_j = \eta$ for all $j \in \{1, \ldots, n\}$, and fix any non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$. Due to Proposition 2.6 in chapter 2, it is straightforward to verify that [231, Theorem 5] and [231, Corollary 6] hold for

any f-divergence with non-linear operator convex f (rather than just KL divergence). As a result, [231, Corollary 6] yields the tensorization bound:

$$\eta_f(P_{Y_1^n|X_1^n}) \leq 1 - (1-\eta)^n \tag{3.25}$$

which can be construed as an analogue of part 5 of Proposition 2.3 for contraction coefficients of channels. Thus, for every pair of input distributions $P_{X_1^n}$ and $Q_{X_1^n}$, we have the tensorized SDPI:

$$D_f(P_{Y_1^n}||Q_{Y_1^n}) \leq (1-(1-\eta)^n)\, D_f(P_{X_1^n}||Q_{X_1^n}) \tag{3.26}$$

where $P_{Y_1^n}$ and $Q_{Y_1^n}$ are the output distributions after passing $P_{X_1^n}$ and $Q_{X_1^n}$ through the channel $P_{Y_1^n|X_1^n}$, respectively. Likewise, for any joint distribution P_{U,X_1^n} such that $U \to X_1^n \to Y_1^n$ form a Markov chain, we have the tensorized SDPI (see (2.18) and (2.54)):

$$I_f(U;Y_1^n) \leq (1-(1-\eta)^n)\, I_f(U;X_1^n)\,. \tag{3.27}$$

However, as argued in [231, Section 6.2], "stronger [single-letter bounds] can be given if we have finer knowledge" about the pair $(P_{X_1^n}, Q_{X_1^n})$ or the distribution P_{U,X_1^n}. In this vein, the following theorem presents tighter bounds on $D_f(P_{Y_1^n}||Q_{Y_1^n})$ and $I_f(U;Y_1^n)$ using the single-letter contraction coefficients $\{\eta_j : j \in \{1,\dots,n\}\}$, and terms representing the "average" input f-divergence and "average" mutual f-information contained in subsets of X_1^n, respectively.

Theorem 3.2 (Generalized Samorodnitsky's SDPI). *Consider any non-linear operator convex function $f : (0,\infty) \to \mathbb{R}$ such that $f(1) = 0$. Suppose U, $X_1,\dots,X_n$, $Y_1,\dots,Y_n$ are discrete random variables with finite alphabets such that the channel $P_{Y_1^n|X_1^n}$ is fixed and memoryless (see (3.23)). Let S be a random subset of $\{1,\dots,n\}$ that is constructed by independently including each element $j \in \{1,\dots,n\}$ with probability η_j (defined in (3.24)), and assume that S is independent of (U, X_1^n, Y_1^n). Then, for every pair of input distributions $P_{X_1^n}$ and $Q_{X_1^n}$:*

$$D_f(P_{Y_1^n}||Q_{Y_1^n}) \leq \sum_{T \subseteq \{1,\dots,n\}} P_S(T)\, D_f(P_{X_T}||Q_{X_T})$$

where P_S is the distribution of S, $X_T \triangleq \{X_k : k \in T\}$ for any subset $T \subseteq \{1,\dots,n\}$, $D_f(P_{X_\varnothing}||Q_{X_\varnothing}) = 0$, and $P_{Y_1^n}$ and $Q_{Y_1^n}$ are the output distributions after passing $P_{X_1^n}$ and $Q_{X_1^n}$ through the channel $P_{Y_1^n|X_1^n}$, respectively. Similarly, for any joint distribution P_{U,X_1^n} such that $U \to X_1^n \to Y_1^n$ form a Markov chain:

$$I_f(U;Y_1^n) \leq I_f(U;X_S,S) = \sum_{T \subseteq \{1,\dots,n\}} P_S(T)\, I_f(U;X_T)$$

where $I_f(U;X_\varnothing) = 0$. Moreover, if $\eta_j = \eta$ for all $j \in \{1,\dots,n\}$, then we have:

$$I_f(U;Y_1^n) \leq \sum_{k=0}^{n} \binom{n}{k} \eta^k (1-\eta)^{n-k} I_k$$

where for every $k \in \{0, \dots, n\}$, we define I_k to be the "average" mutual f-information contained in subsets of X_1^n with cardinality k:

$$I_k \triangleq \binom{n}{k}^{-1} \sum_{\substack{T \subseteq \{1,\dots,n\} \\ |T|=k}} I_f(U; X_T) \,.$$

We prove Theorem 3.2 using Theorem 3.1 in appendix B.4 by mimicking the proof technique of [231]. The KL divergence case of Theorem 3.2 was first derived by Samorodnitsky in [241] using linear programming techniques to prove a special case of the *Courtade-Kumar conjecture* from [164]. It was then distilled into its present form in [231, Theorem 20, Remark 6], where a simpler proof was also given. Our result in Theorem 3.2 generalizes this KL divergence case to all non-linear operator convex f-divergences (which includes, for example, all Hellinger divergences of order $\alpha \in (0, 2]$, as mentioned earlier). We also remark that the KL divergence case of this result has other applications such as the strengthening of *Mrs. Gerber's Lemma* (cf. [81, Section 2.1]) in [231, Remark 5]. Lastly, as explained in [231, Remark 4], if $\eta_j = \eta$ for all $j \in \{1, \dots, n\}$ in Theorem 3.2, then approximating the $\mathsf{binomial}(n, \eta)$ distribution with its expectation $n\eta$ yields:

$$I_f(U; Y_1^n) \lesssim I_{n\eta} \tag{3.28}$$

for any Markov chain $U \to X_1^n \to Y_1^n$, which conveys that only information about U contained in subsets of X_1^n with cardinality bounded by $n\eta$ can be inferred from Y_1^n.

While our development in this thesis is mostly for finite alphabets, Theorem 3.1 can be generalized for arbitrary measurable spaces. We close this subsection by providing one example of such a generalization for the special case of χ^2-divergence. To state and prove an equivalent characterization of $\succeq_{\mathsf{ln}}$ via χ^2-divergence for general measurable spaces, we introduce some additional notation pertinent only to this result. Let $(\mathcal{X}, \mathcal{F})$, $(\mathcal{Y}_1, \mathcal{H}_1)$, and $(\mathcal{Y}_2, \mathcal{H}_2)$ be three measurable spaces, and let $W : \mathcal{H}_1 \times \mathcal{X} \to [0, 1]$ and $V : \mathcal{H}_2 \times \mathcal{X} \to [0, 1]$ be two *Markov kernels* (or channels) acting on the same source space $(\mathcal{X}, \mathcal{F})$. Given any probability measure P_X on $(\mathcal{X}, \mathcal{F})$, we denote by $P_X W$ the probability measure on $(\mathcal{Y}_1, \mathcal{H}_1)$ induced by the push-forward of P_X.[42] Recall that for any two probability measures P_X and Q_X on $(\mathcal{X}, \mathcal{F})$, their KL divergence is given by:

$$D(P_X||Q_X) \triangleq \begin{cases} \displaystyle\int_{\mathcal{X}} \log\left(\frac{dP_X}{dQ_X}\right) dP_X & , \quad P_X \ll Q_X \\ +\infty & , \quad \text{otherwise} \end{cases} \tag{3.29}$$

and their χ^2-divergence is given by:

$$\chi^2(P_X||Q_X) \triangleq \begin{cases} \displaystyle\int_{\mathcal{X}} \left(\frac{dP_X}{dQ_X}\right)^2 dQ_X - 1 & , \quad P_X \ll Q_X \\ +\infty & , \quad \text{otherwise} \end{cases} \tag{3.30}$$

[42] Here, we can think of X and Y as random variables with codomains $\mathcal{X}$ and $\mathcal{Y}$, respectively. The Markov kernel W behaves like the conditional distribution of Y given X (under regularity conditions). Moreover, when the distribution of X is P_X, the corresponding distribution of Y is $P_Y = P_X W$.

where $P_X \ll Q_X$ denotes that P_X is absolutely continuous with respect to Q_X, and $\frac{dP_X}{dQ_X}$ denotes the Radon-Nikodym derivative of P_X with respect to Q_X. Furthermore, the characterization of $\succeq_{\mathsf{ln}}$ in Proposition 3.1 extends naturally to general Markov kernels; indeed, $W \succeq_{\mathsf{ln}} V$ if and only if $D(P_X W||Q_X W) \geq D(P_X V||Q_X V)$ for every pair of probability measures P_X and Q_X on $(\mathcal{X}, \mathcal{F})$. The next theorem presents the χ^2-divergence characterization of $\succeq_{\mathsf{ln}}$.

Theorem 3.3 (χ^2-Divergence Characterization of $\succeq_{\mathsf{ln}}$). *For any Markov kernels $W : \mathcal{H}_1 \times \mathcal{X} \to [0,1]$ and $V : \mathcal{H}_2 \times \mathcal{X} \to [0,1]$ acting on the same source space, $W \succeq_{\mathsf{ln}} V$ if and only if:*

$$\chi^2(P_X W||Q_X W) \geq \chi^2(P_X V||Q_X V)$$

for every pair of probability measures P_X and Q_X on $(\mathcal{X}, \mathcal{F})$.

Theorem 3.3 is proved in subsection 3.6.2.

■ 3.3.2 Less Noisy Domination by Symmetric Channels

Our remaining results are all concerned with less noisy (and degraded) domination by q-ary symmetric channels. Suppose we are given a q-ary symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ with $\delta \in [0,1]$, and another channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ with common input and output alphabets. Then, the next result provides a sufficient condition for when $W_\delta \succeq_{\mathsf{deg}} V$.

Theorem 3.4 (Sufficient Condition for Degradation by Symmetric Channels). *Given a channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ with $q \geq 2$ and minimum probability entry $\nu = \min\{[V]_{i,j} : i, j \in \{1, \ldots, q\}\}$, we have:*

$$0 \leq \delta \leq \frac{\nu}{1 - (q-1)\nu + \frac{\nu}{q-1}} \quad \Rightarrow \quad W_\delta \succeq_{\mathsf{deg}} V \, .$$

Theorem 3.4 is proved in section 3.8. We note that the sufficient condition in Theorem 3.4 is tight as there exist channels V that violate $W_\delta \succeq_{\mathsf{deg}} V$ when $\delta > \nu/(1 - (q-1)\nu + \frac{\nu}{q-1})$. Furthermore, Theorem 3.4 also provides a sufficient condition for $W_\delta \succeq_{\mathsf{ln}} V$ due to Proposition 3.3.

■ 3.3.3 Structure of Additive Noise Channels

Our next major result is concerned with understanding when q-ary symmetric channels operating on an Abelian group $(\mathcal{X}, \oplus)$ dominate other additive noise channels on $(\mathcal{X}, \oplus)$, which are defined in (3.16), in the less noisy and degraded senses. Given a symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ with $\delta \in [0,1]$, we define the *additive less noisy domination region* of W_δ as:

$$\mathcal{L}_{W_\delta}^{\mathsf{add}} \triangleq \{v \in \mathcal{P}_q : W_\delta = \mathsf{circ}_{\mathcal{X}}(w_\delta) \succeq_{\mathsf{ln}} \mathsf{circ}_{\mathcal{X}}(v)\} \tag{3.31}$$

which is the set of all noise pmfs whose corresponding channel transition probability matrices are dominated by W_δ in the less noisy sense. Likewise, we define the *additive degradation region* of W_δ as:

$$\mathcal{D}_{W_\delta}^{\mathsf{add}} \triangleq \{v \in \mathcal{P}_q : W_\delta = \mathsf{circ}_{\mathcal{X}}(w_\delta) \succeq_{\mathsf{deg}} \mathsf{circ}_{\mathcal{X}}(v)\} \tag{3.32}$$

which is the set of all noise pmfs whose corresponding channel transition probability matrices are degraded versions of W_δ. The next theorem exactly characterizes $\mathcal{D}_{W_\delta}^{\mathsf{add}}$, and "bounds" $\mathcal{L}_{W_\delta}^{\mathsf{add}}$ in a set theoretic sense.

Theorem 3.5 (Additive Less Noisy Domination and Degradation Regions for Symmetric Channels). *Given a symmetric channel $W_\delta = \mathsf{circ}_{\mathcal{X}}(w_\delta) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $\delta \in [0, \frac{q-1}{q}]$ and $q \geq 2$, we have:*

$$\begin{aligned}\mathcal{D}_{W_\delta}^{\mathsf{add}} &= \mathsf{conv}\Big(\Big\{w_\delta P_q^k : k \in [q]\Big\}\Big) \\ &\subseteq \mathsf{conv}\Big(\Big\{w_\delta P_q^k : k \in [q]\Big\} \cup \Big\{w_\gamma P_q^k : k \in [q]\Big\}\Big) \\ &\subseteq \mathcal{L}_{W_\delta}^{\mathsf{add}} \subseteq \{v \in \mathcal{P}_q : \|v - \boldsymbol{u}\|_2 \leq \|w_\delta - \boldsymbol{u}\|_2\}\end{aligned}$$

where the first set inclusion is strict for $\delta \in (0, \frac{q-1}{q})$ and $q \geq 3$, P_q denotes the generator cyclic permutation matrix as defined in (3.15), and:

$$\gamma = \frac{1-\delta}{1-\delta+\frac{\delta}{(q-1)^2}}.$$

Furthermore, $\mathcal{L}_{W_\delta}^{\mathsf{add}}$ is a closed and convex set that is invariant under the permutations $\{P_x \in \mathbb{R}^{q\times q} : x \in \mathcal{X}\}$ defined in (3.9) corresponding to the underlying Abelian group $(\mathcal{X}, \oplus)$ (i.e. $v \in \mathcal{L}_{W_\delta}^{\mathsf{add}} \Rightarrow vP_x \in \mathcal{L}_{W_\delta}^{\mathsf{add}}$ for every $x \in \mathcal{X}$).

Theorem 3.5 is a compilation of several results. As explained at the very end of subsection 3.7.2, Proposition 3.6 (in subsection 3.5.1), Corollary 3.1 (in subsection 3.5.2), part 1 of Proposition 3.9 (in subsection 3.7.1), and Proposition 3.11 (in subsection 3.7.2) make up Theorem 3.5. We remark that according to numerical evidence, the second and third set inclusions in Theorem 3.5 appear to be strict, and $\mathcal{L}_{W_\delta}^{\mathsf{add}}$ seems to be a strictly convex set. The content of Theorem 3.5 and these observations are illustrated in Figure 3.2, which portrays the probability simplex of noise pmfs for $q = 3$ and the pertinent regions which capture less noisy domination and degradation by a q-ary symmetric channel.

■ 3.3.4 Comparison of Dirichlet Forms

As mentioned in section 3.2, one of the reasons we study q-ary symmetric channels and prove Theorems 3.4 and 3.5 is because less noisy domination implies useful bounds between Dirichlet forms. Recall that the q-ary symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $\delta \in [0, 1]$ has uniform stationary distribution $\mathbf{u} \in \mathcal{P}_q$ (see part 3 of Proposition 3.4). For any channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ that is doubly stochastic and has uniform stationary distribution, we may define a corresponding *Dirichlet form*:

$$\forall f \in \mathbb{R}^q, \quad \mathcal{E}_V(f, f) = \frac{1}{q} f^T (I_q - V) f \tag{3.33}$$

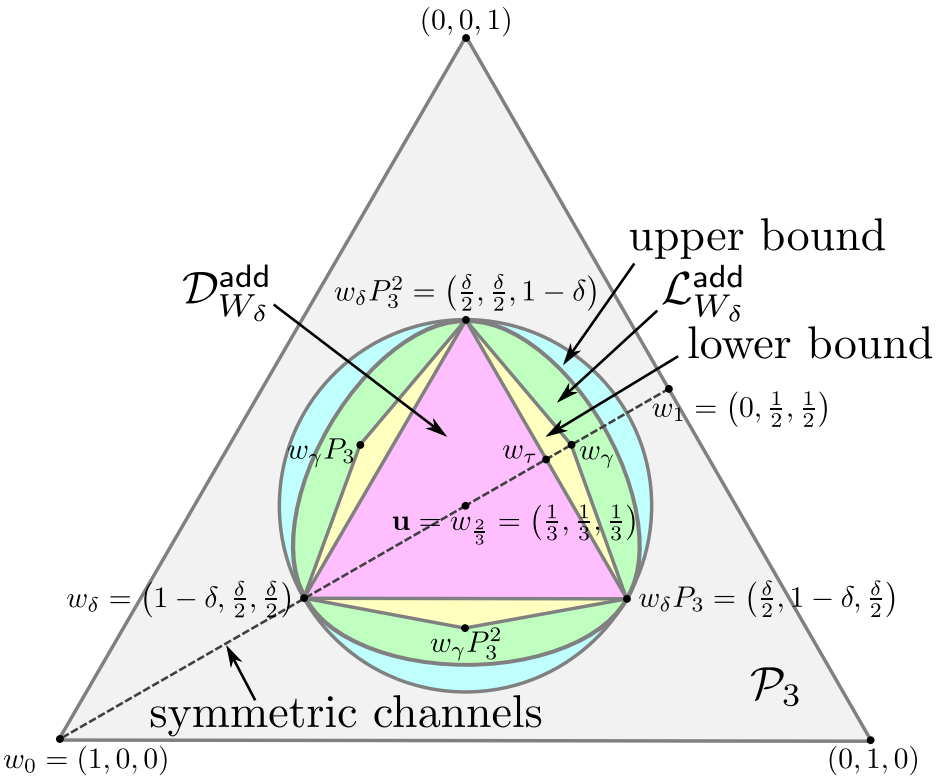


Figure 3.2. Illustration of the additive less noisy domination region and the additive degradation region for a q-ary symmetric channel when $q = 3$ and $\delta \in (0, 2/3)$: The gray triangle denotes the probability simplex of noise pmfs $\mathcal{P}_3$. The dotted line denotes the parametrized family of noise pmfs of 3-ary symmetric channels $\{w_\delta \in \mathcal{P}_3 : \delta \in [0,1]\}$; its noteworthy points are w_0 (corner of simplex, W_0 is less noisy than every channel), w_δ for some fixed $\delta \in (0, 2/3)$ (noise pmf of 3-ary symmetric channel W_δ under consideration), $w_{2/3} = \mathbf{u}$ (uniform pmf, $W_{2/3}$ is more noisy than every channel), w_τ with $\tau = 1 - (\delta/2)$ (W_τ is the extremal symmetric channel that is degraded by W_δ), w_γ with $\gamma = (1-\delta)/(1-\delta+(\delta/4))$ (W_γ is a 3-ary symmetric channel that is not degraded by W_δ but $W_\delta \succeq_{\mathsf{ln}} W_\gamma$), and w_1 (edge of simplex). The magenta triangle denotes the additive degradation region $\mathsf{conv}(\{w_\delta, w_\delta P_3, w_\delta P_3^2\})$ of W_δ. The green convex region denotes the additive less noisy domination region of W_δ, and the yellow region $\mathsf{conv}(\{w_\delta, w_\delta P_3, w_\delta P_3^2, w_\gamma, w_\gamma P_3, w_\gamma P_3^2\})$ is its lower bound while the circular cyan region $\{v \in \mathcal{P}_3 : \|v - \mathbf{u}\|_2 \leq \|w_\delta - \mathbf{u}\|_2\}$ (which is a hypersphere for general $q \geq 3$) is its upper bound. Note that we do not need to specify the underlying group because there is only one group of order 3.

where $f = [f_1 \cdots f_q]^T \in \mathbb{R}^q$ are column vectors (as shown in [69] or [206]). Our final theorem portrays that $W_\delta \succeq_{\mathsf{ln}} V$ implies that the Dirichlet form corresponding to V dominates the Dirichlet form corresponding to W_δ pointwise. The Dirichlet form corresponding to W_δ is in fact a scaled version of the so called *standard Dirichlet form*:

$$\forall f \in \mathbb{R}^q, \;\; \mathcal{E}_{\mathsf{std}}(f,f) \triangleq \mathbb{VAR}_{\mathbf{u}}(f) = \frac{1}{q}\sum_{k=1}^{q} f_k^2 - \left(\frac{1}{q}\sum_{k=1}^{q} f_k\right)^2 \tag{3.34}$$

which is the Dirichlet form corresponding to the q-ary symmetric channel $W_{(q-1)/q} = \mathbf{1u}$ with all uniform conditional pmfs. Indeed, using $I_q - W_\delta = \frac{q\delta}{q-1}(I_q - \mathbf{1u})$, we have:

$$\forall f \in \mathbb{R}^q, \;\; \mathcal{E}_{W_\delta}(f,f) = \frac{q\delta}{q-1}\,\mathcal{E}_{\mathsf{std}}(f,f)\,. \tag{3.35}$$

The standard Dirichlet form is the usual choice for Dirichlet form comparison because its logarithmic Sobolev constant has been precisely computed in [69, Appendix, Theorem A.1]. So, we present Theorem 3.6 using $\mathcal{E}_{\mathsf{std}}$ rather than $\mathcal{E}_{W_\delta}$.

Theorem 3.6 (Domination of Dirichlet Forms). *Given the doubly stochastic channels $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $\delta \in \left[0, \frac{q-1}{q}\right]$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$, if $W_\delta \succeq_{\mathsf{ln}} V$, then:*

$$\forall f \in \mathbb{R}^q, \;\; \mathcal{E}_V(f,f) \geq \frac{q\delta}{q-1}\,\mathcal{E}_{\mathsf{std}}(f,f)\,.$$

An extension of Theorem 3.6 is proved in section 3.9. The domination of Dirichlet forms shown in Theorem 3.6 has several useful consequences. A major consequence is that we can immediately establish Poincaré (spectral gap) inequalities and logarithmic Sobolev inequalities (LSIs) for the channel V using the corresponding inequalities for q-ary symmetric channels. For example, the LSI for $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $q > 2$ is:

$$D(f^2\mathbf{u}||\mathbf{u}) \leq \frac{(q-1)\log(q-1)}{(q-2)\delta}\,\mathcal{E}_{W_\delta}(f,f) \tag{3.36}$$

for all $f \in \mathbb{R}^q$ such that $\sum_{k=1}^q f_k^2 = q$, where we use (3.76) and the logarithmic Sobolev constant computed in part 1 of Proposition 3.12. As shown in appendix B.8, (3.36) is easily established using the known logarithmic Sobolev constant corresponding to the standard Dirichlet form. Using the LSI for V that follows from (3.36) and Theorem 3.6, we immediately obtain guarantees on the convergence rate and *hypercontractivity* properties of the associated *Markov semigroup* $\{\exp(-t(I_q - V)) : t \geq 0\}$. We refer readers to [69] and [206] for comprehensive accounts of such topics.

■ 3.4 Chapter Outline

We briefly outline the content of the ensuing sections in this chapter. In section 3.5, we study the structure of less noisy domination and degradation regions of channels. In section 3.6, we prove Theorems 3.1 and 3.3, and present some other equivalent characterizations of $\succeq_{\mathsf{ln}}$. We then derive several necessary and sufficient conditions for less noisy domination among additive noise channels in section 3.7, which together with the results of section 3.5, culminates in a proof of Theorem 3.5. Section 3.8 provides a proof of Theorem 3.4, and section 3.9 introduces LSIs and proves an extension of Theorem 3.6. Finally, we conclude our discussion and propose future research directions in section 3.10.

■ 3.5 Less Noisy Domination and Degradation Regions

In this section, we focus on understanding the "geometric" aspects of less noisy domination and degradation by channels. We begin by deriving some simple characteristics of the sets of channels that are dominated by some fixed channel in the less noisy and degraded senses. We then specialize our results for additive noise channels, and this culminates in a complete characterization of $\mathcal{D}_{W_\delta}^{\mathsf{add}}$ and derivations of certain properties of $\mathcal{L}_{W_\delta}^{\mathsf{add}}$ presented in Theorem 3.5.

Let $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ be a fixed channel with $q, r \in \mathbb{N}$, and define its *less noisy domination region*:

$$\mathcal{L}_W \triangleq \left\{V \in \mathbb{R}_{\mathsf{sto}}^{q\times r} : W \succeq_{\mathsf{ln}} V\right\} \tag{3.37}$$

as the set of all channels on the same input and output alphabets that are dominated by W in the less noisy sense. Moreover, we define the *degradation region* of W:

$$\mathcal{D}_W \triangleq \left\{V \in \mathbb{R}_{\mathsf{sto}}^{q\times r} : W \succeq_{\mathsf{deg}} V\right\} \tag{3.38}$$

as the set of all channels on the same input and output alphabets that are degraded versions of W. Then, $\mathcal{L}_W$ and $\mathcal{D}_W$ satisfy the properties delineated below.

Proposition 3.5 (Less Noisy Domination and Degradation Regions). *Given the channel $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$, its less noisy domination region $\mathcal{L}_W$ and its degradation region $\mathcal{D}_W$ are non-empty, closed, convex, and output alphabet permutation symmetric (i.e. $V \in \mathcal{L}_W \Rightarrow VP \in \mathcal{L}_W$ and $V \in \mathcal{D}_W \Rightarrow VP \in \mathcal{D}_W$ for every permutation matrix $P \in \mathbb{R}^{r\times r}$).*

Proof.

Non-Emptiness of $\mathcal{L}_W$ and $\mathcal{D}_W$: $W \succeq_{\mathsf{ln}} W \Rightarrow W \in \mathcal{L}_W$, and $W \succeq_{\mathsf{deg}} W \Rightarrow W \in \mathcal{D}_W$. So, $\mathcal{L}_W$ and $\mathcal{D}_W$ are non-empty.

Closure of $\mathcal{L}_W$: Fix any two pmfs $P_X, Q_X \in \mathcal{P}_q$, and consider a sequence of channels $V_k \in \mathcal{L}_W$ such that $V_k \to V \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ as $k \to \infty$ (with respect to the Frobenius norm). Then, we also have $P_X V_k \to P_X V$ and $Q_X V_k \to Q_X V$ as $k \to \infty$ (with respect to the ℓ^2-norm). Hence, we get:

$$\begin{aligned} D(P_X V||Q_X V) &\leq \liminf_{k\to\infty} D(P_X V_k||Q_X V_k) \\ &\leq D(P_X W||Q_X W) \end{aligned}$$

where the first line follows from the lower semicontinuity of KL divergence [232, Theorem 1], [230, Theorem 3.6, Section 3.5], and the second line holds because $V_k \in \mathcal{L}_W$. This implies that for any two pmfs $P_X, Q_X \in \mathcal{P}_q$, the set:

$$\mathcal{S}\left(P_X, Q_X\right) = \left\{V \in \mathbb{R}_{\mathsf{sto}}^{q\times r} : D(P_X W||Q_X W) \geq D(P_X V||Q_X V)\right\}$$

is actually closed. Using Proposition 3.1, we have that:

$$\mathcal{L}_W = \bigcap_{P_X, Q_X \in \mathcal{P}_q} \mathcal{S}\left(P_X, Q_X\right).$$

So, $\mathcal{L}_W$ is closed since it is an intersection of closed sets [239].

Closure of $\mathcal{D}_W$: Consider a sequence of channels $V_k \in \mathcal{D}_W$ such that $V_k \to V \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ as $k \to \infty$. Since each $V_k = WA_k$ for some channel $A_k \in \mathbb{R}_{\mathsf{sto}}^{r\times r}$ belonging to the compact set $\mathbb{R}_{\mathsf{sto}}^{r\times r}$, there exists a subsequence A_{k_m} that converges by (sequential) compactness [239]: $A_{k_m} \to A \in \mathbb{R}_{\mathsf{sto}}^{r\times r}$ as $m \to \infty$. Hence, $V \in \mathcal{D}_W$ since $V_{k_m} = WA_{k_m} \to WA = V$ as $m \to \infty$, and $\mathcal{D}_W$ is a closed set.

Convexity of $\mathcal{L}_W$: Suppose $V_1, V_2 \in \mathcal{L}_W$, and fix any $\lambda \in [0,1]$. Then, for every $P_X, Q_X \in \mathcal{P}_q$, we have:

$$D(P_X W||Q_X W) \geq D(P_X(\lambda V_1 + (1-\lambda)V_2)||Q_X(\lambda V_1 + (1-\lambda)V_2))$$

by the convexity of KL divergence. Hence, $\mathcal{L}_W$ is convex.

Convexity of $\mathcal{D}_W$: If $V_1, V_2 \in \mathcal{D}_W$ so that $V_1 = WA_1$ and $V_2 = WA_2$ for some $A_1, A_2 \in \mathbb{R}_{\mathsf{sto}}^{r\times r}$, then $\lambda V_1 + (1-\lambda)V_2 = W(\lambda A_1 + (1-\lambda)A_2) \in \mathcal{D}_W$ for all $\lambda \in [0,1]$, and $\mathcal{D}_W$ is convex.

Symmetry of $\mathcal{L}_W$: This is obvious from Proposition 3.1 because KL divergence is invariant to permutations of its input pmfs.

Symmetry of $\mathcal{D}_W$: Given $V \in \mathcal{D}_W$ so that $V = WA$ for some $A \in \mathbb{R}_{\mathsf{sto}}^{r\times r}$, we have that $VP = WAP \in \mathcal{D}_W$ for every permutation matrix $P \in \mathbb{R}^{r\times r}$. This completes the proof. ■

While the channels in $\mathcal{L}_W$ and $\mathcal{D}_W$ all have the same output alphabet as W, as defined in (3.37) and (3.38), we may extend the output alphabet of W by adding zero probability letters. So, separate less noisy domination and degradation regions can be defined for each output alphabet size that is at least as large as the original output alphabet size of W. On a separate note, given a channel $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$, it is straightforward to verify that $\mathcal{L}_W = \mathcal{L}_{WP}$ and $\mathcal{D}_W = \mathcal{D}_{WP}$ for every permutation matrix $P \in \mathbb{R}^{r\times r}$. (Indeed, we have $W \succeq_{\mathsf{ln}} WP \succeq_{\mathsf{ln}} W$ and $W \succeq_{\mathsf{deg}} WP \succeq_{\mathsf{deg}} W$ for every permutation matrix $P \in \mathbb{R}^{r\times r}$.) Therefore, a channel W and any output alphabet permutation of it are equivalent from the perspective of less noisy domination and degradation.

■ 3.5.1 Less Noisy Domination and Degradation Regions for Additive Noise Channels

Often in information theory, we are concerned with additive noise channels on an Abelian group $(\mathcal{X}, \oplus)$ with $\mathcal{X} = [q]$ and $q \in \mathbb{N}$, as defined in (3.16). Such channels are completely defined by a noise pmf $P_Z \in \mathcal{P}_q$ with corresponding channel transition probability matrix $\mathsf{circ}_{\mathcal{X}}(P_Z) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$. Suppose $W = \mathsf{circ}_{\mathcal{X}}(w) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ is an additive noise channel with noise pmf $w \in \mathcal{P}_q$. Then, we are often only interested in the set of additive noise channels that are dominated by W. We define the *additive less noisy domination region* of W:

$$\mathcal{L}_W^{\mathsf{add}} \triangleq \{v \in \mathcal{P}_q : W \succeq_{\mathsf{ln}} \mathsf{circ}_{\mathcal{X}}(v)\} \tag{3.39}$$

as the set of all noise pmfs whose corresponding channel transition matrices are dominated by W in the less noisy sense. Likewise, we define the *additive degradation region* of W:

$$\mathcal{D}_W^{\mathsf{add}} \triangleq \{v \in \mathcal{P}_q : W \succeq_{\mathsf{deg}} \mathsf{circ}_{\mathcal{X}}(v)\} \tag{3.40}$$

as the set of all noise pmfs whose corresponding channel transition matrices are degraded versions of W. (These definitions generalize (3.31) and (3.32), and can also hold for any non-additive noise channel W.) The next proposition illustrates certain properties of $\mathcal{L}_W^{\mathsf{add}}$ and explicitly characterizes $\mathcal{D}_W^{\mathsf{add}}$.

Proposition 3.6 (Additive Less Noisy Domination and Degradation Regions). *Given the additive noise channel $W = \mathsf{circ}_{\mathcal{X}}(w) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with noise pmf $w \in \mathcal{P}_q$, we have:*

1. *$\mathcal{L}_W^{\mathsf{add}}$ and $\mathcal{D}_W^{\mathsf{add}}$ are non-empty, closed, convex, and invariant under the permutations $\{P_x \in \mathbb{R}^{q\times q} : x \in \mathcal{X}\}$ defined in (3.9) (i.e. $v \in \mathcal{L}_W^{\mathsf{add}} \Rightarrow vP_x \in \mathcal{L}_W^{\mathsf{add}}$ and $v \in \mathcal{D}_W^{\mathsf{add}} \Rightarrow vP_x \in \mathcal{D}_W^{\mathsf{add}}$ for every $x \in \mathcal{X}$).*
2. *$\mathcal{D}_W^{\mathsf{add}} = \mathsf{conv}(\{wP_x : x \in \mathcal{X}\}) = \{v \in \mathcal{P}_q : w \succeq_{\mathcal{X}} v\}$, where $\succeq_{\mathcal{X}}$ denotes the group majorization preorder as defined in appendix B.1.*

To prove Proposition 3.6, we will need the following lemma.

Lemma 3.1 (Additive Noise Channel Degradation). *Given two additive noise channels $W = \mathsf{circ}_{\mathcal{X}}(w) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ and $V = \mathsf{circ}_{\mathcal{X}}(v) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with noise pmfs $w, v \in \mathcal{P}_q$, $W \succeq_{\mathsf{deg}} V$ if and only if $V = W\mathsf{circ}_{\mathcal{X}}(z) = \mathsf{circ}_{\mathcal{X}}(z)W$ for some $z \in \mathcal{P}_q$ (i.e. for additive noise channels $W \succeq_{\mathsf{deg}} V$, the channel that degrades W to produce V is also an additive noise channel without loss of generality).*

Proof. Since $\mathcal{X}$-circulant matrices commute, we must have $W\mathsf{circ}_{\mathcal{X}}(z) = \mathsf{circ}_{\mathcal{X}}(z)W$ for every $z \in \mathcal{P}_q$. Furthermore, $V = W\mathsf{circ}_{\mathcal{X}}(z)$ for some $z \in \mathcal{P}_q$ implies that $W \succeq_{\mathsf{deg}} V$ by Definition 3.1. So, it suffices to prove that $W \succeq_{\mathsf{deg}} V$ implies $V = W\mathsf{circ}_{\mathcal{X}}(z)$ for some $z \in \mathcal{P}_q$. By Definition 3.1, $W \succeq_{\mathsf{deg}} V$ implies that $V = WR$ for some doubly stochastic channel $R \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ (as V and W are doubly stochastic). Let r with $r^T \in \mathcal{P}_q$ be the first column of R, and $s = Wr$ with $s^T \in \mathcal{P}_q$ be the first column of V. Then, it is straightforward to verify using (3.14) that:

$$\begin{aligned} V &= \begin{bmatrix} s & P_1 s & P_2 s & \cdots & P_{q-1}s \end{bmatrix} \\ &= \begin{bmatrix} Wr & P_1 Wr & P_2 Wr & \cdots & P_{q-1}Wr \end{bmatrix} \\ &= W \begin{bmatrix} r & P_1 r & P_2 r & \cdots & P_{q-1}r \end{bmatrix} \end{aligned}$$

where the third equality holds because $\{P_x : x \in \mathcal{X}\}$ are $\mathcal{X}$-circulant matrices which commute with W. Hence, V is the product of W and an $\mathcal{X}$-circulant stochastic matrix, i.e. $V = W\mathsf{circ}_{\mathcal{X}}(z)$ for some $z \in \mathcal{P}_q$. This concludes the proof. An alternative proof is provided in appendix B.5. ■

We emphasize that in Lemma 3.1, the channel that degrades W to produce V is only an additive noise channel without loss of generality. We can certainly have $V = WR$ with a non-additive noise channel R. Consider for instance, $V = W = \mathbf{1}\mathbf{1}^T/q$, where every doubly stochastic matrix R satisfies $V = WR$. However, when we consider $V = WR$ with an additive noise channel R, V corresponds to the channel W with an additional independent additive noise term associated with R. We now prove Proposition 3.6.

Proof of Proposition 3.6.

Part 1: Non-emptiness, closure, and convexity of $\mathcal{L}_W^{\mathsf{add}}$ and $\mathcal{D}_W^{\mathsf{add}}$ can be proved in exactly the same way as in Proposition 3.5, with the additional observation that the set of $\mathcal{X}$-circulant matrices is closed and convex. Moreover, for every $x \in \mathcal{X}$:

$$W \succeq_{\mathsf{ln}} WP_x = \mathsf{circ}_{\mathcal{X}}(wP_x) \succeq_{\mathsf{ln}} W$$
$$W \succeq_{\mathsf{deg}} WP_x = \mathsf{circ}_{\mathcal{X}}(wP_x) \succeq_{\mathsf{deg}} W$$

where the equalities follow from (3.14). These inequalities and the transitive properties of $\succeq_{\mathsf{ln}}$ and $\succeq_{\mathsf{deg}}$ yield the invariance of $\mathcal{L}_W^{\mathsf{add}}$ and $\mathcal{D}_W^{\mathsf{add}}$ with respect to the permutations $\{P_x \in \mathbb{R}^{q\times q} : x \in \mathcal{X}\}$.

Part 2: Lemma 3.1 is equivalent to the fact that $v \in \mathcal{D}_W^{\mathsf{add}}$ if and only if $\mathsf{circ}_{\mathcal{X}}(v) = \mathsf{circ}_{\mathcal{X}}(w)\,\mathsf{circ}_{\mathcal{X}}(z)$ for some $z \in \mathcal{P}_q$. This implies that $v \in \mathcal{D}_W^{\mathsf{add}}$ if and only if $v = w\,\mathsf{circ}_{\mathcal{X}}(z)$ for some $z \in \mathcal{P}_q$ (due to (3.14) and the fact that $\mathcal{X}$-circulant matrices commute). Applying Proposition B.2 from appendix B.1 completes the proof. ■

We remark that part 1 of Proposition 3.6 does not require W to be an additive noise channel. The proofs of closure, convexity, and invariance with respect to $\{P_x \in \mathbb{R}^{q\times q} : x \in \mathcal{X}\}$ hold for general $W \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$. Moreover, $\mathcal{L}_W^{\mathsf{add}}$ and $\mathcal{D}_W^{\mathsf{add}}$ are non-empty because $\mathbf{u} \in \mathcal{L}_W^{\mathsf{add}}$ and $\mathbf{u} \in \mathcal{D}_W^{\mathsf{add}}$.

■ 3.5.2 Less Noisy Domination and Degradation Regions for Symmetric Channels

Since q-ary symmetric channels for $q \in \mathbb{N}$ are additive noise channels, Proposition 3.6 holds for symmetric channels. In this subsection, we deduce some simple results that are unique to symmetric channels. The first of these is a specialization of part 2 of Proposition 3.6 which states that the additive degradation region of a symmetric channel can be characterized by traditional majorization instead of group majorization.

Corollary 3.1 (Degradation Region of Symmetric Channel). *The q-ary symmetric channel $W_\delta = \mathsf{circ}_{\mathcal{X}}(w_\delta) \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ for $\delta \in [0,1]$ has additive degradation region:*

$$\mathcal{D}_{W_\delta}^{\mathsf{add}} = \{v \in \mathcal{P}_q : w_\delta \succeq_{\mathsf{maj}} v\} = \mathsf{conv}\Big(\Big\{w_\delta P_q^k : k \in [q]\Big\}\Big)$$

where $\succeq_{\mathsf{maj}}$ denotes the majorization preorder defined in appendix B.1, and $P_q \in \mathbb{R}^{q\times q}$ is defined in (3.15).

Proof. From part 2 of Proposition 3.6, we have that:

$$\begin{aligned}\mathcal{D}_{W_\delta}^{\mathsf{add}} &= \mathsf{conv}(\{w_\delta P_x : x \in \mathcal{X}\}) = \mathsf{conv}\Big(\Big\{w_\delta P_q^k : k \in [q]\Big\}\Big) \\ &= \mathsf{conv}(\{w_\delta P : P \in \mathbb{R}^{q\times q} \text{ is a permutation matrix}\}) \\ &= \{v \in \mathcal{P}_q : w \succeq_{\mathsf{maj}} v\}\end{aligned}$$

where the second and third equalities hold regardless of the choice of group $(\mathcal{X}, \oplus)$, because the sets of all cyclic or regular permutations of w_δ (see (3.18)) equal $\{w_\delta P_x : x \in \mathcal{X}\}$. The final equality follows from the definition of majorization in appendix B.1. This completes the proof. ■

With this geometric characterization of the additive degradation region, it is easy to find the extremal symmetric channel W_τ that is a degraded version of W_δ for some fixed $\delta \in [0,1]\backslash\{\frac{q-1}{q}\}$. Indeed, we compute τ by using the fact that the noise pmf $w_\tau \in \mathsf{conv}(\{w_\delta P_q^k : k \in \{1,\dots,q-1\}\})$:

$$w_\tau = \sum_{i=1}^{q-1} \lambda_i w_\delta P_q^i \tag{3.41}$$

for some $\lambda_1,\dots,\lambda_{q-1} \in [0,1]$ such that $\lambda_1 + \cdots + \lambda_{q-1} = 1$. Solving (3.41) for τ and $\lambda_1,\dots,\lambda_{q-1}$ yields:

$$\tau = 1 - \frac{\delta}{q-1} \tag{3.42}$$

and $\lambda_1 = \cdots = \lambda_{q-1} = \frac{1}{q-1}$, which means that:

$$w_\tau = \frac{1}{q-1}\sum_{i=1}^{q-1} w_\delta P_q^i\,. \tag{3.43}$$

This is illustrated in Figure 3.2 for the case where $\delta \in (0, \frac{q-1}{q})$ and $\tau > \frac{q-1}{q} > \delta$. For $\delta \in (0, \frac{q-1}{q})$, the symmetric channels that are degraded versions of W_δ are $\{W_\gamma : \gamma \in [\delta,\tau]\}$. In particular, for such $\gamma \in [\delta,\tau]$, $W_\gamma = W_\delta W_\beta$ with $\beta = (\gamma-\delta)/(1-\delta-\frac{\delta}{q-1})$ using the proof of part 5 of Proposition 3.4 in appendix B.3.

In the spirit of comparing symmetric and erasure channels as done in [97] for the binary input case, our next result shows that a q-ary symmetric channel can never be less noisy than a q-ary erasure channel.

Proposition 3.7 (Symmetric Channel $\not\succeq_{\mathsf{ln}}$ Erasure Channel). *For $q \in \mathbb{N}\backslash\{1\}$, given a q-ary erasure channel $E_\epsilon \in \mathbb{R}_{\mathsf{sto}}^{q\times(q+1)}$ with erasure probability $\epsilon \in (0,1)$, there does not exist $\delta \in (0,1)$ such that the corresponding q-ary symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ on the same input alphabet satisfies $W_\delta \succeq_{\mathsf{ln}} E_\epsilon$.*

Proof. For a q-ary erasure channel E_ϵ with $\epsilon \in (0,1)$, we always have $D(\mathbf{u}E_\epsilon||\Delta_0 E_\epsilon) = +\infty$ for $\mathbf{u}, \Delta_0 = (1,0,\ldots,0) \in \mathcal{P}_q$. On the other hand, for any q-ary symmetric channel W_δ with $\delta \in (0,1)$, we have $D(P_X W_\delta||Q_X W_\delta) < +\infty$ for every $P_X, Q_X \in \mathcal{P}_q$. Thus, $W_\delta \not\succeq_{\mathsf{ln}} E_\epsilon$ for any $\delta \in (0,1)$. ■

In fact, the argument for Proposition 3.7 conveys that a symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $\delta \in (0,1)$ satisfies $W_\delta \succeq_{\mathsf{ln}} V$ for some channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ only if the KL divergence $D(P_X V||Q_X V) < +\infty$ for every $P_X, Q_X \in \mathcal{P}_q$. Typically, we are only interested in studying q-ary symmetric channels with $q \geq 2$ and $\delta \in (0, \frac{q-1}{q})$. For example, the BSC with crossover probability δ is usually studied for $\delta \in (0, \frac{1}{2})$. Indeed, the less noisy domination characteristics of the extremal q-ary symmetric channels with $\delta = 0$ or $\delta = \frac{q-1}{q}$ are quite elementary. Given $q \geq 2$, $W_0 = I_q \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ satisfies $W_0 \succeq_{\mathsf{ln}} V$, and $W_{(q-1)/q} = \mathbf{1u} \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ satisfies $V \succeq_{\mathsf{ln}} W_{(q-1)/q}$, for every channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ on a common input alphabet. For the sake of completeness, we also note that for $q \geq 2$, the extremal q-ary erasure channels $E_0 \in \mathbb{R}_{\mathsf{sto}}^{q\times(q+1)}$ and $E_1 \in \mathbb{R}_{\mathsf{sto}}^{q\times(q+1)}$, with $\epsilon = 0$ and $\epsilon = 1$ respectively, satisfy $E_0 \succeq_{\mathsf{ln}} V$ and $V \succeq_{\mathsf{ln}} E_1$ for every channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ on a common input alphabet.

The result that the q-ary symmetric channel with uniform noise pmf $W_{(q-1)/q}$ is more noisy than every channel on the same input alphabet has an analogue concerning AWGN channels. Consider all additive noise channels of the form:

$$Y = X + Z \tag{3.44}$$

where $X, Y \in \mathbb{R}$, the input X is uncorrelated with the additive noise Z: $\mathbb{E}[XZ] = 0$, and the noise Z has power constraint $\mathbb{E}[Z^2] \leq \sigma_Z^2$ for some fixed $\sigma_Z > 0$. Let $X = X_{\mathsf{g}} \sim \mathcal{N}(0, \sigma_X^2)$ for some $\sigma_X > 0$. Then, we have:

$$I(X_{\mathsf{g}}; X_{\mathsf{g}} + Z) \geq I(X_{\mathsf{g}}; X_{\mathsf{g}} + Z_{\mathsf{g}}) \tag{3.45}$$

where $Z_{\mathsf{g}} \sim \mathcal{N}(0, \sigma_Z^2)$, Z_{g} is independent of X_{g}, and equality occurs if and only if $Z = Z_{\mathsf{g}}$ in distribution [230, Section 4.7]. This states that Gaussian noise is the "worst case additive noise" for a Gaussian source. Hence, the AWGN channel is *not more capable* than any other additive noise channel with the same constraints. As a result, the AWGN channel is *not less noisy* than any other additive noise channel with the same constraints (using Proposition 3.3).

■ 3.6 Equivalent Characterizations of Less Noisy Preorder

Having studied the structure of less noisy domination and degradation regions of channels, we now consider the problem of verifying whether a channel W is less noisy than another channel V. Since using Definition 3.2 or Proposition 3.1 directly is difficult, we often start by checking whether V is a degraded version of W. When this fails, we typically resort to verifying van Dijk's condition in Proposition 3.2, cf. [282, Theorem 2]. In this section, we prove the equivalent characterizations of the less noisy preorder in Theorems 3.1 and 3.3, and then present some useful corollaries of van Dijk's condition.

■ 3.6.1 Characterization using Operator Convex f-Divergences

It is well-known that sufficiently smooth f-divergences, such as KL divergence, can be locally approximated by χ^2-divergence, cf. (2.24) in chapter 2. While this approximation sometimes fails globally, see e.g. [12], (2.55) and Proposition 2.6 convey that $\eta_{\mathsf{KL}}(W) = \eta_{\chi^2}(W) = \eta_f(W)$ for any channel $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ and non-linear operator convex function $f : (0,\infty) \to \mathbb{R}$ with $f(1) = 0$. Since η_{KL} characterizes less noisy domination with respect to erasure channels as mentioned in section 3.2 (see e.g. (3.22)), (2.55) and Proposition 2.6 portray that η_f for any non-linear operator convex f, and in particular, η_{χ^2}, also characterize this domination. This begs the question: *Do non-linear operator convex f-divergences, and more specifically, χ^2-divergence, characterize less noisy domination by an arbitrary channel (rather than an erasure channel)?* To answer this question, Theorems 3.1 and 3.3 in subsection 3.3.1 characterize $\succeq_{\mathsf{ln}}$ using non-linear operator convex f-divergences, and hence, also χ^2-divergence (thereby generalizing the results in (2.55) and Proposition 2.6).

We now use Lemma B.2 in appendix B.2 to prove the equivalent characterizations of $\succeq_{\mathsf{ln}}$ using operator convexity in Theorem 3.1.

Proof of Theorem 3.1. Our proof is inspired by the proof technique of [46, Theorem 1] (also see [234, Section III-C]). Fix any non-linear operator convex function $f : (0,\infty) \to \mathbb{R}$ such that $f(1) = 0$, where the non-linearity ensures that the corresponding f-divergence is not identically zero (see the affine invariance property in subsection 2.2.1). For any two channels $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ with the same input alphabet, we first establish that:

$$\forall P_X, Q_X \in \mathcal{P}_q,\ \ \chi^2(P_X W||Q_X W) \geq \chi^2(P_X V||Q_X V) \tag{3.46}$$

if and only if:

$$\forall P_X, Q_X \in \mathcal{P}_q,\ \ D_f(P_X W||Q_X W) \geq D_f(P_X V||Q_X V)\,. \tag{3.47}$$

To prove the forward direction, we utilize Lemma B.2 and the equivalent form of Vincze-Le Cam divergences in (2.12) in chapter 2 to obtain the following integral representation of our f-divergence in terms of χ^2-divergence, cf. [46, p.33]:

$$D_f(P_X||Q_X) = b\,\chi^2(P_X||Q_X) + \int_{(0,1)} \frac{1+\lambda^2}{(1-\lambda)^2}\,\chi^2(P_X||\lambda P_X + (1-\lambda)Q_X)\,d\tau(\lambda) \tag{3.48}$$

for all $P_X, Q_X \in \mathcal{P}_{\mathcal{X}}$, where $b \geq 0$ is some constant and τ is a finite positive measure on $(0,1)$. Since (3.46) holds, we also have:

$$\forall P_X, Q_X \in \mathcal{P}_q,\ \ \chi^2(P_X W||(\lambda P_X + (1-\lambda)Q_X)W) \geq \chi^2(P_X V||(\lambda P_X + (1-\lambda)Q_X)V) \tag{3.49}$$

for every $\lambda \in (0,1)$. Therefore, using (3.46) and (3.49) along with the integral representation in (3.48) yields (3.47), as desired.

To prove the converse direction, observe that Löwner's integral representation in (B.5) (see Lemma B.1 in appendix B.2) ensures that f is infinitely differentiable and $f''(1) > 0$. Since (3.47) holds, we also have:

$$\forall P_X, Q_X \in \mathcal{P}_q, \;\; D_f(((1-\lambda)Q_X + \lambda P_X)W||Q_X W) \geq D_f(((1-\lambda)Q_X + \lambda P_X)V||Q_X V) \tag{3.50}$$

for every $\lambda \in (0,1)$. Therefore, we can scale both sides of (3.50) by $2/(f''(1)\lambda^2) > 0$ and let $\lambda \to 0^+$ so that the local approximation of f-divergences in (2.24) in chapter 2 yields (3.46) for all $P_X \in \mathcal{P}_q$ and all $Q_X \in \mathcal{P}_q^\circ$. Although our version of (2.24) requires the $Q_X \in \mathcal{P}_q^\circ$ assumption, (3.46) also holds for all $Q_X \in \mathcal{P}_q \backslash \mathcal{P}_q^\circ$ due to the continuity of χ^2-divergence in its second argument with fixed first argument.

Now notice that the equivalence between (3.46) and (3.47) illustrates that all channel preorders that are defined using non-linear operator convex f-divergences via (3.47) (in a manner analogous to the characterization of $\succeq_{\mathsf{ln}}$ in Proposition 3.1) are equivalent. Indeed, they are all characterized by χ^2-divergence, cf. (3.46). Since KL divergence is a non-linear operator convex f-divergence (due to part 3 of Theorem B.1 in appendix B.2), $\succeq_{\mathsf{ln}}$ is equivalent to the preorder defined by (3.46), and hence, to the preorder defined by (3.47) for any non-linear operator convex f-divergence. This completes the proof. ■

We next derive Proposition 2.6 from chapter 2 to illustrate that it is a straightforward corollary of Theorem 3.1.

Proof of Proposition 2.6. Fix any non-linear operator convex function $f : (0,\infty) \to \mathbb{R}$ such that $f(1) = 0$, and any channel $P_{Y|X}$ with row stochastic transition probability matrix $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Using Theorem 3.1, the $|\mathcal{X}|$-ary erasure channel $E_{1-\beta}$ with erasure probability $1-\beta \in [0,1]$ is less noisy than $P_{Y|X}$ if and only if for every pair of input pmfs $P_X, Q_X \in \mathcal{P}_{\mathcal{X}}$:

$$D_f(P_X W||Q_X W) \leq D_f(P_X E_{1-\beta}||Q_X E_{1-\beta}) = \beta\, D_f(P_X||Q_X)$$

where the equality is shown near the end of subsection 2.2.2 in chapter 2. This equivalence yields the following generalization of (3.22):

$$\eta_f(P_{Y|X}) = \min\Big\{\beta \in [0,1] : E_{1-\beta} \succeq_{\mathsf{ln}} P_{Y|X}\Big\}. \tag{3.51}$$

Therefore, the contraction coefficients $\eta_f(P_{Y|X})$ for all non-linear operator convex f are equal, and in particular, they all equal $\eta_{\chi^2}(P_{Y|X})$ (since $f(t) = t^2 - 1$ is operator convex; see part 2 of Theorem B.1 in appendix B.2). ■

Finally, we remark that an analogue of (3.22) and (3.51) for degradation by erasure channels is derived in [103, Lemma 4]:

$$\eta_{\mathsf{Doeblin}}(V) \triangleq \sum_{j=1}^{s} \min_{i\in\{1,\ldots,q\}} [V]_{i,j} = \max\{\beta \in [0,1] : E_\beta \succeq_{\mathsf{deg}} V\} \tag{3.52}$$

for every channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times s}$, where $\eta_{\mathsf{Doeblin}}(V)$ is known as *Doeblin's coefficient of ergodicity*, cf. [49, Definition 5.1]. Equation (3.52) demonstrates that the characterization of degradation in (3.8) does not hold for channels with input alphabets of cardinality greater than 2. Indeed, if degradation was characterized by (3.8) for arbitrary input alphabet sizes, then an argument similar to the proof of Proposition 2.6 would yield that $\eta_{\mathsf{Doeblin}}(V) = 1 - \eta_{\mathsf{TV}}(V) = \max\{\beta \in [0,1] : E_\beta \succeq_{\mathsf{deg}} V\}$ (using part 7 of Proposition 2.5 in chapter 2). However, it is well-known that $\eta_{\mathsf{Doeblin}}(V) \neq 1 - \eta_{\mathsf{TV}}(V)$ in general.

■ 3.6.2 Characterization using χ^2-Divergence

Recall the general measure theoretic setup pertinent to Theorem 3.3 from subsection 3.3.1. We now prove Theorem 3.3, which generalizes (2.55) from chapter 2 and illustrates that χ^2-divergence characterizes the less noisy preorder.

Proof of Theorem 3.3. In order to prove the forward direction, we recall the local approximation of KL divergence using χ^2-divergence from [230, Proposition 4.2], which states that for any two probability measures P_X and Q_X on $(\mathcal{X}, \mathcal{F})$:

$$\lim_{\lambda \to 0^+} \frac{2}{\lambda^2} D(\lambda P_X + (1-\lambda)Q_X || Q_X) = \chi^2(P_X || Q_X) \tag{3.53}$$

where both sides of (3.53) are finite or infinite together (cf. (2.24) in chapter 2). Then, we observe that for any two probability measures P_X and Q_X, and any $\lambda \in [0,1]$, we have:

$$D(\lambda P_X W + (1-\lambda)Q_X W || Q_X W) \geq D(\lambda P_X V + (1-\lambda)Q_X V || Q_X V)$$

since $W \succeq_{\mathsf{ln}} V$. Scaling this inequality by $\frac{2}{\lambda^2}$ and letting $\lambda \to 0^+$ produces:

$$\chi^2(P_X W || Q_X W) \geq \chi^2(P_X V || Q_X V)$$

as shown in (3.53). This proves the forward direction.

To establish the converse direction, we recall an integral representation of KL divergence using χ^2-divergence presented in [231, Appendix A.2] (which can be distilled from the argument in [46, Theorem 1], cf. Lemma B.2 in appendix B.2):[43]

$$D(P_X || Q_X) = \int_0^\infty \frac{\chi^2(P_X || Q_X^t)}{t+1}\, dt \tag{3.54}$$

for any two probability measures P_X and Q_X on $(\mathcal{X}, \mathcal{F})$, where $Q_X^t = \frac{t}{1+t} P_X + \frac{1}{t+1} Q_X$ for $t \in [0, \infty)$, and both sides of (3.54) are finite or infinite together (as a close inspection of the proof in [231, Appendix A.2] reveals). Hence, for every P_X and Q_X, we have by assumption:

$$\chi^2(P_X W || Q_X^t W) \geq \chi^2(P_X V || Q_X^t V)$$

[43] Note that [231, Equation (78)] is missing a factor of $\frac{1}{t+1}$ inside the integral.

which implies via (3.54) that:

$$\int_0^\infty \frac{\chi^2(P_X W||Q_X^t W)}{t+1}\,dt \geq \int_0^\infty \frac{\chi^2(P_X V||Q_X^t V)}{t+1}\,dt$$
$$\Rightarrow \quad D(P_X W||Q_X W) \geq D(P_X V||Q_X V)\,.$$

Hence, $W \succeq_{\mathsf{ln}} V$, which completes the proof. ■

■ 3.6.3 Characterizations via the Löwner Partial Order and Spectral Radius

We will use the finite alphabet setup of subsection 3.1.1 for the remaining discussion in this chapter. In the finite alphabet setting, Theorem 3.3 states that $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ is less noisy than $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ if and only if (3.46) holds. This characterization has the flavor of a Löwner partial order condition. Indeed, it is straightforward to verify that for any $P_X \in \mathcal{P}_q$ and $Q_X \in \mathcal{P}_q^\circ$, we can write their χ^2-divergence as:

$$\chi^2(P_X||Q_X) = J_X \mathsf{diag}(Q_X)^{-1} J_X^T \tag{3.55}$$

where $J_X = P_X - Q_X$. Hence, we can express the inequality in (3.46) as:

$$J_X W \mathsf{diag}(Q_X W)^{-1} W^T J_X^T \geq J_X V \mathsf{diag}(Q_X V)^{-1} V^T J_X^T \tag{3.56}$$

for every $J_X = P_X - Q_X$ such that $P_X \in \mathcal{P}_q$ and $Q_X \in \mathcal{P}_q^\circ$. This suggests that (3.46) is equivalent to:

$$W \mathsf{diag}(Q_X W)^{-1} W^T \succeq_{\mathsf{PSD}} V \mathsf{diag}(Q_X V)^{-1} V^T \tag{3.57}$$

for every $Q_X \in \mathcal{P}_q^\circ$. It turns out that (3.57) indeed characterizes $\succeq_{\mathsf{ln}}$, and this is straightforward to prove directly. The next proposition illustrates that (3.57) also follows as a corollary of van Dijk's characterization in Proposition 3.2, and presents an equivalent spectral characterization of $\succeq_{\mathsf{ln}}$.

Proposition 3.8 (Löwner and Spectral Characterizations of $\succeq_{\mathsf{ln}}$). *For any pair of channels $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ on the same input alphabet $[q]$, the following are equivalent:*

1. $W \succeq_{\mathsf{ln}} V$.
2. *For every $P_X \in \mathcal{P}_q^\circ$, we have:*
$$W \mathsf{diag}(P_X W)^{-1} W^T \succeq_{\mathsf{PSD}} V \mathsf{diag}(P_X V)^{-1} V^T.$$
3. *For every $P_X \in \mathcal{P}_q^\circ$, we have $\mathcal{R}(V \mathsf{diag}(P_X V)^{-1} V^T) \subseteq \mathcal{R}(W \mathsf{diag}(P_X W)^{-1} W^T)$ and:*
$$\rho\Big(\big(W \mathsf{diag}(P_X W)^{-1} W^T\big)^\dagger V \mathsf{diag}(P_X V)^{-1} V^T\Big) = 1\,.$$

Proof.

Equivalence between Parts 1 and 2: Recall the functional $F : \mathcal{P}_q \to \mathbb{R}$, $F(P_X) = I(P_X, W_{Y|X}) - I(P_X, V_{Y|X})$ defined in Proposition 3.2, cf. [282, Theorem 2]. Since $F : \mathcal{P}_q \to \mathbb{R}$ is continuous on its domain $\mathcal{P}_q$, and twice differentiable on $\mathcal{P}_q^\circ$, F is concave if and only if its Hessian is negative semidefinite for every $P_X \in \mathcal{P}_q^\circ$ (i.e. $-\nabla^2 F(P_X) \succeq_{\mathsf{PSD}} 0$ for every $P_X \in \mathcal{P}_q^\circ$) [34, Section 3.1.4]. The Hessian matrix of F, $\nabla^2 F : \mathcal{P}_q^\circ \to \mathbb{R}^{q\times q}_{\mathsf{sym}}$, is defined entry-wise for every $x, x' \in [q]$ as:

$$\left[\nabla^2 F(P_X)\right]_{x+1,x'+1} = \frac{\partial^2 F}{\partial P_X(x)\partial P_X(x')}(P_X)\,.$$

Furthermore, a straightforward calculation shows that:

$$\nabla^2 F(P_X) = V\mathsf{diag}(P_X V)^{-1} V^T - W\mathsf{diag}(P_X W)^{-1} W^T$$

for every $P_X \in \mathcal{P}_q^\circ$. (Note that the matrix inverses here are well-defined because $P_X \in \mathcal{P}_q^\circ$). Therefore, F is concave if and only if for every $P_X \in \mathcal{P}_q^\circ$:

$$W\mathsf{diag}(P_X W)^{-1} W^T \succeq_{\mathsf{PSD}} V\mathsf{diag}(P_X V)^{-1} V^T.$$

This establishes the equivalence between parts 1 and 2 due to van Dijk's characterization of $\succeq_{\mathsf{ln}}$ in Proposition 3.2.

Equivalence between Parts 2 and 3: We now derive the spectral characterization of $\succeq_{\mathsf{ln}}$ using part 2. To this end, we recall a well-known fact from [265, Theorem 1 parts (a),(f)] that provides a direct connection between the Löwner partial order and spectral radius (also see [129, Theorem 7.7.3(a)]).

Lemma 3.2 (Löwner Domination and Spectral Radius [265]). *Given positive semidefinite matrices $A, B \in \mathbb{R}^{q\times q}_{\succeq 0}$, $A \succeq_{\mathsf{PSD}} B$ if and only if $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ and $\rho(A^\dagger B) \leq 1$.*

We provide a proof of Lemma 3.2 in appendix B.6 for completeness. Note that when A is invertible, $\rho(A^{-1}B)$ in Lemma 3.2 is the largest *generalized eigenvalue* of the *matrix pencil* (B, A).

Since $W\mathsf{diag}(P_X W)^{-1} W^T$ and $V\mathsf{diag}(P_X V)^{-1} V^T$ are positive semidefinite for every $P_X \in \mathcal{P}_q^\circ$, applying Lemma 3.2 shows that part 2 holds if and only if for every $P_X \in \mathcal{P}_q^\circ$, we have $\mathcal{R}(V\mathsf{diag}(P_X V)^{-1} V^T) \subseteq \mathcal{R}(W\mathsf{diag}(P_X W)^{-1} W^T)$ and:

$$\rho\left(\left(W\mathsf{diag}(P_X W)^{-1} W^T\right)^\dagger V\mathsf{diag}(P_X V)^{-1} V^T\right) \leq 1\,.$$

To prove that this inequality is an equality, for any $P_X \in \mathcal{P}_q^\circ$, define:

$$A = W\mathsf{diag}(P_X W)^{-1} W^T\,,$$
$$B = V\mathsf{diag}(P_X V)^{-1} V^T\,.$$

It suffices to prove that: $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ and $\rho(A^\dagger B) \leq 1$ if and only if $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ and $\rho(A^\dagger B) = 1$. The converse direction is trivial, so we only establish the forward direction. Observe that $P_X A = \mathbf{1}^T$ and $P_X B = \mathbf{1}^T$. This implies that:

$$\mathbf{1}^T A^\dagger B = P_X \left(AA^\dagger\right) B = P_X B = \mathbf{1}^T$$

where $(AA^\dagger)B = B$ because $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ and $AA^\dagger$ is the orthogonal projection matrix onto $\mathcal{R}(A)$. Since $\rho(A^\dagger B) \leq 1$ and $A^\dagger B$ has an eigenvalue of 1, we have $\rho(A^\dagger B) = 1$. Thus, we have proved that part 2 holds if and only if for every $P_X \in \mathcal{P}_q^\circ$, we have $\mathcal{R}(V\mathsf{diag}(P_X V)^{-1} V^T) \subseteq \mathcal{R}(W\mathsf{diag}(P_X W)^{-1} W^T)$ and:

$$\rho\left(\left(W\mathsf{diag}(P_X W)^{-1} W^T\right)^\dagger V\mathsf{diag}(P_X V)^{-1} V^T\right) = 1\,.$$

This completes the proof. ■

The Löwner characterization of $\succeq_{\mathsf{ln}}$ in part 2 of Proposition 3.8, which is essentially a version of the χ^2-divergence characterization in Theorem 3.3 (as explained earlier), will be useful for proving some of our ensuing results. We remark that the equivalence between parts 1 and 2 can be derived by considering several other functionals. For instance, for any fixed pmf $Q_X \in \mathcal{P}_q^\circ$, we may consider the functional $F_2 : \mathcal{P}_q \to \mathbb{R}$ defined by:

$$F_2(P_X) = D(P_X W||Q_X W) - D(P_X V||Q_X V) \tag{3.58}$$

which has Hessian matrix, $\nabla^2 F_2 : \mathcal{P}_q^\circ \to \mathbb{R}_{\mathsf{sym}}^{q\times q}$, $\nabla^2 F_2(P_X) = W\mathsf{diag}(P_X W)^{-1} W^T - V\mathsf{diag}(P_X V)^{-1} V^T$, that does not depend on Q_X. Much like van Dijk's functional F, F_2 is *convex* (for all $Q_X \in \mathcal{P}_q^\circ$) if and only if $W \succeq_{\mathsf{ln}} V$. This is reminiscent of Ahlswede and Gács' technique to prove (2.55) (see chapter 2), where the convexity of a similar functional is established [5].

As another example, for any fixed pmf $Q_X \in \mathcal{P}_q^\circ$, consider the functional $F_3 : \mathcal{P}_q \to \mathbb{R}$ defined by:

$$F_3(P_X) = \chi^2(P_X W||Q_X W) - \chi^2(P_X V||Q_X V) \tag{3.59}$$

which has Hessian matrix, $\nabla^2 F_3 : \mathcal{P}_q^\circ \to \mathbb{R}_{\mathsf{sym}}^{q\times q}$, $\nabla^2 F_3(P_X) = 2\,W\mathsf{diag}(Q_X W)^{-1} W^T - 2\,V\mathsf{diag}(Q_X V)^{-1} V^T$, that does not depend on P_X. Much like F and F_2, F_3 is *convex* for all $Q_X \in \mathcal{P}_q^\circ$ if and only if $W \succeq_{\mathsf{ln}} V$.

Finally, we also mention some specializations of the spectral radius condition in part 3 of Proposition 3.8. If $q \geq r$ and W has full column rank, the expression for spectral radius in the proposition statement can be simplified to:

$$\rho\left((W^\dagger)^T\mathsf{diag}(P_X W)\, W^\dagger V\mathsf{diag}(P_X V)^{-1} V^T\right) = 1 \tag{3.60}$$

using basic properties of the Moore-Penrose pseudoinverse. Moreover, if $q = r$ and W is non-singular, then the Moore-Penrose pseudoinverses in (3.60) can be written as inverses, and the inclusion relation between the ranges in part 3 of Proposition 3.8 is

trivially satisfied (and can be omitted from the proposition statement). We have used the spectral radius condition in this latter setting to (numerically) compute the additive less noisy domination region in Figure 3.2.

■ 3.7 Conditions for Less Noisy Domination over Additive Noise Channels

We now turn our attention to deriving several conditions for determining when q-ary symmetric channels are less noisy than other channels. Our interest in q-ary symmetric channels arises from their analytical tractability; Proposition 3.4 from subsection 3.1.2, Proposition 3.12 from section 3.9, and [95, Theorem 4.5.2] (which conveys that q-ary symmetric channels have uniform capacity achieving input distributions) serve as illustrations of this tractability. We focus on additive noise channels in this section, and on general channels in the next section.

■ 3.7.1 Necessary Conditions

We first present some straightforward necessary conditions for when an additive noise channel $W \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $q \in \mathbb{N}$ is less noisy than another additive noise channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ on an Abelian group $(\mathcal{X}, \oplus)$. These conditions can obviously be specialized for less noisy domination by symmetric channels.

Proposition 3.9 (Necessary Conditions for $\succeq_{\mathsf{ln}}$ Domination over Additive Noise Channels). *Suppose $W = \mathsf{circ}_{\mathcal{X}}(w)$ and $V = \mathsf{circ}_{\mathcal{X}}(v)$ are additive noise channels with noise pmfs $w, v \in \mathcal{P}_q$ such that $W \succeq_{\mathsf{ln}} V$. Then, the following are true:*

1. (Circle Condition) $\|w - \boldsymbol{u}\|_2 \geq \|v - \boldsymbol{u}\|_2$.
2. (Contraction Condition) $\eta_f(W) \geq \eta_f(V)$ *for all non-linear operator convex functions* $f : (0, \infty) \to \mathbb{R}$ *such that* $f(1) = 0$.
3. (Entropy Condition) $H(v) \geq H(w)$, *where* $H : \mathcal{P}_q \to \mathbb{R}$ *is the Shannon entropy function.*

Proof.

Part 1: Letting $P_X = \Delta_0 = (1, 0, \ldots, 0)$ and $Q_X = \mathbf{u}$ in the χ^2-divergence characterization of $\succeq_{\mathsf{ln}}$ in Theorem 3.3 produces:

$$q \|w - \mathbf{u}\|_2^2 = \chi^2(w||\mathbf{u}) \geq \chi^2(v||\mathbf{u}) = q \|v - \mathbf{u}\|_2^2$$

since $\mathbf{u}W = \mathbf{u}V = \mathbf{u}$, and $\Delta_0 W = w$ and $\Delta_0 V = v$ using (3.14). Hence, we have $\|w - \mathbf{u}\|_2 \geq \|v - \mathbf{u}\|_2$. An alternative proof of this result using Fourier analysis is given in appendix B.7.

Part 2: This easily follows from Theorem 3.1 and Definition 2.5 (in chapter 2).

Part 3: Letting $P_X = \Delta_0 = (1, 0, \ldots, 0)$ and $Q_X = \mathbf{u}$ in the KL divergence characterization of $\succeq_{\mathsf{ln}}$ in Proposition 3.1 produces:

$$\log(q) - H(w) = D(w||\mathbf{u}) \geq D(v||\mathbf{u}) = \log(q) - H(v)$$

via the same reasoning as part 1. This completes the proof. ■

We remark that the aforementioned necessary conditions have many generalizations. Firstly, if $W, V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ are doubly stochastic matrices, then the generalized circle condition holds:

$$\left\|W - W_{\frac{q-1}{q}}\right\|_{\mathsf{Fro}} \geq \left\|V - W_{\frac{q-1}{q}}\right\|_{\mathsf{Fro}} \tag{3.61}$$

where $W_{(q-1)/q} = \mathbf{1u}$ is the q-ary symmetric channel whose conditional pmfs are all uniform. Indeed, letting $P_X = \Delta_x$ for $x \in [q]$ in the proof of part 1 and then adding the inequalities corresponding to every $x \in [q]$ produces (3.61). Secondly, the contraction condition in Proposition 3.9 actually holds for any pair of general channels $W \in \mathbb{R}_{\mathsf{sto}}^{q\times r}$ and $V \in \mathbb{R}_{\mathsf{sto}}^{q\times s}$ on a common input alphabet (not necessarily additive noise channels). Moreover, we can start with Theorem 3.1 and use Definition 2.2 (from chapter 2) to get another variant of the contraction condition for general channels:

$$\eta_f(P_X, W) \geq \eta_f(P_X, V) \tag{3.62}$$

for all $P_X \in \mathcal{P}_q$, and all non-linear operator convex f-divergences. (In particular, the χ^2-divergence case of this inequality yields a maximal correlation condition via (2.37).)

■ 3.7.2 Sufficient Conditions

We next portray a sufficient condition for when an additive noise channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ is a degraded version of a symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$. By Proposition 3.3, this is also a sufficient condition for $W_\delta \succeq_{\mathsf{ln}} V$.

Proposition 3.10 (Degradation by Symmetric Channels). *Given an additive noise channel $V = \mathsf{circ}_{\mathcal{X}}(v)$ with noise pmf $v \in \mathcal{P}_q$ and minimum probability entry $\tau = \min\{[V]_{i,j} : i, j \in \{1, \ldots, q\}\}$, we have:*

$$0 \leq \delta \leq (q-1)\,\tau \quad \Rightarrow \quad W_\delta \succeq_{\mathsf{deg}} V$$

where $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ is a q-ary symmetric channel.

Proof. Using Corollary 3.1, it suffices to prove that the noise pmf $w_{(q-1)\tau} \succeq_{\mathsf{maj}} v$. Since $0 \leq \tau \leq 1/q$, we must have $0 \leq (q-1)\tau \leq (q-1)/q$. So, all entries of $w_{(q-1)\tau}$, except (possibly) the first, are equal to its minimum entry of τ. As $v \geq \tau$ (entry-wise), $w_{(q-1)\tau} \succeq_{\mathsf{maj}} v$ because the conditions of part 3 in Proposition B.1 in appendix B.1 are satisfied. ■

It is compelling to find a sufficient condition for $W_\delta \succeq_{\mathsf{ln}} V$ that does not simply ensure $W_\delta \succeq_{\mathsf{deg}} V$ (such as Proposition 3.10 and Theorem 3.4). The ensuing proposition elucidates such a sufficient condition for additive noise channels. The general strategy for finding such a condition for additive noise channels is to identify a noise pmf that belongs to $\mathcal{L}_{W_\delta}^{\mathsf{add}} \backslash \mathcal{D}_{W_\delta}^{\mathsf{add}}$. One can then use Proposition 3.6 to explicitly construct a set of noise pmfs that is a subset of $\mathcal{L}_{W_\delta}^{\mathsf{add}}$ but strictly includes $\mathcal{D}_{W_\delta}^{\mathsf{add}}$. The proof of Proposition 3.11 finds such a noise pmf (that corresponds to a q-ary symmetric channel).

Proposition 3.11 (Less Noisy Domination by Symmetric Channels). *Given an additive noise channel $V = \mathsf{circ}_{\mathcal{X}}(v)$ with noise pmf $v \in \mathcal{P}_q$ and $q \geq 2$, if for $\delta \in [0, \frac{q-1}{q}]$ we have:*

$$v \in \mathsf{conv}\Big(\Big\{w_\delta P_q^k : k \in [q]\Big\} \cup \Big\{w_\gamma P_q^k : k \in [q]\Big\}\Big)$$

then $W_\delta \succeq_{\mathsf{ln}} V$, where $P_q \in \mathbb{R}^{q\times q}$ is defined in (3.15), and:

$$\gamma = \frac{1-\delta}{1-\delta+\frac{\delta}{(q-1)^2}} \in \left[1-\frac{\delta}{q-1}, 1\right].$$

Proof. Due to Proposition 3.6 and $\{w_\gamma P_x : x \in \mathcal{X}\} = \{w_\gamma P_q^k : k \in [q]\}$, it suffices to prove that $W_\delta \succeq_{\mathsf{ln}} W_\gamma$. Since $\delta = 0 \Rightarrow \gamma = 1$ and $\delta = \frac{q-1}{q} \Rightarrow \gamma = \frac{q-1}{q}$, $W_\delta \succeq_{\mathsf{ln}} W_\gamma$ is certainly true for $\delta \in \{0, \frac{q-1}{q}\}$. So, we assume that $\delta \in (0, \frac{q-1}{q})$, which implies that:

$$\gamma = \frac{1-\delta}{1-\delta+\frac{\delta}{(q-1)^2}} \in \left(\frac{q-1}{q}, 1\right).$$

Since our goal is to show $W_\delta \succeq_{\mathsf{ln}} W_\gamma$, we prove the equivalent condition in part 2 of Proposition 3.8 that for every $P_X \in \mathcal{P}_q^\circ$:

$$\begin{aligned}
& W_\delta\, \mathsf{diag}(P_X W_\delta)^{-1}\, W_\delta^T \succeq_{\mathsf{PSD}} W_\gamma\, \mathsf{diag}(P_X W_\gamma)^{-1}\, W_\gamma^T \\
\Leftrightarrow\quad & W_\gamma^{-1}\mathsf{diag}(P_X W_\gamma)\, W_\gamma^{-1} \succeq_{\mathsf{PSD}} W_\delta^{-1}\mathsf{diag}(P_X W_\delta)\, W_\delta^{-1} \\
\Leftrightarrow\quad & \mathsf{diag}(P_X W_\gamma) \succeq_{\mathsf{PSD}} W_\gamma W_\delta^{-1}\mathsf{diag}(P_X W_\delta)\, W_\delta^{-1} W_\gamma \\
\Leftrightarrow\quad & I_q \succeq_{\mathsf{PSD}} \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}}\, W_\tau \mathsf{diag}(P_X W_\delta)\, W_\tau \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}} \\
\Leftrightarrow\quad & 1 \geq \left\| \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}}\, W_\tau \mathsf{diag}(P_X W_\delta)\, W_\tau \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}} \right\|_{\mathsf{op}} \\
\Leftrightarrow\quad & 1 \geq \left\| \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}}\, W_\tau \mathsf{diag}(P_X W_\delta)^{\frac{1}{2}} \right\|_{\mathsf{op}}
\end{aligned}$$

where the second equivalence holds because W_δ and W_γ are symmetric and invertible (see part 4 of Proposition 3.4 and [129, Corollary 7.7.4]), and the third and fourth equivalences are non-singular $*$-congruences with $W_\tau = W_\delta^{-1} W_\gamma = W_\gamma W_\delta^{-1}$ and:

$$\tau = \frac{\gamma - \delta}{1 - \delta - \frac{\delta}{q-1}} > 0$$

which can be computed as in the proof of Proposition B.4 in appendix B.9.[44]

It is instructive to note that if $W_\tau \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$, then the adjoint of the DTM (see (2.38) in chapter 2), $\mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}}\, W_\tau \mathsf{diag}(P_X W_\delta)^{\frac{1}{2}}$, has right singular vector $\sqrt{P_X W_\delta}^T$ and

[44] Note that we cannot use the strict Löwner partial order $\succ_{\mathsf{PSD}}$ (recall that for $A, B \in \mathbb{R}_{\mathsf{sym}}^{q\times q}$, $A \succ_{\mathsf{PSD}} B$ if and only if $A - B$ is positive definite) for these equivalences as $\mathbf{1}^T W_\gamma^{-1}\mathsf{diag}(P_X W_\gamma)\, W_\gamma^{-1}\mathbf{1} = \mathbf{1}^T W_\delta^{-1}\mathsf{diag}(P_X W_\delta)\, W_\delta^{-1}\mathbf{1}$.

left singular vector $\sqrt{P_X W_\gamma}^T$ corresponding to its maximum singular value of unity (see the proof of Proposition 2.2 in appendix A.1). So, $W_\tau \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ is a sufficient condition for $W_\delta \succeq_{\mathsf{ln}} W_\gamma$. Since $W_\tau \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ if and only if $0 \leq \tau \leq 1$ if and only if $\delta \leq \gamma \leq 1 - \frac{\delta}{q-1}$, the latter condition also implies that $W_\delta \succeq_{\mathsf{ln}} W_\gamma$. However, we recall from (3.42) in subsection 3.5.2 that $W_\delta \succeq_{\mathsf{deg}} W_\gamma$ for $\delta \leq \gamma \leq 1 - \frac{\delta}{q-1}$, while we seek some $1 - \frac{\delta}{q-1} < \gamma \leq 1$ for which $W_\delta \succeq_{\mathsf{ln}} W_\gamma$. When $q = 2$, we only have:

$$\gamma = \frac{1-\delta}{1-\delta+\frac{\delta}{(q-1)^2}} = 1 - \frac{\delta}{q-1} = 1 - \delta$$

which implies that $W_\delta \succeq_{\mathsf{deg}} W_\gamma$ is true for $q = 2$. On the other hand, when $q \geq 3$, it is straightforward to verify that:

$$\gamma = \frac{1-\delta}{1-\delta+\frac{\delta}{(q-1)^2}} \in \left(1 - \frac{\delta}{q-1}, 1\right)$$

since $\delta \in (0, \frac{q-1}{q})$.

From the preceding discussion, it suffices to prove for $q \geq 3$ that for every $P_X \in \mathcal{P}_q^\circ$:

$$\left\| \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}} W_\tau \mathsf{diag}(P_X W_\delta) W_\tau \mathsf{diag}(P_X W_\gamma)^{-\frac{1}{2}} \right\|_{\mathsf{op}} \leq 1 \,.$$

Since $\tau > 0$, and $0 \leq \tau \leq 1$ does not produce $\gamma > 1 - \frac{\delta}{q-1}$, we require that $\tau > 1$ ($\Leftrightarrow \gamma > 1 - \frac{\delta}{q-1}$) so that W_τ has strictly negative entries along the diagonal. Notice that:

$$\forall x \in [q], \;\; \mathsf{diag}(\Delta_x W_\gamma) \succeq_{\mathsf{PSD}} W_\gamma W_\delta^{-1} \mathsf{diag}(\Delta_x W_\delta) W_\delta^{-1} W_\gamma$$

implies that:

$$\forall P_X \in \mathcal{P}_q^\circ, \;\; \mathsf{diag}(P_X W_\gamma) \succeq_{\mathsf{PSD}} W_\gamma W_\delta^{-1} \mathsf{diag}(P_X W_\delta) W_\delta^{-1} W_\gamma$$

because convex combinations preserve the Löwner relation. So, it suffices to prove that for every $x \in [q]$:

$$\left\| \mathsf{diag}\left(w_\gamma P_q^x\right)^{-\frac{1}{2}} W_\tau \mathsf{diag}\left(w_\delta P_q^x\right) W_\tau \mathsf{diag}\left(w_\gamma P_q^x\right)^{-\frac{1}{2}} \right\|_{\mathsf{op}} \leq 1$$

where $P_q \in \mathbb{R}^{q\times q}$ is defined in (3.15), because $\Delta_x M$ extracts the $(x+1)$th row of a matrix $M \in \mathbb{R}^{q\times q}$. Let us define:

$$A_x \triangleq \mathsf{diag}\left(w_\gamma P_q^x\right)^{-\frac{1}{2}} W_\tau \mathsf{diag}\left(w_\delta P_q^x\right) W_\tau \mathsf{diag}\left(w_\gamma P_q^x\right)^{-\frac{1}{2}}$$

for each $x \in [q]$. Observe that for every $x \in [q]$, $A_x \in \mathbb{R}^{q\times q}_{\succeq 0}$ is orthogonally diagonalizable by the real spectral theorem [17, Theorem 7.13], and has a strictly positive eigenvector $\sqrt{w_\gamma P_q^x}$ corresponding to the eigenvalue of unity:

$$\forall x \in [q], \;\; \sqrt{w_\gamma P_q^x} A_x = \sqrt{w_\gamma P_q^x}$$

so that all other eigenvectors of A_x have some strictly negative entries since they are orthogonal to $\sqrt{w_\gamma P_q^x}$. Suppose A_x is entry-wise non-negative for every $x \in [q]$. Then, the largest eigenvalue (known as the Perron-Frobenius eigenvalue) and the spectral radius of each A_x is unity by the *Perron-Frobenius theorem* [129, Theorem 8.3.4], which proves that $\|A_x\|_{\mathsf{op}} \leq 1$ for every $x \in [q]$. Therefore, it is sufficient to prove that A_x is entry-wise non-negative for every $x \in [q]$. Equivalently, we can prove that $W_\tau \mathsf{diag}(w_\delta P_q^x) W_\tau$ is entry-wise non-negative for every $x \in [q]$, since $\mathsf{diag}(w_\gamma P_q^x)^{-1/2}$ scales the rows or columns of the matrix it is pre- or post-multiplied with using strictly positive scalars.

We now show the equivalent condition below that the minimum possible entry of $W_\tau \mathsf{diag}(w_\delta P_q^x) W_\tau$ is non-negative:

$$\begin{aligned} 0 \leq \min_{\substack{x \in [q] \\ i,j \in \{1,\ldots,q\}}} & \underbrace{\sum_{r=1}^{q} [W_\tau]_{i,r} [W_\delta]_{x+1,r} [W_\tau]_{r,j}}_{= \left[W_\tau \mathsf{diag}\left(w_\delta P_q^x\right) W_\tau\right]_{i,j}} \\ &= \frac{\tau(1-\delta)(1-\tau)}{q-1} + \frac{\delta\tau(1-\tau)}{(q-1)^2} + (q-2)\frac{\delta\tau^2}{(q-1)^3} \,. \end{aligned} \tag{3.63}$$

The above equality holds because for $i \neq j$:

$$\frac{\delta}{q-1} \sum_{r=1}^{q} \underbrace{[W_\tau]_{i,r} [W_\tau]_{r,i}}_{= [W_\tau]_{i,r}^2 \geq 0} \geq \frac{\delta}{q-1} \sum_{r=1}^{q} [W_\tau]_{i,r} [W_\tau]_{r,j}$$

is clearly true (using, for example, the rearrangement inequality in [121, Section 10.2]), and adding $\left(1 - \delta - \frac{\delta}{q-1}\right) [W_\tau]_{i,k}^2 \geq 0$ (regardless of the value of $1 \leq k \leq q$) to the left summation increases its value, while adding $\left(1 - \delta - \frac{\delta}{q-1}\right) [W_\tau]_{i,p} [W_\tau]_{p,j} < 0$ (which exists for an appropriate value $1 \leq p \leq q$ as $\tau > 1$) to the right summation decreases its value. As a result, the minimum possible entry of $W_\tau \mathsf{diag}(w_\delta P_q^x) W_\tau$ can be achieved with $x + 1 = i \neq j$ or $i \neq j = x + 1$. We next substitute $\tau = (\gamma - \delta)/\left(1 - \delta - \frac{\delta}{q-1}\right)$ into (3.63) and simplify the resulting expression to get:

$$0 \leq (\gamma - \delta) \left(\left(1 - \frac{\delta}{q-1} - \gamma\right) \left(1 - \delta + \frac{\delta}{q-1}\right) + \frac{(q-2)\,\delta\,(\gamma - \delta)}{(q-1)^2} \right).$$

The right hand side of this inequality is quadratic in γ with roots $\gamma = \delta$ and $\gamma = \frac{1-\delta}{1-\delta+(\delta/(q-1)^2)}$. Since the coefficient of γ^2 in this quadratic is strictly negative:

$$\underbrace{\frac{(q-2)\,\delta}{(q-1)^2} - \left(1 - \delta + \frac{\delta}{q-1}\right)}_{\text{coefficient of } \gamma^2} < 0 \quad \Leftrightarrow \quad 1 - \delta + \frac{\delta}{(q-1)^2} > 0$$

the minimum possible entry of $W_\tau \mathsf{diag}(w_\delta P_q^x) W_\tau$ is non-negative if and only if:

$$\delta \leq \gamma \leq \frac{1-\delta}{1 - \delta + \frac{\delta}{(q-1)^2}}$$

where we use the fact that $\frac{1-\delta}{1-\delta+(\delta/(q-1)^2)} \geq 1 - \frac{\delta}{q-1} \geq \delta$. Therefore, $\gamma = \frac{1-\delta}{1-\delta+(\delta/(q-1)^2)}$ produces $W_\delta \succeq_{\mathsf{ln}} W_\gamma$, which completes the proof. ■

Heretofore we have derived results concerning less noisy domination and degradation regions in section 3.5, and proven several necessary and sufficient conditions for less noisy domination of additive noise channels by symmetric channels in this section. We finally have all the pieces in place to establish Theorem 3.5 from section 3.3. In closing this section, we indicate the pertinent results that coalesce to justify it.

Proof of Theorem 3.5. The first equality follows from Corollary 3.1. The first set inclusion is obvious, and its strictness follows from the proof of Proposition 3.11. The second set inclusion follows from Proposition 3.11. The third set inclusion follows from the circle condition (part 1) in Proposition 3.9. Lastly, the properties of $\mathcal{L}^{\mathsf{add}}_{W_\delta}$ are derived in Proposition 3.6. ■

■ 3.8 Sufficient Conditions for Degradation over General Channels

While Propositions 3.10 and 3.11 present sufficient conditions for a symmetric channel $W_\delta \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ to be less noisy than an additive noise channel, our more comprehensive objective is to find the maximum $\delta \in \left[0, \frac{q-1}{q}\right]$ such that $W_\delta \succeq_{\mathsf{ln}} V$ for any given general channel $V \in \mathbb{R}^{q\times r}_{\mathsf{sto}}$ on a common input alphabet. We may formally define this maximum δ (that characterizes the extremal symmetric channel that is less noisy than V) as:

$$\delta^\star(V) \triangleq \sup\left\{\delta \in \left[0, \frac{q-1}{q}\right] : W_\delta \succeq_{\mathsf{ln}} V\right\} \tag{3.64}$$

and for every $0 \leq \delta < \delta^\star(V)$, $W_\delta \succeq_{\mathsf{ln}} V$. Alternatively, we can define a non-negative (less noisy) *domination factor* function for any channel $V \in \mathbb{R}^{q\times r}_{\mathsf{sto}}$:

$$\mu_V(\delta) \triangleq \sup_{\substack{P_X, Q_X \in \mathcal{P}_q: \\ 0<D(P_XW_\delta||Q_XW_\delta)<+\infty}} \frac{D(P_XV||Q_XV)}{D(P_XW_\delta||Q_XW_\delta)} \geq 0 \tag{3.65}$$

with $\delta \in \left[0, \frac{q-1}{q}\right)$, which is analogous to the contraction coefficient for KL divergence since $\mu_V(0) \triangleq \eta_{\mathsf{KL}}(V)$. Indeed, we may perceive P_XW_δ and Q_XW_δ in the denominator of (3.65) as pmfs inside the "shrunk" simplex $\mathsf{conv}(\{w_\delta P_q^k : k \in [q]\})$, and (3.65) represents a contraction coefficient of V where the supremum is taken over this "shrunk" simplex.[45] For simplicity, consider a channel $V \in \mathbb{R}^{q\times r}_{\mathsf{sto}}$ that is strictly positive entry-wise, and has domination factor function $\mu_V : \left(0, \frac{q-1}{q}\right) \to \mathbb{R}$, where the domain excludes zero because μ_V is only interesting for $\delta \in \left(0, \frac{q-1}{q}\right)$, and this exclusion also affords us some analytical simplicity. It is shown in Proposition B.4 in appendix B.9 that μ_V is always finite

[45] Pictorially, the "shrunk" simplex is the magenta triangle in Figure 3.2 while the simplex itself is the larger gray triangle.

on $(0, \frac{q-1}{q})$, continuous, convex, strictly increasing, and has a vertical asymptote at $\delta = \frac{q-1}{q}$. Since for every $P_X, Q_X \in \mathcal{P}_q$:

$$\mu_V(\delta)\, D(P_X W_\delta || Q_X W_\delta) \geq D(P_X V || Q_X V) \tag{3.66}$$

we have $\mu_V(\delta) \leq 1$ if and only if $W_\delta \succeq_{\mathsf{ln}} V$. Hence, using the strictly increasing property of $\mu_V : (0, \frac{q-1}{q}) \to \mathbb{R}$, we can also characterize $\delta^\star(V)$ as:

$$\delta^\star(V) = \mu_V^{-1}(1) \tag{3.67}$$

where μ_V^{-1} denotes the inverse function of μ_V, and unity is in the range of μ_V by Theorem 3.4 since V is strictly positive entry-wise.

We next briefly delineate how one might computationally approximate $\delta^\star(V)$ for a given general channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times r}$. From part 2 of Proposition 3.8, it is straightforward to obtain the following minimax characterization of $\delta^\star(V)$:

$$\delta^\star(V) = \inf_{P_X \in \mathcal{P}_q^\circ} \sup_{\delta \in \mathcal{S}(P_X)} \delta \tag{3.68}$$

where $\mathcal{S}(P_X) = \{\delta \in [0, \frac{q-1}{q}] : W_\delta\, \mathsf{diag}(P_X W_\delta)^{-1}\, W_\delta^T \succeq_{\mathsf{PSD}} V \mathsf{diag}(P_X V)^{-1}\, V^T\}$. The infimum in (3.68) can be naïvely approximated by sampling several $P_X \in \mathcal{P}_q^\circ$. The supremum in (3.68) can be estimated by verifying collections of rational (ratio of polynomials) inequalities in δ. This is because the positive semidefiniteness of a matrix is equivalent to the non-negativity of all its principal minors by *Sylvester's criterion* [129, Theorem 7.2.5]. Unfortunately, this procedure appears to be rather cumbersome.

Since analytically computing $\delta^\star(V)$ also seems intractable, we now prove Theorem 3.4 from section 3.3. Theorem 3.4 provides a sufficient condition for $W_\delta \succeq_{\mathsf{deg}} V$ (which implies $W_\delta \succeq_{\mathsf{ln}} V$ using Proposition 3.3) by restricting its attention to the case where $V \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ with $q \geq 2$. Moreover, it can be construed as a lower bound on $\delta^\star(V)$:

$$\delta^\star(V) \geq \frac{\nu}{1 - (q-1)\nu + \frac{\nu}{q-1}} \tag{3.69}$$

where $\nu = \min\{[V]_{i,j} : i, j \in \{1, \ldots, q\}\}$ is the minimum conditional probability in V.

Proof of Theorem 3.4. Let the channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ be consisted of the conditional pmfs $v_1, \ldots, v_q \in \mathcal{P}_q$ as its rows:

$$V = \begin{bmatrix} v_1^T & v_2^T & \cdots & v_q^T \end{bmatrix}^T.$$

From the proof of Proposition 3.10, we know that $w_{(q-1)\nu} \succeq_{\mathsf{maj}} v_i$ for every $i \in \{1, \ldots, q\}$. Using part 1 of Proposition B.1 in appendix B.1 (and the fact that the set of all permutations of $w_{(q-1)\nu}$ is exactly the set of all cyclic permutations of $w_{(q-1)\nu}$), we can write this as:

$$\forall i \in \{1, \ldots, q\}, \quad v_i = \sum_{j=1}^{q} p_{i,j}\, w_{(q-1)\nu} P_q^{j-1}$$

where the matrix $P_q \in \mathbb{R}^{q\times q}$ is given in (3.15), and $\{p_{i,j} \geq 0 : i,j \in \{1,\ldots,q\}\}$ are the convex weights such that $\sum_{j=1}^{q} p_{i,j} = 1$ for every $i \in \{1,\ldots,q\}$. Defining $P \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ entry-wise as $[P]_{i,j} = p_{i,j}$ for every $1 \leq i,j \leq q$, we can also write this equation as $V = PW_{(q-1)\nu}$.[46] Observe that:

$$P = \sum_{1\leq j_1,\ldots,j_q\leq q} \left(\prod_{i=1}^{q} p_{i,j_i}\right) E_{j_1,\ldots,j_q}$$

where $\{\prod_{i=1}^{q} p_{i,j_i} : j_1,\ldots,j_q \in \{1,\ldots,q\}\}$ form a product pmf of convex weights, and for every $1 \leq j_1,\ldots,j_q \leq q$:

$$E_{j_1,\ldots,j_q} \triangleq \begin{bmatrix} e_{j_1} & e_{j_2} & \cdots & e_{j_q} \end{bmatrix}^T$$

where $e_i \in \mathbb{R}^q$ is the ith standard basis vector for each $i \in \{1,\ldots,q\}$. Hence, we get:

$$V = \sum_{1\leq j_1,\ldots,j_q\leq q} \left(\prod_{i=1}^{q} p_{i,j_i}\right) E_{j_1,\ldots,j_q} W_{(q-1)\nu}\,.$$

Suppose there exists $\delta \in \left[0, \frac{q-1}{q}\right]$ such that for all $j_1,\ldots,j_q \in \{1,\ldots,q\}$:

$$\exists M_{j_1,\ldots,j_q} \in \mathbb{R}_{\mathsf{sto}}^{q\times q},\ \ E_{j_1,\ldots,j_q} W_{(q-1)\nu} = W_\delta M_{j_1,\ldots,j_q}$$

i.e. $W_\delta \succeq_{\mathsf{deg}} E_{j_1,\ldots,j_q} W_{(q-1)\nu}$. Then, we would have:

$$V = W_\delta \underbrace{\sum_{1\leq j_1,\ldots,j_q\leq q} \left(\prod_{i=1}^{q} p_{i,j_i}\right) M_{j_1,\ldots,j_q}}_{\text{stochastic matrix}}$$

which implies that $W_\delta \succeq_{\mathsf{deg}} V$.

We will demonstrate that for every $j_1,\ldots,j_q \in \{1,\ldots,q\}$, there exists $M_{j_1,\ldots,j_q} \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ such that $E_{j_1,\ldots,j_q} W_{(q-1)\nu} = W_\delta M_{j_1,\ldots,j_q}$ when $0 \leq \delta \leq \nu/(1-(q-1)\nu+\frac{\nu}{q-1})$. Since $0 \leq \nu \leq \frac{1}{q}$, the preceding inequality implies that $0 \leq \delta \leq \frac{q-1}{q}$, where $\delta = \frac{q-1}{q}$ is possible if and only if $\nu = \frac{1}{q}$. When $\nu = \frac{1}{q}$, $V = W_{(q-1)/q}$ is the channel with all uniform conditional pmfs, and $W_{(q-1)/q} \succeq_{\mathsf{deg}} V$ clearly holds. Hence, we assume that $0 \leq \nu < \frac{1}{q}$ so that $0 \leq \delta < \frac{q-1}{q}$, and establish the equivalent condition that for every $j_1,\ldots,j_q \in \{1,\ldots,q\}$:

$$M_{j_1,\ldots,j_q} = W_\delta^{-1} E_{j_1,\ldots,j_q} W_{(q-1)\nu}$$

[46] Matrices of the form $V = PW_{(q-1)\nu}$ with $P \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ are *not* necessarily degraded versions of $W_{(q-1)\nu}$: $W_{(q-1)\nu} \not\succeq_{\mathsf{deg}} V$ (although we certainly have input-output degradation: $W_{(q-1)\nu} \succeq_{\mathsf{iod}} V$). As a counterexample, consider $W_{1/2}$ for $q = 3$, and $P = [1\,0\,0;\ 1\,0\,0;\ 0\,1\,0]$, where the semicolons separate the rows of the matrix. If $W_{1/2} \succeq_{\mathsf{deg}} PW_{1/2}$, then there exists $A \in \mathbb{R}_{\mathsf{sto}}^{3\times 3}$ such that $PW_{1/2} = W_{1/2}A$. However, $A = W_{1/2}^{-1}PW_{1/2} = (1/4)\,[3\,0\,1;\ 3\,0\,1;\ -1\,4\,1]$ has a strictly negative entry, which leads to a contradiction.

is a valid stochastic matrix. Recall that $W_\delta^{-1} = W_\tau$ with $\tau = \frac{-\delta}{1-\delta-(\delta/(q-1))}$ using part 4 of Proposition 3.4. Clearly, all the rows of each $M_{j_1,\ldots,j_q}$ sum to unity. So, it remains to verify that each $M_{j_1,\ldots,j_q}$ has non-negative entries. For any $j_1,\ldots,j_q \in \{1,\ldots,q\}$ and any $i,j \in \{1,\ldots,q\}$:

$$\left[M_{j_1,\ldots,j_q}\right]_{i,j} \geq \nu\,(1-\tau) + \tau\,(1-(q-1)\,\nu)$$

where the right hand side is the minimum possible entry of any $M_{j_1,\ldots,j_q}$ (with equality when $j_1 > 1$ and $j_2 = j_3 = \cdots = j_q = 1$ for example) as $\tau < 0$ and $1-(q-1)\,\nu > \nu$. To ensure each $M_{j_1,\ldots,j_q}$ is entry-wise non-negative, the minimum possible entry must satisfy:

$$\begin{aligned} & \nu\,(1-\tau) + \tau\,(1-(q-1)\,\nu) \geq 0 \\ \Leftrightarrow \quad & \nu + \frac{\delta\nu}{1-\delta-\frac{\delta}{q-1}} - \frac{\delta\,(1-(q-1)\,\nu)}{1-\delta-\frac{\delta}{q-1}} \geq 0 \end{aligned}$$

and the latter inequality is equivalent to:

$$\delta \leq \frac{\nu}{1-(q-1)\,\nu + \frac{\nu}{q-1}}\,.$$

This completes the proof. ■

We remark that if $V = E_{2,1,\ldots,1}W_{(q-1)\nu} \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$, then this proof illustrates that $W_\delta \succeq_{\mathsf{deg}} V$ if and only if $0 \leq \delta \leq \nu/(1-(q-1)\nu+\frac{\nu}{q-1})$. Hence, the condition in Theorem 3.4 is tight when no further information about V is known. However, if further information is available, then other sufficient conditions can be derived, cf. [215, Proposition 8.1, Equations (8.1) and (8.2)]. It is worth juxtaposing Theorem 3.4 and Proposition 3.10. The upper bounds on δ from these results satisfy:

$$\underbrace{\frac{\nu}{1-(q-1)\nu+\frac{\nu}{q-1}}}_{\text{upper bound in Theorem 3.4}} \quad \leq \quad \underbrace{(q-1)\,\nu}_{\substack{\text{upper bound in} \\ \text{Proposition 3.10}}} \tag{3.70}$$

where we have equality if and only if $\nu = 1/q$, and it is straightforward to verify that (3.70) is equivalent to $\nu \leq 1/q$. Moreover, assuming that q is large and $\nu = o(1/q)$, the upper bound in Theorem 3.4 is $\nu/(1+o(1)+o(1/q^2)) = \Theta(\nu)$, while the upper bound in Proposition 3.10 is $\Theta(q\nu)$. (Note that both bounds are $\Theta(1)$ if $\nu = 1/q$.) Therefore, when $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ is an additive noise channel, $\delta = O(q\nu)$ is enough for $W_\delta \succeq_{\mathsf{deg}} V$, but a general channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ requires $\delta = O(\nu)$ for such degradation. So, in order to account for q different conditional pmfs in the general case (as opposed to a single conditional pmf which characterizes the channel in the additive noise case), we loose a factor of q in the upper bound on δ. Furthermore, we can check using simulations that $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ is not in general less noisy than $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ for $\delta = (q-1)\nu$. Indeed, counterexamples can be easily obtained by letting $V = E_{j_1,\ldots,j_q}W_\delta$ for specific values

of $1 \leq j_1, \ldots, j_q \leq q$, and computationally verifying that $W_\delta \not\succeq_{\mathsf{ln}} V + J \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ for appropriate choices of perturbation matrices $J \in \mathbb{R}^{q \times q}$ with sufficiently small Frobenius norm.

We have now proved Theorems 3.3, 3.4, and 3.5 from section 3.3. The next section relates our results regarding less noisy and degradation preorders to LSIs, and proves Theorem 3.6.

■ 3.9 Less Noisy Domination and Logarithmic Sobolev Inequalities

Logarithmic Sobolev inequalities (LSIs) are a class of functional inequalities that shed light on several important phenomena such as isoperimetry, concentration of measure, and ergodicity and hypercontractivity of Markov semigroups. We refer readers to [168] and [19] for a general treatment of such inequalities, and more pertinently to [69] and [206], which present LSIs in the context of finite state space Markov chains. In this section, we illustrate that proving a channel $W \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ is less noisy than a channel $V \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ allows us to translate an LSI for W to an LSI for V. Thus, important information about V can be deduced (from its LSI) by proving $W \succeq_{\mathsf{ln}} V$ for an appropriate channel W (such as a q-ary symmetric channel) that has a known LSI.

We commence by introducing some appropriate notation and terminology associated with LSIs. For fixed input and output alphabet $\mathcal{X} = \mathcal{Y} = [q]$ with $q \in \mathbb{N}$, we think of a channel $W \in \mathbb{R}_{\mathsf{sto}}^{q \times q}$ as a Markov kernel on $\mathcal{X}$. We assume that the time homogeneous discrete-time Markov chain defined by W is *irreducible*, and has unique stationary distribution (or invariant measure) $\pi \in \mathcal{P}_q$ such that $\pi W = \pi$. Furthermore, we define the Hilbert space $\mathcal{L}^2(\mathcal{X}, \pi)$ of all real-valued functions with domain $\mathcal{X}$ endowed with the inner product:

$$\forall f, g \in \mathcal{L}^2(\mathcal{X}, \pi), \quad \langle f, g \rangle_\pi \triangleq \sum_{x \in \mathcal{X}} \pi(x) f(x) g(x) \tag{3.71}$$

and induced norm $\|\cdot\|_\pi$. We construe $W : \mathcal{L}^2(\mathcal{X}, \pi) \to \mathcal{L}^2(\mathcal{X}, \pi)$ as a conditional expectation operator that takes a function $f \in \mathcal{L}^2(\mathcal{X}, \pi)$, which we can write as a column vector $f = [f(0) \cdots f(q-1)]^T \in \mathbb{R}^q$, to another function $Wf \in \mathcal{L}^2(\mathcal{X}, \pi)$, which we can also write as a column vector $Wf \in \mathbb{R}^q$. Corresponding to the discrete-time Markov chain W, we may also define a continuous-time *Markov semigroup*:

$$\forall t \geq 0, \quad H_t \triangleq \exp\left(-t\left(I_q - W\right)\right) \in \mathbb{R}_{\mathsf{sto}}^{q \times q} \tag{3.72}$$

where the "discrete-time derivative" $W - I_q$ is the *Laplacian operator* that forms the *generator* of the Markov semigroup. The unique stationary distribution of this Markov semigroup is also π, and we may interpret $H_t : \mathcal{L}^2(\mathcal{X}, \pi) \to \mathcal{L}^2(\mathcal{X}, \pi)$ as a conditional expectation operator for each $t \geq 0$ as well.

In order to present LSIs, we define the *Dirichlet form* $\mathcal{E}_W : \mathcal{L}^2(\mathcal{X}, \pi) \times \mathcal{L}^2(\mathcal{X}, \pi) \to \mathbb{R}$:

$$\forall f, g \in \mathcal{L}^2(\mathcal{X}, \pi), \quad \mathcal{E}_W(f, g) \triangleq \langle (I_q - W) f, g \rangle_\pi \tag{3.73}$$

which is used to study properties of the Markov chain W and its associated Markov semigroup $\{H_t \in \mathbb{R}_{\mathsf{sto}}^{q\times q} : t \geq 0\}$. ($\mathcal{E}_W$ is technically only a Dirichlet form when W is a reversible Markov chain, i.e. W is a self-adjoint operator, or equivalently, W and π satisfy the detailed balance conditions [69, Section 2.3, p.705].) Moreover, the quadratic form defined by $\mathcal{E}_W$ represents the energy of its input function, and satisfies:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \pi)\,, \;\; \mathcal{E}_W(f,f) = \left\langle \left(I_q - \frac{W + W^*}{2}\right) f, f \right\rangle_\pi \tag{3.74}$$

where $W^* : \mathcal{L}^2(\mathcal{X}, \pi) \to \mathcal{L}^2(\mathcal{X}, \pi)$ is the adjoint operator of W. Finally, we introduce a particularly important Dirichlet form corresponding to the channel $W_{(q-1)/q} = \mathbf{1u}$, which has all uniform conditional pmfs and uniform stationary distribution $\pi = \mathbf{u}$, known as the *standard Dirichlet form*:

$$\mathcal{E}_{\mathsf{std}}(f,g) \triangleq \mathcal{E}_{\mathbf{1u}}(f,g) = \mathbb{COV}_{\mathbf{u}}(f,g) = \sum_{x\in\mathcal{X}} \frac{f(x)g(x)}{q} - \left(\sum_{x\in\mathcal{X}} \frac{f(x)}{q}\right)\left(\sum_{x\in\mathcal{X}} \frac{g(x)}{q}\right) \tag{3.75}$$

for any $f, g \in \mathcal{L}^2(\mathcal{X}, \mathbf{u})$. The quadratic form defined by the standard Dirichlet form is presented in (3.34) in subsection 3.3.4.

We now present the LSIs associated with the Markov chain W and the Markov semigroup $\{H_t \in \mathbb{R}_{\mathsf{sto}}^{q\times q} : t \geq 0\}$ it defines. The LSI for the Markov semigroup with constant $\alpha \in \mathbb{R}$ states that for every $f \in \mathcal{L}^2(\mathcal{X}, \pi)$ such that $\|f\|_\pi = 1$, we have:

$$D(f^2\pi||\pi) = \sum_{x\in\mathcal{X}} \pi(x) f^2(x) \log\Big(f^2(x)\Big) \leq \frac{1}{\alpha}\,\mathcal{E}_W(f,f) \tag{3.76}$$

where we construe $\mu = f^2\pi \in \mathcal{P}_q$ as a pmf such that $\mu(x) = f(x)^2\pi(x)$ for every $x \in \mathcal{X}$, and f^2 behaves like the Radon-Nikodym derivative (or density) of μ with respect to π. The largest constant α such that (3.76) holds:

$$\alpha(W) \triangleq \inf_{\substack{f\in\mathcal{L}^2(\mathcal{X},\pi):\\ \|f\|_\pi=1 \\ D(f^2\pi||\pi)\neq 0}} \frac{\mathcal{E}_W(f,f)}{D(f^2\pi||\pi)} \tag{3.77}$$

is called the *logarithmic Sobolev constant* (LSI constant) of the Markov chain W (or the Markov chain $(W + W^*)/2$). Likewise, the LSI for the discrete-time Markov chain with constant $\alpha \in \mathbb{R}$ states that for every $f \in \mathcal{L}^2(\mathcal{X}, \pi)$ such that $\|f\|_\pi = 1$, we have:

$$D(f^2\pi||\pi) \leq \frac{1}{\alpha}\,\mathcal{E}_{WW^*}(f,f) \tag{3.78}$$

where $\mathcal{E}_{WW^*} : \mathcal{L}^2(\mathcal{X}, \pi) \times \mathcal{L}^2(\mathcal{X}, \pi) \to \mathbb{R}$ is the "discrete" Dirichlet form. The largest constant α such that (3.78) holds is the LSI constant of the Markov chain WW^*, $\alpha(WW^*)$, and we refer to it as the *discrete logarithmic Sobolev constant* of the Markov

chain W. As we mentioned earlier, there are many useful consequences of such LSIs. For example, if (3.76) holds with constant (3.77), then for every pmf $\mu \in \mathcal{P}_q$:

$$\forall t \geq 0, \;\; D(\mu H_t || \pi) \leq e^{-2\alpha(W)t} D(\mu||\pi) \tag{3.79}$$

where the exponent $2\alpha(W)$ can be improved to $4\alpha(W)$ if W is reversible [69, Theorem 3.6]. This is a measure of ergodicity of the semigroup $\{H_t \in \mathbb{R}_{\mathsf{sto}}^{q\times q} : t \geq 0\}$. Likewise, if (3.78) holds with constant $\alpha(WW^*)$, then for every pmf $\mu \in \mathcal{P}_q$:

$$\forall n \in \mathbb{N}, \;\; D(\mu W^n||\pi) \leq (1-\alpha(WW^*))^n \, D(\mu||\pi) \tag{3.80}$$

as mentioned in [69, Remark, p.725] and proved in [203]. This is also a measure of ergodicity of the Markov chain W.

Although LSIs have many useful consequences, LSI constants are difficult to compute analytically. Fortunately, the LSI constant corresponding to $\mathcal{E}_{\mathsf{std}}$ has been computed in [69, Appendix, Theorem A.1]. Therefore, using the relation in (3.35), we can compute LSI constants for q-ary symmetric channels as well. The next proposition collects the LSI constants for q-ary symmetric channels (which are irreducible for $\delta \in (0,1]$) as well as some other related quantities.

Proposition 3.12 (Constants of Symmetric Channels). *The q-ary symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with $q \geq 2$ has:*

1. *LSI constant:*
$$\alpha(W_\delta) = \begin{cases} \delta & , \quad q = 2 \\ \frac{(q-2)\delta}{(q-1)\log(q-1)} & , \quad q > 2 \end{cases}$$
for $\delta \in (0,1]$.

2. *discrete LSI constant:*
$$\alpha(W_\delta W_\delta^*) = \alpha(W_{\delta'}) = \begin{cases} 2\delta(1-\delta) & , \quad q = 2 \\ \frac{(q-2)(2q-2-q\delta)\delta}{(q-1)^2\log(q-1)} & , \quad q > 2 \end{cases}$$
for $\delta \in (0,1]$, where $\delta' = \delta(2 - \frac{q\delta}{q-1})$.

3. *maximal correlation corresponding to the uniform stationary distribution $\boldsymbol{u} \in \mathcal{P}_q$ (see Definition 2.3 in chapter 2):*
$$\rho_{\mathsf{max}}(X;Y) = \left|1 - \delta - \frac{\delta}{q-1}\right|$$
for $\delta \in [0,1]$, where the random variables $X, Y \in [q]$ have joint distribution defined by $P_X = \boldsymbol{u}$ and $P_{Y|X} = W_\delta$.

4. *contraction coefficient for KL divergence bounded by:*

$$\left(1-\delta-\frac{\delta}{q-1}\right)^2 \leq \eta_{\mathsf{KL}}(W_\delta) \leq \left|1-\delta-\frac{\delta}{q-1}\right|$$

for $\delta \in [0,1]$.

Proof. See appendix B.8. ■

In view of Proposition 3.12 and the intractability of computing LSI constants for general Markov chains, we often "compare" a given irreducible channel $V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ with a q-ary symmetric channel $W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ to try and establish an LSI for it. We assume for the sake of simplicity that V is doubly stochastic and has uniform stationary pmf (just like q-ary symmetric channels). Usually, such a comparison between W_δ and V requires us to prove domination of Dirichlet forms, such as:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u}), \;\; \mathcal{E}_V(f,f) \geq \mathcal{E}_{W_\delta}(f,f) = \frac{q\delta}{q-1}\mathcal{E}_{\mathsf{std}}(f,f) \tag{3.81}$$

where we use (3.35). Such pointwise domination results immediately produce LSIs, (3.76) and (3.78), for V. Furthermore, they also lower bound the LSI constants of V; for example:

$$\alpha(V) \geq \alpha(W_\delta). \tag{3.82}$$

This in turn begets other results such as (3.79) and (3.80) for the channel V (albeit with worse constants in the exponents since the LSI constants of W_δ are used instead of those for V). More general versions of Dirichlet form domination between Markov chains on different state spaces with different stationary distributions, and the resulting bounds on their LSI constants are presented in [69, Lemmata 3.3 and 3.4]. We next illustrate that the information theoretic notion of less noisy domination is a sufficient condition for various kinds of pointwise Dirichlet form domination.

Theorem 3.7 (Domination of Dirichlet Forms). *Let* $W, V \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ *be doubly stochastic channels, and* $\pi = \boldsymbol{u}$ *be the uniform stationary distribution. Then, the following are true:*

1. *If* $W \succeq_{\mathsf{ln}} V$, *then:*

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \boldsymbol{u}), \;\; \mathcal{E}_{VV^*}(f,f) \geq \mathcal{E}_{WW^*}(f,f).$$

2. *If* $W \in \mathbb{R}_{\succeq 0}^{q\times q}$ *is positive semidefinite,* V *is normal (i.e.* $V^TV = VV^T$*), and* $W \succeq_{\mathsf{ln}} V$, *then:*

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \boldsymbol{u}), \;\; \mathcal{E}_V(f,f) \geq \mathcal{E}_W(f,f).$$

3. *If* $W = W_\delta \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ *is any* q*-ary symmetric channel with* $\delta \in \left[0, \frac{q-1}{q}\right]$ *and* $W_\delta \succeq_{\mathsf{ln}} V$, *then:*

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \boldsymbol{u}), \;\; \mathcal{E}_V(f,f) \geq \frac{q\delta}{q-1}\mathcal{E}_{\mathsf{std}}(f,f).$$

Proof.

Part 1: First observe that:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u}), \;\; \mathcal{E}_{WW^*}(f,f) = \frac{1}{q} f^T \left(I_q - WW^T\right) f$$
$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u}), \;\; \mathcal{E}_{VV^*}(f,f) = \frac{1}{q} f^T \left(I_q - VV^T\right) f$$

where we use the facts that $W^T = W^*$ and $V^T = V^*$ because the stationary distribution is uniform. This implies that $\mathcal{E}_{VV^*}(f,f) \geq \mathcal{E}_{WW^*}(f,f)$ for every $f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u})$ if and only if $I_q - VV^T \succeq_{\mathsf{PSD}} I_q - WW^T$, which is true if and only if $WW^T \succeq_{\mathsf{PSD}} VV^T$. Since $W \succeq_{\mathsf{ln}} V$, we get $WW^T \succeq_{\mathsf{PSD}} VV^T$ from part 2 of Proposition 3.8 after letting $P_X = \mathbf{u} = P_X W = P_X V$.

Part 2: Once again, we first observe using (3.74) that:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u}), \;\; \mathcal{E}_{W}(f,f) = \frac{1}{q} f^T \left(I_q - \frac{W + W^T}{2}\right) f \,,$$
$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u}), \;\; \mathcal{E}_{V}(f,f) = \frac{1}{q} f^T \left(I_q - \frac{V + V^T}{2}\right) f \,.$$

So, $\mathcal{E}_V(f,f) \geq \mathcal{E}_W(f,f)$ for every $f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u})$ if and only if $(W + W^T)/2 \succeq_{\mathsf{PSD}} (V + V^T)/2$. Since $WW^T \succeq_{\mathsf{PSD}} VV^T$ from the proof of part 1, it is sufficient to prove that:

$$WW^T \succeq_{\mathsf{PSD}} VV^T \quad \Rightarrow \quad \frac{W + W^T}{2} \succeq_{\mathsf{PSD}} \frac{V + V^T}{2} \,. \tag{3.83}$$

Lemma B.5 in appendix B.10 establishes the claim in (3.83) because $W \in \mathbb{R}^{q \times q}_{\succeq 0}$ and V is a normal matrix.

Part 3: We note that when V is a normal matrix, this result follows from part 2 because $W_\delta \in \mathbb{R}^{q \times q}_{\succeq 0}$ for $\delta \in [0, \frac{q-1}{q}]$, as can be seen from part 2 of Proposition 3.4. For a general doubly stochastic channel V, we need to prove that:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbf{u}), \;\; \mathcal{E}_V(f,f) \geq \mathcal{E}_{W_\delta}(f,f) = \frac{q\delta}{q-1} \mathcal{E}_{\mathsf{std}}(f,f)$$

where we use (3.35). Following the proof of part 2, it is sufficient to prove (3.83) with $W = W_\delta$:[47]

$$W_\delta^2 \succeq_{\mathsf{PSD}} VV^T \quad \Rightarrow \quad W_\delta \succeq_{\mathsf{PSD}} \frac{V + V^T}{2}$$

where $W_\delta^2 = W_\delta W_\delta^T$ and $W_\delta = (W_\delta + W_\delta^T)/2$. Recall part 1 of Theorem B.1 (the Löwner-Heinz theorem) from appendix B.2, which states that for $A, B \in \mathbb{R}^{q \times q}_{\succeq 0}$ and $p \in [0,1]$:[48]

$$A \succeq_{\mathsf{PSD}} B \quad \Rightarrow \quad A^p \succeq_{\mathsf{PSD}} B^p \,. \tag{3.84}$$

[47] Note that (3.83) trivially holds for $W = W_\delta$ with $\delta = (q-1)/q$, because $W_{(q-1)/q} = W^2_{(q-1)/q} = \mathbf{1u} \succeq_{\mathsf{PSD}} VV^T$ implies that $V = W_{(q-1)/q}$.

[48] See [224] for a short operator theoretic proof of this result.

Using (3.84) with $p = \frac{1}{2}$ (cf. [129, Corollary 7.7.4(b)]), we have:

$$W_\delta^2 \succeq_{\mathsf{PSD}} VV^T \quad \Rightarrow \quad W_\delta \succeq_{\mathsf{PSD}} \left(VV^T\right)^{\frac{1}{2}}$$

because the Gramian matrix $VV^T \in \mathbb{R}^{q\times q}_{\succeq 0}$ is positive semidefinite.

Let $VV^T = Q\Lambda Q^T$ and $(V+V^T)/2 = U\Sigma U^T$ be the spectral decompositions of VV^T and $(V+V^T)/2$, where $Q, U \in \mathcal{V}_q(\mathbb{R}^q)$ are orthogonal matrices with eigenvectors as columns, and $\Lambda, \Sigma \in \mathbb{R}^{q\times q}$ are diagonal matrices of eigenvalues. Since VV^T and $(V+V^T)/2$ are both doubly stochastic, they both have the unit norm eigenvector $\mathbf{1}/\sqrt{q}$ corresponding to the maximum eigenvalue of unity. In fact, we have:

$$\left(VV^T\right)^{\frac{1}{2}} \frac{\mathbf{1}}{\sqrt{q}} = \frac{\mathbf{1}}{\sqrt{q}} \quad \text{and} \quad \left(\frac{V+V^T}{2}\right) \frac{\mathbf{1}}{\sqrt{q}} = \frac{\mathbf{1}}{\sqrt{q}}$$

where we use the fact that $(VV^T)^{\frac{1}{2}} = Q\Lambda^{\frac{1}{2}}Q^T$ is the spectral decomposition of $(VV^T)^{\frac{1}{2}}$. For any matrix $A \in \mathbb{R}^{q\times q}_{\mathsf{sym}}$, let $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_q(A)$ denote the eigenvalues of A in descending order. Without loss of generality, we assume that for every $j \in \{1, \ldots, q\}$:

$$[\Lambda]_{j,j} = \lambda_j\left(VV^T\right),$$
$$[\Sigma]_{j,j} = \lambda_j\left(\frac{V+V^T}{2}\right).$$

So, $\lambda_1((VV^T)^{\frac{1}{2}}) = \lambda_1((V+V^T)/2) = 1$, and the first columns of both Q and U are equal to $\mathbf{1}/\sqrt{q}$.

From part 2 of Proposition 3.4, we have $W_\delta = QDQ^T = UDU^T$, where $D \in \mathbb{R}^{q\times q}$ is the diagonal matrix of eigenvalues such that $[D]_{1,1} = \lambda_1(W_\delta) = 1$ and $[D]_{j,j} = \lambda_j(W_\delta) = 1 - \delta - \frac{\delta}{q-1}$ for $j \in \{2, \ldots, q\}$. Note that we may use either of the eigenbases, Q or U, because they both have first column $\mathbf{1}/\sqrt{q}$, which is the eigenvector of W_δ corresponding to $\lambda_1(W_\delta) = 1$ since W_δ is doubly stochastic, and the remaining eigenvector columns are permitted to be any orthonormal basis of $\mathsf{span}(\mathbf{1}/\sqrt{q})^\perp$ (which is the orthogonal complement subspace of the span of $\mathbf{1}/\sqrt{q}$) as $\lambda_j(W_\delta) = 1 - \delta - \frac{\delta}{q-1}$ for $j \in \{2, \ldots, q\}$. Hence, we have:

$$W_\delta \succeq_{\mathsf{PSD}} \left(VV^T\right)^{\frac{1}{2}} \quad \Leftrightarrow \quad QDQ^T \succeq_{\mathsf{PSD}} Q\Lambda^{\frac{1}{2}}Q^T \quad \Leftrightarrow \quad D \succeq_{\mathsf{PSD}} \Lambda^{\frac{1}{2}},$$
$$W_\delta \succeq_{\mathsf{PSD}} \frac{V+V^T}{2} \quad \Leftrightarrow \quad UDU^T \succeq_{\mathsf{PSD}} U\Sigma U^T \quad \Leftrightarrow \quad D \succeq_{\mathsf{PSD}} \Sigma.$$

In order to show that $D \succeq_{\mathsf{PSD}} \Lambda^{\frac{1}{2}} \Rightarrow D \succeq_{\mathsf{PSD}} \Sigma$, it suffices to prove that $\Lambda^{\frac{1}{2}} \succeq_{\mathsf{PSD}} \Sigma$. Recall the following eigenvalue domination lemma, cf. [128, Corollary 3.1.5], which states that for any matrix $A \in \mathbb{R}^{q\times q}$, the ith largest eigenvalue of the symmetric part of A is less than or equal to the ith largest singular value of A (which is the ith largest eigenvalue of the unique positive semidefinite part $(AA^T)^{1/2}$ in the polar decomposition of A) for every $i \in \{1, \ldots, q\}$.

Lemma 3.3 (Eigenvalue Domination [128]). *Given a matrix $A \in \mathbb{R}^{q\times q}$, we have for every $i \in \{1,\dots,q\}$:*

$$\lambda_i\left(\left(AA^T\right)^{\frac{1}{2}}\right) \geq \lambda_i\left(\frac{A+A^T}{2}\right).$$

Hence, Lemma 3.3 implies that $\Lambda^{\frac{1}{2}} \succeq_{\mathsf{PSD}} \Sigma$ is true, cf. [70, Lemma 2.5]. This completes the proof. ■

Theorem 3.7 includes Theorem 3.6 from section 3.3 as part 3, and also provides two other useful pointwise Dirichlet form domination results. Part 1 of Theorem 3.7 states that less noisy domination implies discrete Dirichlet form domination. In particular, if we have $W_\delta \succeq_{\mathsf{ln}} V$ for some irreducible q-ary symmetric channel $W_\delta \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ and irreducible doubly stochastic channel $V \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$, then part 1 implies that:

$$\forall n \in \mathbb{N},\ \ D(\mu V^n||\mathbf{u}) \leq (1-\alpha(W_\delta W_\delta^*))^n\, D(\mu||\mathbf{u}) \tag{3.85}$$

for all pmfs $\mu \in \mathcal{P}_q$, where $\alpha(W_\delta W_\delta^*)$ is computed in part 2 of Proposition 3.12. However, it is worth mentioning that (3.80) for W_δ and Proposition 3.1 directly produce (3.85). So, such ergodicity results for the discrete-time Markov chain V do not require the full power of the Dirichlet form domination in part 1. Regardless, Dirichlet form domination results, such as in parts 2 and 3, yield several functional inequalities (like Poincaré inequalities and LSIs) which have many other potent consequences as well.

Parts 2 and 3 of Theorem 3.7 convey that less noisy domination also implies the usual (continuous) Dirichlet form domination under regularity conditions. We note that in part 2, the channel W is more general than that in part 3, but the channel V is restricted to be normal (which includes the case where V is an additive noise channel). The proofs of these parts essentially consist of two segments. The first segment uses part 1, and the second segment illustrates that pointwise domination of discrete Dirichlet forms implies pointwise domination of Dirichlet forms (as shown in (3.81)). This latter segment is encapsulated in Lemma B.5 in appendix B.10 for part 2, and requires a slightly more sophisticated proof pertaining to q-ary symmetric channels in part 3.

■ 3.10 Conclusion and Future Directions

In closing this chapter, we briefly reiterate our main results by delineating a possible program for proving LSIs for certain Markov chains. Given an arbitrary irreducible doubly stochastic channel $V \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ with minimum probability entry $\nu = \min\{[V]_{i,j} : i,j \in \{1,\dots,q\}\} > 0$ and $q \geq 2$, we can first use Theorem 3.4 to generate a q-ary symmetric channel $W_\delta \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ with $\delta = \nu/(1-(q-1)\nu+\frac{\nu}{q-1})$ such that $W_\delta \succeq_{\mathsf{deg}} V$. This also means that $W_\delta \succeq_{\mathsf{ln}} V$, using Proposition 3.3. Moreover, the δ parameter can be improved using Theorem 3.5 (or Propositions 3.10 and 3.11) if V is an additive noise channel. We can then use Theorem 3.7 to deduce a pointwise domination of Dirichlet forms. Since W_δ satisfies the LSIs (3.76) and (3.78) with corresponding LSI constants

given in Proposition 3.12, Theorem 3.7 establishes the following LSIs for the channel V:

$$D(f^2\mathbf{u}||\mathbf{u}) \leq \frac{1}{\alpha(W_\delta)}\,\mathcal{E}_V(f,f) \tag{3.86}$$

$$D(f^2\mathbf{u}||\mathbf{u}) \leq \frac{1}{\alpha(W_\delta W_\delta^*)}\,\mathcal{E}_{VV^*}(f,f) \tag{3.87}$$

for every $f \in \mathcal{L}^2(\mathcal{X},\mathbf{u})$ such that $\|f\|_{\mathbf{u}} = 1$. These inequalities can be used to derive a myriad of important facts about V. We note that the equivalent characterizations of the less noisy preorder via non-linear operator convex f-divergences in Theorem 3.1, and specifically Theorem 3.3 and Proposition 3.8, are particularly useful for proving some of these results. Furthermore, Theorem 3.1 generalizes Proposition 2.6 in chapter 2, which is a well-known result in the contraction coefficients literature, and has other applications in information theory as well, such as our generalization to Samorodnitsky's SDPI in Theorem 3.2. Finally, we accentuate that Theorems 3.4 and 3.5 address our motivation in section 3.2 by providing analogues of the relationship between less noisy domination by q-ary erasure channels and contraction coefficients in the context of q-ary symmetric channels.

Many of the results in this chapter could possibly be extended or generalized. For example, Theorem 3.4 only holds for square channel transition probability matrices. However, degradation does not require input and output alphabet sizes to match, and Theorem 3.4 could potentially be extended for rectangular channel transition probability matrices. In the special context of additive noise channels, Theorem 3.5 does not completely characterize $\mathcal{L}_{W_\delta}^{\mathsf{add}}$. Hence, another direction of future research is to establish better "bounds" on $\mathcal{L}_{W_\delta}^{\mathsf{add}}$. In particular, our lower bound on $\mathcal{L}_{W_\delta}^{\mathsf{add}}$ (proved in Proposition 3.11) could be improved by finding other noise pmfs that belong to $\mathcal{L}_{W_\delta}^{\mathsf{add}} \backslash \mathcal{D}_{W_\delta}^{\mathsf{add}}$ (and then applying Proposition 3.6). Lastly, the pointwise Dirichlet form domination results in Theorem 3.7 are only derived for doubly stochastic Markov chains with uniform stationary distribution. These results could probably be generalized for Markov chains with non-uniform stationary distributions.

■ 3.11 Bibliographical Notes

Chapter 3 and appendix B are based primarily on the journal paper [188], and partly on the manuscript [192]. The work in [188] was also published in part at the Proceedings of the IEEE International Symposium on Information Theory (ISIT) 2017 [187].

Chapter 4

Modal Decomposition of Mutual χ^2-Information

In this chapter, we delve into the elegant geometric structure of the contraction coefficient for χ^2-divergence. Recall from (2.37) in chapter 2 that the contraction coefficient for χ^2-divergence $\eta_{\chi^2}(P_X, P_{Y|X})$ of a source-channel pair $(P_X, P_{Y|X})$ is equal to the squared maximal correlation $\rho_{\mathsf{max}}(X;Y)^2$ of the input and output random variables X and Y. Since Proposition 2.2 illustrates that $\rho_{\mathsf{max}}(X;Y)$ can be characterized as a singular value of the so called DTM associated to the joint distribution $P_{X,Y}$, the DPI for χ^2-divergence can be geometrically understood using the SVD of the DTM. We will develop "modal decompositions" of bivariate distributions and mutual χ^2-information based on SVDs of DTMs in the ensuing discussion. To emphasize the utility of these ideas in statistical inference and machine learning applications, we will adopt a statistical perspective in this chapter instead of the usual information contraction perspective of previous chapters.

Recall from Definition 2.3 in chapter 2 that the Hirschfeld-Gebelein-Rényi maximal correlation is a variational generalization of the well-known Pearson correlation coefficient, and was originally introduced as a normalized measure of dependence between two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ [96, 125, 236, 242]:

$$\rho_{\mathsf{max}}(X;Y) \triangleq \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\, g:\mathcal{Y}\to\mathbb{R}:\\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0\\ \mathbb{E}[f(X)^2]=\mathbb{E}[g(Y)^2]=1}} \mathbb{E}\left[f(X)g(Y)\right]. \tag{4.1}$$

Indeed, it is easily verified that $0 \le \rho_{\mathsf{max}}(X;Y) \le 1$, and $\rho_{\mathsf{max}}(X;Y) = 0$ if and only if X is independent of Y (cf. Proposition 2.3). It turns out that the variational formulation of maximal correlation in (4.1) shares deep ties with a class of statistical inference problems. We consider inference problems with the general structure of a Markov chain $U \to X \to Y$, where the conditional distributions $P_{Y|U}$ form the (overall) observation model and the conditional distributions $P_{Y|X}$ form the noise model. Our goal is to make decisions about the *latent variable* U corresponding to the data X based on some noisy observation of the data Y. In many applications, the true observation model $P_{Y|U}$ is unknown. Therefore, a natural way to address the statistical inference problem

of extracting information about U based on the noisy observation Y is to first learn the observation model $P_{Y|U}$ from training data, and then employ standard decoding techniques that use knowledge of the likelihoods $P_{Y|U}$ to solve the inference problem.

Unfortunately, it is difficult to learn the observation model $P_{Y|U}$ in many applications. For instance, in the popular setting of the "Netflix problem" of recommending movies to subscribers [23], if we let X denote the subscriber index and Y denote the movie index, it is challenging to identify what latent variable U of a subscriber is relevant to their choice of movies. So, we cannot obtain (labeled) training data to learn $P_{Y|U}$. A different approach to such problems is to focus only on the noise model, since we can easily obtain bivariate training data samples of the form (X, Y) (which represents subscriber X streaming movie Y). In this spirit, we try to find features of the noisy observation Y that carry as much information about X as possible, and yet are simple enough so that further processing, such as clustering or kernel methods, can be applied to make final decisions. Most *dimensionality reduction* algorithms follow this approach. For example, one way to establish the information theoretic optimality of principal component analysis is to assume that X is jointly Gaussian distributed and Y is a noisy observation of X after adding independent white Gaussian noise. In this case, the principle components of the observed Y can be shown to carry the maximum amount of mutual information about X, cf. [176].

We can interpret maximal correlation as a general formulation for this approach. The optimization problem in (4.1) tries to find a feature $g(Y)$ that is highly correlated with some feature $f(X)$, or equivalently, has high predictive power towards some aspects of X. The advantage of finding such a feature (or embedding) is that $g(Y)$ can be a general real-valued function. In particular, it need not be a linear function of the data, and the data itself need not be real-valued, i.e. we can have *categorical data*. Thus, the maximal correlation formulation in (4.1) provides a general basis for performing *feature extraction* from high-dimensional categorical data.[49] Our goal in this chapter is to extend this framework and develop a practical algorithm to learn useful features from categorical data. In the ensuing discussion, we will present an efficient algorithm that solves the optimization problem in (4.1) using training samples, show that both the formulation of maximal correlation and the associated algorithm can be generalized to produce an arbitrary number of features, and explain how the resulting approach is different from existing methods such as principal component analysis and canonical correlation analysis.

■ 4.1 Chapter Outline

We briefly delineate the content of the ensuing sections. In section 4.2, we will introduce modal decompositions of bivariate distributions—the key players of this chapter. Then, we will construe such modal decompositions under a local information geometric lens

[49] In the categorical data setting of this chapter, "high-dimensional" refers to the large cardinalities of $\mathcal{X}$ and $\mathcal{Y}$.

in section 4.3. We will present an algorithm to learn the maximal correlation functions that make up modal decompositions from training data in section 4.4. Moreover, we will compare this algorithm with related statistical techniques and analyze its sample complexity in sections 4.5 and 4.6, respectively. Finally, we will conclude this discussion and present some future research directions in section 4.7. Additionally, at the end of this chapter in section 4.8, we will shortly digress and analyze reliable communication through permutation channels.

■ 4.2 Modal Decomposition of Bivariate Distributions

In order to present modal decompositions of bivariate distributions, we first introduce some relevant notation and assumptions. Consider the finite alphabet sets $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ and $\mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$ (without loss of generality) such that $2 \leq |\mathcal{X}|, |\mathcal{Y}| < +\infty$. Let $\mathcal{P}_{\mathcal{X}} \subseteq \left(\mathbb{R}^{|\mathcal{X}|}\right)^*$ and $\mathcal{P}_{\mathcal{Y}} \subseteq \left(\mathbb{R}^{|\mathcal{Y}|}\right)^*$ denote the probability simplices of pmfs corresponding to the discrete random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, respectively. Similarly, let $\mathcal{P}_{\mathcal{X}}^{\circ}$ and $\mathcal{P}_{\mathcal{Y}}^{\circ}$ denote the relative interiors of $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$, respectively. Since any element of $\mathcal{X}$ or $\mathcal{Y}$ whose marginal probability mass is zero can be dropped from the sample space altogether, we will restrict our attention to the following subset of joint pmfs between X and Y (with abuse of notation):

$$\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \triangleq \{P_{X,Y} : P_{X,Y} \text{ is a joint pmf on } \mathcal{X} \times \mathcal{Y} \text{ such that } P_X \in \mathcal{P}_{\mathcal{X}}^{\circ} \text{ and } P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}\} \tag{4.2}$$

where P_X and P_Y denote the marginal distributions of the bivariate distribution $P_{X,Y}$. Furthermore, let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}^{\circ}$ denote the relative interior of $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, i.e. the set of all entry-wise strictly positive joint pmfs of X and Y. Lastly, we note that the aforementioned simplices can be perceived as metric spaces (with respect to e.g. the standard Euclidean ℓ^2-norm), and topological statements in the sequel should be understood in terms of these metric spaces.

■ 4.2.1 Divergence Transition and Canonical Dependence Matrices

Fix any bivariate distribution $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. Let $\mathcal{L}^2(\mathcal{X}, P_X)$ denote the Hilbert space of real-valued functions on $\mathcal{X}$ with correlation as the inner product (also see (3.71) in chapter 3):

$$\forall f, f' \in \mathcal{L}^2(\mathcal{X}, P_X), \;\; \langle f, f' \rangle_{P_X} \triangleq \mathbb{E}[f(X)f'(X)] = \sum_{x\in\mathcal{X}} P_X(x)f(x)f'(x)\,, \tag{4.3}$$

and second moment as the induced norm:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, P_X), \;\; \|f\|_{P_X} \triangleq \mathbb{E}\Big[f(X)^2\Big]^{\frac{1}{2}} = \left(\sum_{x\in\mathcal{X}} P_X(x)f(x)^2\right)^{\frac{1}{2}}. \tag{4.4}$$

Similarly, let $\mathcal{L}^2(\mathcal{Y}, P_Y)$ denote the Hilbert space of real-valued functions on $\mathcal{Y}$ with correlation as the inner product. We will analyze two equivalent linear operators corresponding to the joint pmf $P_{X,Y}$. The first of these is the usual *conditional expectation*

operator, $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$, which maps any function $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to the function $C(f) \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ given by:

$$\forall y \in \mathcal{Y}, \;\; (C(f))(y) \triangleq \mathbb{E}[f(X)|Y = y] \,. \tag{4.5}$$

The second operator is given by the DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$, cf. [139], which defines a linear map $B : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{Y}|}$ via matrix-vector multiplication between the Euclidean spaces $\mathbb{R}^{|\mathcal{X}|}$ and $\mathbb{R}^{|\mathcal{Y}|}$.

Definition 4.1 (Divergence Transition Matrix). *For any joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, we define its corresponding* divergence transition matrix $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ *entry-wise as:*

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; [B]_{y,x} \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}} \,.$$

We remark that (2.38) in chapter 2 defines B^T as the DTM (instead of B), because B^T maps row vectors $v \in \left(\mathbb{R}^{|\mathcal{X}|}\right)^*$ to row vectors $vB^T \in \left(\mathbb{R}^{|\mathcal{Y}|}\right)^*$, and chapter 2 mainly deals with perturbation vectors of pmfs in $\mathcal{P}_{\mathcal{X}}$, which are row vectors. In contrast, the version of the DTM in Definition 4.1 is more useful in this chapter, since we will construe the DTM as linear map on column vectors associated with functions in $\mathcal{L}^2(\mathcal{X}, P_X)$.

It is straightforward to verify that $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ and $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ are equivalent maps. Indeed, notice that the Euclidean space $\mathbb{R}^{|\mathcal{X}|}$ (with standard Euclidean inner product) is *isometrically isomorphic* to $\mathcal{L}^2(\mathcal{X}, P_X)$, because for any $\psi \in \mathbb{R}^{|\mathcal{X}|}$, we can construct $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ via:[50]

$$\forall x \in \mathcal{X}, \;\; f(x) = \frac{\psi(x)}{\sqrt{P_X(x)}} \tag{4.6}$$

where we abuse notation and let $\psi(x) = \psi_x$ denote the xth element of ψ. So, for any two vectors $\psi, \psi' \in \mathbb{R}^{|\mathcal{X}|}$ with corresponding functions $f, f' \in \mathcal{L}^2(\mathcal{X}, P_X)$, respectively, defined using (4.6), we have:

$$\psi^T\psi' = \sum_{x\in\mathcal{X}} \psi(x)\psi'(x) = \sum_{x\in\mathcal{X}} P_X(x)f(x)f'(x) = \langle f, f' \rangle_{P_X} \,. \tag{4.7}$$

Likewise, the Euclidean space $\mathbb{R}^{|\mathcal{Y}|}$ (with standard Euclidean inner product) is isometrically isomorphic to $\mathcal{L}^2(\mathcal{Y}, P_Y)$, because for any $\phi \in \mathbb{R}^{|\mathcal{Y}|}$, we can construct $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ via:

$$\forall y \in \mathcal{Y}, \;\; g(y) = \frac{\phi(y)}{\sqrt{P_Y(y)}} \,. \tag{4.8}$$

Thus, the linear operators $B : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{Y}|}$ and $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ are equivalent in the sense that:

$$\phi = B\psi \quad \Leftrightarrow \quad g = C(f) \tag{4.9}$$

[50] It is well-known that any two separable Hilbert spaces with the same dimension are always isometrically isomorphic to each other.

for every $\psi \in \mathbb{R}^{|\mathcal{X}|}$, where $\phi \in \mathbb{R}^{|\mathcal{Y}|}$, $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ is defined by (4.6), and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ is defined by (4.8). To see this, observe that for every $\psi \in \mathbb{R}^{|\mathcal{X}|}$ such that $\phi = B\psi \in \mathbb{R}^{|\mathcal{Y}|}$, we have for all $y \in \mathcal{Y}$:

$$g(y) = \frac{\phi(y)}{\sqrt{P_Y(y)}} = \frac{1}{\sqrt{P_Y(y)}} \sum_{x \in \mathcal{X}} \underbrace{\frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}}}_{[B]_{y,x}} \psi(x) = \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) f(x) \tag{4.10}$$

where the first equality follows from (4.8), the second equality holds because $\phi = B\psi$, and the the third equality follows from (4.6).

Let the SVD of B be:

$$\forall i \in \{1, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}, \;\; B\psi_i = \sigma_i \phi_i \tag{4.11}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \geq 0$ are the ordered singular values of B, $\{\psi_i \in \mathbb{R}^{|\mathcal{X}|} : i \in \{1, \dots, |\mathcal{X}|\}\}$ is the orthonormal basis of corresponding right singular vectors of B, and $\{\phi_i \in \mathbb{R}^{|\mathcal{Y}|} : i \in \{1, \dots, |\mathcal{Y}|\}\}$ is the orthonormal basis of corresponding left singular vectors of B.[51] Then, due to the equivalence in (4.9), the SVD of C is:

$$\forall i \in \{1, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}, \;\; C(f_i) = \sigma_i g_i \tag{4.12}$$

where $\sigma_1, \dots, \sigma_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}$ are also the ordered singular values of C, $\{f_i \in \mathcal{L}^2(\mathcal{X}, P_X) : i \in \{1, \dots, |\mathcal{X}|\}\}$ is the orthonormal basis of right singular vectors of C defined by (4.6):

$$\forall i \in \{1, \dots, |\mathcal{X}|\}, \forall x \in \mathcal{X}, \;\; f_i(x) = \frac{\psi_i(x)}{\sqrt{P_X(x)}}, \tag{4.13}$$

and $\{g_i \in \mathcal{L}^2(\mathcal{Y}, P_Y) : i \in \{1, \dots, |\mathcal{Y}|\}\}$ is the orthonormal basis of left singular vectors of C defined by (4.8):

$$\forall i \in \{1, \dots, |\mathcal{Y}|\}, \forall y \in \mathcal{Y}, \;\; g_i(y) = \frac{\phi_i(y)}{\sqrt{P_Y(y)}}. \tag{4.14}$$

(Note that the orthonormality of the singular vectors of C is defined with respect to the appropriate Hilbert space inner products.) The next theorem presents some simple properties of the SVD of the DTM B and the conditional expectation operator C, cf. [11, 75, 125, 180, 236, 289].

Theorem 4.1 (Properties of DTMs and Conditional Expectation Operators). *For the DTM $B \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ and conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ corresponding to any joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$, the following statements are true:*

[51] Here, we complete the orthonormal bases of right and left singular vectors for convenience. Moreover, if $|\mathcal{X}| > |\mathcal{Y}|$, then $B\psi_i = \mathbf{0}$ for all $i \in \{|\mathcal{Y}| + 1, \dots, |\mathcal{X}|\}$.

1. *The largest singular value is unity:*

$$\|B\|_{\mathsf{op}} = \|C\|_{\mathsf{op}} = \sigma_1 = 1\,.$$

2. *The right and left singular vectors of B corresponding to $\sigma_1 = 1$ are:*

$$\psi_1 = \sqrt{P_X}^T \quad \textit{and} \quad \phi_1 = \sqrt{P_Y}^T,$$

and the right and left singular vectors of C corresponding to $\sigma_1 = 1$ are the constant functions:

$$f_1 = \mathbf{1} \quad \textit{and} \quad g_1 = \mathbf{1}$$

where we use $\mathbf{1}$ to represent everywhere unity functions with appropriate domains.

3. *The second largest singular value is maximal correlation:*

$$\sigma_2 = \rho_{\mathsf{max}}(X;Y)\,.$$

4. *The right and left singular vectors of C corresponding to σ_2, $f_2 \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g_2 \in \mathcal{L}^2(\mathcal{Y}, P_Y)$, are the maximal correlation functions that solve the extremal problem in (4.1):*

$$\rho_{\mathsf{max}}(X;Y) = \mathbb{E}[f_2(X)g_2(Y)]\,.$$

Proof. This follows from Proposition 2.2 in chapter 2 and its proof in appendix A.1, and the relations (4.6), (4.8), and (4.9). ■

Part 1 of Theorem 4.1 portrays that $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ is a *contraction*, i.e. $\|C\|_{\mathsf{op}} = 1$. It turns out that this implies the DPI for χ^2-divergence (with fixed input pmf P_X). Indeed, for any input pmf $R_X \in \mathcal{P}_{\mathcal{X}}$, let $R_Y \in \mathcal{P}_{\mathcal{Y}}$ denote the induced output pmf after passing R_X through the channel $P_{Y|X}$, and construct the functions:

$$\forall x \in \mathcal{X},\ \ f(x) = \frac{R_X(x) - P_X(x)}{P_X(x)}\,, \tag{4.15}$$

$$\forall y \in \mathcal{Y},\ \ g(y) = (C(f))(y) = \frac{R_Y(y) - P_Y(y)}{P_Y(y)}\,. \tag{4.16}$$

Then, we have:

$$\chi^2(R_Y||P_Y) = \mathbb{E}\Big[g(Y)^2\Big] = \|C(f)\|^2_{P_Y} \le \|f\|^2_{P_X} = \mathbb{E}\Big[f(X)^2\Big] = \chi^2(R_X||P_X) \tag{4.17}$$

where we use (2.9) from chapter 2, and the inequality holds because C is a contraction. In fact, since the functions defined in (4.15) and (4.16) are zero mean, $\mathbb{E}[f(X)] = 0$ and $\mathbb{E}[g(Y)] = 0$, we can also obtain the SDPI for χ^2-divergence (with fixed input pmf P_X):

$$\chi^2(R_Y||P_Y) = \|C(f)\|^2_{P_Y} \le \sigma_2^2\,\|f\|^2_{P_X} = \rho_{\mathsf{max}}(X;Y)^2\chi^2(R_X||P_X) \tag{4.18}$$

where the inequality holds due to part 2 of Theorem 4.1 (since f is orthogonal to the dominant right singular vector $f_1 = \mathbf{1}$), and the final equality holds due to part 3 of Theorem 4.1 (where $\rho_{\mathsf{max}}(X;Y)^2$ is precisely the contraction coefficient $\eta_{\chi^2}(P_X, P_{Y|X})$ using (2.37)). Therefore, the contraction property of the conditional expectation operator C corresponds to the DPI for χ^2-divergence, and finer knowledge about the SVD of C yields "stronger" DPIs for χ^2-divergence. In appendix C.3, we convey that our choices of inner products to define the input and output Hilbert spaces of C are essential in ensuring that C is a contraction.

Part 2 of Theorem 4.1 portrays that ψ_1 and ϕ_1 are determined by the marginal pmfs P_X and P_Y, respectively. This suggests that the statistical dependence between X and Y is (at least intuitively) captured by the remaining pairs of singular vectors of B. Therefore, in our ensuing discussion, we will sometimes focus on the so called canonical dependence matrix (CDM) corresponding to $P_{X,Y}$, which removes the first pair of singular vectors from B. The CDM corresponding to $P_{X,Y}$ is defined next.

Definition 4.2 (Canonical Dependence Matrix). *For any joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, we define its corresponding* canonical dependence matrix $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ *entry-wise as:*

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \quad \left[\tilde{B}\right]_{y,x} \triangleq \frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}},$$

or equivalently, in matrix notation as:

$$\tilde{B} = B - \sqrt{P_Y}^T \sqrt{P_X}\,.$$

So far, we have shown that the DTM B is an equivalent description of the conditional expectation operator C, and explored some properties of its SVD. However, it is not obvious whether the DTM B uniquely identifies the joint pmf $P_{X,Y}$. To address this question, we consider the matrix-valued function $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ that maps joint pmfs in $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ to their corresponding DTMs. In particular, for every $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, we define $\beta(P_{X,Y})$ entry-wise according to Definition 4.1:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \quad [\beta(P_{X,Y})]_{y,x} \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}}\,. \tag{4.19}$$

Furthermore, we let $\mathcal{B}$ denote the range of β:

$$\mathcal{B} \triangleq \left\{B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \exists P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}},\, \beta(P_{X,Y}) = B\right\} \tag{4.20}$$

which is the set of all possible DTMs, and $\mathcal{B}^\circ$ denote the range of β restricted to the domain $\mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}}$:

$$\mathcal{B}^\circ \triangleq \left\{B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : \exists P_{X,Y} \in \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}},\, \beta(P_{X,Y}) = B\right\}. \tag{4.21}$$

The next theorem characterizes both $\mathcal{B}$ and $\mathcal{B}^\circ$, and proves that β is a bijective and continuous map.

Theorem 4.2 (Characterization of DTMs). *The following statements are true:*

1. *A matrix $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a joint pmf in $\mathcal{P}^{\circ}_{\mathcal{X}\times\mathcal{Y}}$ if and only if it is entry-wise strictly positive and has largest singular value of unity:*

$$\mathcal{B}^{\circ} = \left\{B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : B > 0 \text{ entry-wise and } \|B\|_{\mathsf{op}} = 1\right\}.$$

2. *A matrix $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is a DTM corresponding to a joint pmf in $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ if and only if it is entry-wise non-negative, it has largest singular value of unity, and both B^TB and BB^T have entry-wise strictly positive eigenvectors corresponding to the eigenvalue of unity:*

$$\begin{aligned}\mathcal{B} = \Big\{B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : B \geq 0 \text{ entry-wise},\\ \|B\|_{\mathsf{op}} = 1,\\ \exists \text{ entry-wise strictly positive } \psi \in \mathbb{R}^{|\mathcal{X}|},\ B^TB\psi = \psi,\\ \text{and } \exists \text{ entry-wise strictly positive } \phi \in \mathbb{R}^{|\mathcal{Y}|},\ BB^T\phi = \phi\Big\}.\end{aligned}$$

3. *The map $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}$ is bijective and continuous.*

Proof.

Part 1: The prove the $\subseteq$ direction, consider any matrix $B \in \mathcal{B}^{\circ}$. Then, there exists $P_{X,Y} \in \mathcal{P}^{\circ}_{\mathcal{X}\times\mathcal{Y}}$ such that $\beta(P_{X,Y}) = B$. Clearly, $B > 0$ entry-wise, and part 1 of Theorem 4.1 implies that $\|B\|_{\mathsf{op}} = 1$.

To prove the $\supseteq$ direction, consider any matrix $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ satisfying $B > 0$ entry-wise and $\|B\|_{\mathsf{op}} = 1$. Then, its Gramian and dual Gramian matrices are entry-wise strictly positive, and have spectral radius (and largest eigenvalue) of 1: $B^TB > 0$ entry-wise, $BB^T > 0$ entry-wise, and $\rho(B^TB) = \rho(BB^T) = \|B\|^2_{\mathsf{op}} = 1$. Applying the *Perron-Frobenius theorem* [129, Theorem 8.2.2], we get:

$$B^TB\psi = \psi \quad \text{and} \quad BB^T\phi = \phi$$

where $\psi \in \mathbb{R}^{|\mathcal{X}|}$ and $\phi \in \mathbb{R}^{|\mathcal{Y}|}$ are entry-wise strictly positive eigenvectors corresponding to the spectral radius (or largest eigenvalue) of 1 such that $\|\psi\|_2^2 = \|\phi\|_2^2 = 1$. This implies that ψ and ϕ are the right and left singular vectors corresponding to $\|B\|_{\mathsf{op}} = 1$, respectively, of B:

$$B\psi = \phi \quad \text{and} \quad B^T\phi = \psi\,. \tag{4.22}$$

Define the matrix:

$$P \triangleq \mathsf{diag}(\phi)B\,\mathsf{diag}(\psi) \tag{4.23}$$

and the corresponding "candidate" joint pmf:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ P_{X,Y}(x,y) \triangleq [P]_{y,x}\,. \tag{4.24}$$

We now verify that this candidate $P_{X,Y}$ is truly a joint pmf with the desired properties. Observe that:

$$\mathbf{1}^T P \mathbf{1} = \mathbf{1}^T \mathsf{diag}(\phi) B \,\mathsf{diag}(\psi) \mathbf{1} = \phi^T B \psi = \phi^T \phi = 1 \tag{4.25}$$

$$P\mathbf{1} = \mathsf{diag}(\phi) B \,\mathsf{diag}(\psi) \mathbf{1} = \mathsf{diag}(\phi) B \psi = \mathsf{diag}(\phi)\phi = \phi^2 \tag{4.26}$$

$$P^T \mathbf{1} = \mathsf{diag}(\psi) B^T \mathsf{diag}(\phi) \mathbf{1} = \mathsf{diag}(\psi) B^T \phi = \mathsf{diag}(\psi)\psi = \psi^2 \tag{4.27}$$

where we repeatedly use (4.22), and $\psi^2 \in \mathbb{R}^{|\mathcal{X}|}$ and $\phi^2 \in \mathbb{R}^{|\mathcal{Y}|}$ are vectors whose entries are the element-wise squares of ψ and ϕ, respectively. Then, $P_{X,Y} \in \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}}$ since (4.25) holds and $P > 0$ entry-wise. Furthermore, the corresponding marginals are:

$$\forall x \in \mathcal{X},\;\; P_X(x) = \psi(x)^2 \quad \text{and} \quad \forall y \in \mathcal{Y},\;\; P_Y(y) = \phi(y)^2$$

using (4.26) and (4.27). Finally, since $\beta(P_{X,Y}) = \mathsf{diag}(\phi)^{-1} P \,\mathsf{diag}(\psi)^{-1} = B$, where the inverses are well-defined because ψ and ϕ are entry-wise strictly positive, we have that $B \in \mathcal{B}^\circ$.

Part 2: To prove the $\subseteq$ direction, consider any matrix $B \in \mathcal{B}$. Then, there exists $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ such that $\beta(P_{X,Y}) = B$. Clearly, $B \geq 0$ entry-wise, and parts 1 and 2 of Theorem 4.1 imply that $\|B\|_{\mathsf{op}} = 1$ with corresponding right and left singular vectors $\sqrt{P_X}^T$ and $\sqrt{P_Y}^T$, respectively:

$$B\sqrt{P_X}^T = \sqrt{P_Y}^T \quad \text{and} \quad B^T\sqrt{P_Y}^T = \sqrt{P_X}^T.$$

Hence, $B^T B$ and BB^T have entry-wise strictly positive eigenvectors $\sqrt{P_X}^T$ and $\sqrt{P_Y}^T$, respectively, corresponding to the eigenvalue of 1 (where the strict positivity holds by definition of $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$).

To prove the $\supseteq$ direction, we can follow the proof of part 1 mutatis mutandis. However, we must be careful when applying the Perron-Frobenius theorem [129, Theorem 8.3.1] to $B^T B \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ and $BB^T \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ as it only guarantees that the eigenvectors $\psi \in \mathbb{R}^{|\mathcal{X}|}$ and $\phi \in \mathbb{R}^{|\mathcal{Y}|}$ are entry-wise non-negative. If an entry of ψ or ϕ is zero, then the corresponding column or row of $P = \mathsf{diag}(\phi) B \,\mathsf{diag}(\psi)$ is zero. Since we use P to define the joint pmf $P_{X,Y}$ via (4.24), this implies that $P_X \notin \mathcal{P}^\circ_{\mathcal{X}}$ or $P_Y \notin \mathcal{P}^\circ_{\mathcal{Y}}$, which means that $P_{X,Y} \notin \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$—a contradiction. Hence, we enforce the entry-wise strict positivity constraints on ψ and ϕ in the theorem statement to ensure that the proof in part 1 holds.

Part 3: The map $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}$ is bijective because its range is defined as $\mathcal{B}$ and the proofs of parts 1 and 2 delineate the inverse function (see e.g. (4.23)). (Note that (4.23) is uniquely defined because the Gramian and dual Gramian matrices of B have only one entry-wise strictly positive eigenvector each.)

To prove that $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}$ is continuous, consider any sequence of joint pmfs $\{Q^n_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} : n \in \mathbb{N}\}$ that converge to $Q_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\;\; \lim_{n\to\infty} Q^n_{X,Y}(x,y) = Q_{X,Y}(x,y)\,.$$

By the triangle inequality, we have for all $x \in \mathcal{X}$:

$$|Q_X^n(x) - Q_X(x)| = \left| \sum_{y \in \mathcal{Y}} Q_{X,Y}^n(x,y) - Q_{X,Y}(x,y) \right| \leq \sum_{y \in \mathcal{Y}} \left| Q_{X,Y}^n(x,y) - Q_{X,Y}(x,y) \right|$$

which implies that for all $x \in \mathcal{X}$, $\lim_{n\to\infty} Q_X^n(x) = Q_X(x)$. Likewise, for all $y \in \mathcal{Y}$, $\lim_{n\to\infty} Q_Y^n(y) = Q_Y(y)$. Hence, we have:

$$\lim_{n\to\infty} \left[\beta(Q_{X,Y}^n)\right]_{y,x} = \lim_{n\to\infty} \frac{Q_{X,Y}^n(x,y)}{\sqrt{Q_X^n(x) Q_Y^n(y)}} = \frac{Q_{X,Y}(x,y)}{\sqrt{Q_X(x) Q_Y(y)}} = \left[\beta(Q_{X,Y})\right]_{y,x}$$

for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. (Note that the denominators are strictly positive as $Q_{X,Y}^n, Q_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.) Therefore, $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}$ is continuous. ■

The proofs of parts 1 and 2, and the statement of part 3 of Theorem 4.2 illustrate that the DTM B is indeed an equivalent description of the joint pmf $P_{X,Y}$. Furthermore, in the context of part 2 of Theorem 4.2, we remark that an entry-wise non-negative square matrix A has strictly positive left and right eigenvectors corresponding to its Perron-Frobenius eigenvalue (or spectral radius) $\rho(A)$ if and only if the *triangular block form* of A is a direct sum of *irreducible* entry-wise non-negative square matrices whose spectral radii are also $\rho(A)$—see Theorem 3.14 and the preceding discussion in [26, Chapter 2, Section 3]. This means that $B^T B$ and BB^T have strictly positive eigenvectors corresponding to their spectral radius of unity if and only if they have the aforementioned direct form structure after suitable similarity transformations using permutation matrices.

■ 4.2.2 Variational Characterizations of Maximal Correlation Functions

In this subsection, under the setup of subsection 4.2.1, we present two well-known variational characterizations of the SVD structure of the DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ and the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ corresponding to a fixed bivariate distribution $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. These characterizations will be useful in the development of future sections. Our first proposition characterizes the singular values and singular vectors of B and C using a variant of the Courant-Fischer-Weyl min-max theorem (cf. Theorem C.1 in appendix C.1, [128, 129, 270]).

Proposition 4.1 (Courant-Fischer-Weyl Variational Characterization). *For any $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$, define the sets of k-tuples of zero mean orthonormal functions:*

$$\begin{aligned} \mathcal{S}_k(\mathcal{X}, P_X) \triangleq \Big\{ (r_2, \ldots, r_{k+1}) \in \mathcal{L}^2(\mathcal{X}, P_X)^k : \\ \forall i \in \{2, \ldots, k+1\},\ \mathbb{E}[r_i(X)] = 0\,, \\ \forall i, j \in \{2, \ldots, k+1\},\ \mathbb{E}[r_i(X) r_j(X)] = \mathbb{1}\{i = j\} \Big\}\,, \end{aligned} \tag{4.28}$$

$$\mathcal{S}_k(\mathcal{Y}, P_Y) \triangleq \Big\{(s_2, \ldots, s_{k+1}) \in \mathcal{L}^2(\mathcal{Y}, P_Y)^k : \\ \forall i \in \{2, \ldots, k+1\},\ \mathbb{E}[s_i(Y)] = 0\,, \\ \forall i, j \in \{2, \ldots, k+1\},\ \mathbb{E}[s_i(Y)s_j(Y)] = \mathbb{1}\{i = j\}\Big\}. \tag{4.29}$$

Then, the $(k+1)$th largest singular value of B and C is given by:

$$\begin{aligned}\sigma_{k+1} &= \max_{\substack{V_k=[v_2 \cdots v_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|}),\\ U_k=[u_2 \cdots u_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|}):\\ \sqrt{P_X}V_k = \boldsymbol{0}^T,\, \sqrt{P_Y}U_k = \boldsymbol{0}^T}} \ \min_{i \in \{2,\ldots,k+1\}} u_i^T B v_i \\ &= \max_{\substack{(r_2,\ldots,r_{k+1}) \in \mathcal{S}_k(\mathcal{X}, P_X),\\ (s_2,\ldots,s_{k+1}) \in \mathcal{S}_k(\mathcal{Y}, P_Y)}} \ \min_{i \in \{2,\ldots,k+1\}} \mathbb{E}[r_i(X)s_i(Y)]\end{aligned}$$

where the first maximization is over all orthonormal sets $\{v_2, \ldots, v_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{X}|}$ and $\{u_2, \ldots, u_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{Y}|}$ such that $\sqrt{P_X}v_i = 0$ and $\sqrt{P_Y}u_i = 0$ for all $i \in \{2, \ldots, k+1\}$, and the second maximization is over $\mathcal{S}_k(\mathcal{X}, P_X)$ and $\mathcal{S}_k(\mathcal{Y}, P_Y)$. Moreover, the vectors that maximize the first formulation are the singular vectors of B:

$$v_i^* = \psi_i \quad \textit{and} \quad u_i^* = \phi_i$$

for every $i \in \{2, \ldots, k+1\}$, and the functions that maximize the second formulation are the singular vectors of C:

$$r_i^* = f_i \quad \textit{and} \quad s_i^* = g_i$$

for every $i \in \{2, \ldots, k+1\}$.

Proof. The first max-min formulation of σ_{k+1} in terms of the DTM B is an immediate consequence of the alternative version of the Courant-Fischer-Weyl min-max theorem in [42, Theorem 1.2]. The second max-min formulation of σ_{k+1} follows from the first formulation. Indeed, corresponding to each feasible pair of orthonormal sets in the first formulation, $\{v_2, \ldots, v_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{X}|}$ and $\{u_2, \ldots, u_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{Y}|}$, we can construct a feasible pair of k-tuples of zero mean orthonormal functions in the second formulation, $(r_2, \ldots, r_{k+1}) \in \mathcal{S}_k(\mathcal{X}, P_X)$ and $(s_2, \ldots, s_{k+1}) \in \mathcal{S}_k(\mathcal{Y}, P_Y)$, using the relations (4.6) and (4.8), so that:

$$\forall x \in \mathcal{X},\ \ r_i(x) = \frac{v_i(x)}{\sqrt{P_X(x)}} \quad \text{and} \quad \forall y \in \mathcal{Y},\ \ s_i(y) = \frac{u_i(y)}{\sqrt{P_Y(y)}} \tag{4.30}$$

for every $i \in \{2, \ldots, k+1\}$. Furthermore, with these choices of arguments (related by (4.30)), the objective functions of the two formulations are equal since:

$$\forall i \in \{2, \ldots, k+1\},\ \ u_i^T B v_i = \langle s_i, C(r_i)\rangle_{P_Y} = \mathbb{E}[s_i(Y)\mathbb{E}[r_i(X)|Y]] = \mathbb{E}[r_i(X)s_i(Y)]$$

where the first equality holds due to the equivalences (4.30), (4.7), and (4.9), and the final equality follows from the tower property. This proves the second max-min formulation. Lastly, the maximizing arguments of both formulations can be obtained from the SVDs of B and C. ■

We note that Proposition 4.1 can be perceived as a generalization of parts 3 and 4 of Theorem 4.1, because the second max-min characterization of σ_2 in Proposition 4.1 coincides exactly with the variational problem that defines maximal correlation in (4.1). Thus, the second max-min characterization of the kth largest singular value $\sigma_k \in [0,1]$ of B and C for general $k \in \{2,\dots,\min\{|\mathcal{X}|,|\mathcal{Y}|\}\}$ in Proposition 4.1 portrays that σ_k is a generalization of maximal correlation. For these reasons, we refer to the singular vectors $\{f_2,\dots,f_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}\} \subseteq \mathcal{L}^2(\mathcal{X},P_X)$ and $\{g_2,\dots,g_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}\} \subseteq \mathcal{L}^2(\mathcal{Y},P_Y)$ of C as *maximal correlation functions*. It is straightforward to see from Proposition 4.1 that σ_k is given by the Pearson correlation coefficient between the corresponding maximal correlation functions:

$$\forall k \in \{2,\dots,\min\{|\mathcal{X}|,|\mathcal{Y}|\}\},\ \ \sigma_k = \mathbb{E}[f_k(X)g_k(Y)]\,. \tag{4.31}$$

In fact, the Courant-Fischer-Weyl min-max theorem (cf. Theorem C.1 in appendix C.1) also shows that σ_k for $k \in \{3,\dots,\min\{|\mathcal{X}|,|\mathcal{Y}|\}\}$ is obtained by maximizing the Pearson correlation coefficient between the real-valued features $f(X)$ and $g(Y)$ subject to the constraints that $f:\mathcal{X}\to\mathbb{R}$ and $g:\mathcal{Y}\to\mathbb{R}$ are orthogonal to all previous maximal correlation functions $f_2,\dots,f_{k-1}$ and $g_2,\dots,g_{k-1}$, respectively.

From the feature extraction perspective introduced at the outset of this chapter, it is often desirable to find embeddings of the categorical random variables X and Y into $\mathbb{R}^k$ (with $k \in \{2,\dots,\min\{|\mathcal{X}|,|\mathcal{Y}|\}-1\}$) that enable further simple processing in the future, e.g. using the (Lloyd-Max) K-means clustering algorithm [177, 196]. However, the classical maximal correlation functions (f_2,g_2) only provide a single pair of features $(f_2(X),g_2(Y))$ that embed X and Y into $\mathbb{R}$. In order to obtain embeddings of X and Y into $\mathbb{R}^k$, we must use a k-tuple of pairs of real-valued features of X and Y, where we may assume without loss of generality that the feature functions have zero mean and unit variance. It is of course intuitively desirable that:

1. For each pair of features $(f(X),g(Y))$, $f(X)$ and $g(Y)$ have high correlation.
2. The k feature functions of X carry orthogonal modes of information to avoid redundancy, and likewise, the k feature functions of Y carry orthogonal modes of information.

The earlier discussion shows that the pairs of maximal correlation functions $(f_2,g_2),\dots,(f_{k+1},g_{k+1})$ yield an embedding of X and Y into $\mathbb{R}^k$ with the above properties. (In particular, each pair of zero mean, unit variance features $(f_i(X),g_i(Y))$ is maximally correlated subject to being orthogonal to all previous maximal correlation functions.)

While Proposition 4.1 provides a variational characterization for the dominant $k \in \{1,\dots,\min\{|\mathcal{X}|,|\mathcal{Y}|\}-1\}$ maximal correlation functions by maximizing the minimum correlation between zero mean orthonormal pairs of functions, there are several other possible variational characterizations of the top k maximal correlation functions. We present one such alternative characterization based on *Ky Fan's extremum principle*, cf. [128, Theorem 3.4.1], which we will utilize later.

Proposition 4.2 (Ky Fan-von Neumann Variational Characterization). *For any $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$, the Ky Fan k-norm of the CDM $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ is given by:*

$$\left\|\tilde{B}\right\|_{(1,k)} = \sum_{i=2}^{k+1} \sigma_i = \max_{\substack{V_k \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|}),\, U_k \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|}): \\ \sqrt{P_X} V_k = \boldsymbol{0}^T,\, \sqrt{P_Y} U_k = \boldsymbol{0}^T}} \operatorname{tr}\left(U_k^T B V_k\right)$$

$$= \max_{\substack{(r_2, \ldots, r_{k+1}) \in \mathcal{S}_k(\mathcal{X}, P_X), \\ (s_2, \ldots, s_{k+1}) \in \mathcal{S}_k(\mathcal{Y}, P_Y)}} \sum_{i=2}^{k+1} \mathbb{E}[r_i(X) s_i(Y)]$$

where the Ky Fan k-norm $\|\cdot\|_{(1,k)}$ is defined in (C.3) in appendix C.1, the first maximization is over all orthonormal k-frames $V_k = [v_2 \cdots v_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ and $U_k = [u_2 \cdots u_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|})$, with constituent columns $v_2, \ldots, v_{k+1} \in \mathbb{R}^{|\mathcal{X}|}$ and $u_2, \ldots, u_{k+1} \in \mathbb{R}^{|\mathcal{Y}|}$, respectively, such that $\sqrt{P_X} V_k = \boldsymbol{0}^T$ and $\sqrt{P_Y} U_k = \boldsymbol{0}^T$, and the second maximization is over $\mathcal{S}_k(\mathcal{X}, P_X)$ and $\mathcal{S}_k(\mathcal{Y}, P_Y)$ defined in (4.28) and (4.29), respectively. Moreover, the orthonormal k-frames $V_k^* = [v_2^* \cdots v_{k+1}^*] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ and $U_k^* = [u_2^* \cdots u_{k+1}^*] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|})$ that maximize the first formulation are composed of the singular vectors of B:

$$v_i^* = \psi_i \quad \text{and} \quad u_i^* = \phi_i$$

for every $i \in \{2, \ldots, k+1\}$, and the functions that maximize the second formulation are the singular vectors of C:

$$r_i^* = f_i \quad \text{and} \quad s_i^* = g_i$$

for every $i \in \{2, \ldots, k+1\}$.

Proof. The first formulation of $\|\tilde{B}\|_{(1,k)}$ in terms of the DTM B is an immediate consequence of Ky Fan's extremum principle, cf. [128, Theorem 3.4.1]. (Note that Ky Fan's extremum principle can be easily derived from *von Neumann's trace inequality* [129, Theorem 7.4.1.1], which is why we refer to the extremizations in this proposition as "Ky Fan-von Neumann variational characterizations."). The second formulation of $\|\tilde{B}\|_{(1,k)}$ follows from the first formulation. Indeed, much like the proof of Proposition 4.1, corresponding to each feasible pair of orthonormal k-frames in the first formulation, $V_k = [v_2 \cdots v_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ and $U_k = [u_2 \cdots u_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|})$, we can construct a feasible pair of k-tuples of zero mean orthonormal functions in the second formulation, $(r_2, \ldots, r_{k+1}) \in \mathcal{S}_k(\mathcal{X}, P_X)$ and $(s_2, \ldots, s_{k+1}) \in \mathcal{S}_k(\mathcal{Y}, P_Y)$, using the relations (4.30) for all $i \in \{2, \ldots, k+1\}$. Furthermore, with these choices of related arguments, the objective functions of the two formulations are equal since:

$$\operatorname{tr}\left(U_k^T B V_k\right) = \sum_{i=2}^{k+1} u_i^T B v_i = \sum_{i=2}^{k+1} \langle s_i, C(r_i) \rangle_{P_Y} = \sum_{i=2}^{k+1} \mathbb{E}[r_i(X) s_i(Y)]$$

where the second equality holds due to the equivalences (4.30), (4.7), and (4.9), and the final equality holds as before. This proves the second formulation of $\|\tilde{B}\|_{(1,k)}$. Lastly, the maximizing arguments of both formulations can be obtained from the SVDs of B and C. ■

■ 4.2.3 Modal Decompositions

We are finally in a position to present the *modal decomposition* of a bivariate distribution $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, which was discovered by Hirschfeld in [125], and independently developed by Lancaster in [165]. Although the decomposition is an immediate consequence of our discussion so far, we present it as a theorem due to its historical significance.

Theorem 4.3 (Modal Decomposition [125, 165]). *Consider any bivariate distribution $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. Then, the following statements are true:*

1. *$P_{X,Y}$ exhibits the following modal decomposition:*

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; P_{X,Y}(x,y) = P_X(x)\,P_Y(y)\left(1 + \sum_{i=2}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \sigma_i\, f_i(x)\, g_i(y)\right)$$

   where $1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \geq 0$ are the ordered singular values of the CDM $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$, which form a sequence of non-negative correlations:

$$\forall i \in \{2, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}, \;\; \sigma_i = \mathbb{E}[f_i(X) g_i(Y)]\,,$$

   and $f_2, \ldots, f_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g_2, \ldots, g_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ are the corresponding maximal correlation functions, which are singular vectors of the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$.

2. *The mutual χ^2-information between X and Y can be decomposed in terms of the aforementioned non-negative correlations:*

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y}||P_X P_Y) = \left\|\tilde{B}\right\|_{\mathsf{Fro}}^2 = \sum_{i=2}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \sigma_i^2\,.$$

Proof. As we mentioned earlier, part 1 of this result follows from the SVD structure of the DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ or the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$, which is discussed in subsections 4.2.1 and 4.2.2. Part 2 follows from part 1 and the definition of χ^2-divergence in (2.9) in chapter 2.[52] ■

[52] We remark that the proof of Theorem 4.3 in [125] uses the eigen-decomposition of $B^T B$ rather than the SVD of B. Similarly, the proof of part 3 of Theorem 4.1 (which states that the second largest singular value of a compact conditional expectation operator C is equal to maximal correlation) in [236] also analyzes the eigen-decomposition of the composition of C with its adjoint operator rather than the SVD of C. This suggests that although the SVD had been developed in the nineteenth century, it had not gained its modern widespread appeal within the probability and statistics communities in the mid-twentieth century.

Theorem 4.3 elegantly decomposes the statistical dependence between two random variables X and Y into orthogonal modes, and elucidates the relative importance of these modes via the singular values of the CDM. The modal decomposition of $P_{X,Y}$ in Theorem 4.3 has been the basis of a statistical technique known as correspondence analysis, which originated in [125], and was rediscovered and developed by the French school of data analysis [24] (also see [110,111]). While correspondence analysis has mainly been used as an exploratory data visualization tool, cf. [110], our development of modal decompositions reveals their fundamental role in modern data science and machine learning applications. Indeed, we demonstrate that the maximal correlation functions yield real-valued features $\{f_2(X), \ldots, f_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}(X)\}$ and $\{g_2(Y), \ldots, g_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}(Y)\}$ that carry orthogonal modes of information and aptly summarize the salient dependencies between X and Y for an unspecified inference task.

We also remark that the decomposition of mutual χ^2-information in part 2 of Theorem 4.3 yields an illustrative upper bound on (standard) mutual information. Observe that using Lemma 2.3 from chapter 2, we get:

$$I(X;Y) \leq \log\Big(1 + I_{\chi^2}(X;Y)\Big) = \log\left(\sum_{i=1}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \sigma_i^2\right) = 2\log(\|B\|_{\mathsf{Fro}}) \tag{4.32}$$

where $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is the DTM corresponding to $P_{X,Y}$, and we use the fact that $\sigma_1 = 1$ (see part 1 of Theorem 4.1).

■ 4.3 Local Information Geometry

In this section, we develop some local information geometric structure and use it to illustrate why maximal correlation functions yield useful real-valued features for unspecified inference tasks.

■ 4.3.1 Information Vectors and Feature Functions

We commence by recalling aspects of the discussion pertaining to the "local quadratic behavior" of f-divergences at the end of subsection 2.2.1. To introduce the local geometric structure on $\mathcal{P}_{\mathcal{X}}$, we fix any *reference distribution* $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$. Then, we consider a *local perturbation* $R_X^{(\epsilon)} \in \mathcal{P}_{\mathcal{X}}$ of the reference distribution P_X:

$$\forall x \in \mathcal{X}, \quad R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}\,\psi(x) \tag{4.33}$$

$$= P_X(x)\,(1 + \epsilon f(x)) \tag{4.34}$$

where the spherical perturbation vector $\psi = [\psi(1) \cdots \psi(|\mathcal{X}|)]^T \in \mathbb{R}^{|\mathcal{X}|}$, cf. (2.21) and (2.22) in chapter 2, satisfies the constraints:

$$\sqrt{P_X}\psi = 0\,, \tag{4.35}$$

$$\|\psi\|_2^2 = 1\,, \tag{4.36}$$

the *multiplicative perturbation* function $f \in \mathcal{L}^2(\mathcal{X}, P_X)$, which is related to ψ via (4.6), satisfies the equivalent constraints:

$$\langle f, \mathbf{1} \rangle_{P_X} = \mathbb{E}[f(X)] = 0\,, \tag{4.37}$$

$$\|f\|_{P_X}^2 = \mathbb{E}\Big[f(X)^2\Big] = 1\,, \tag{4.38}$$

and $\epsilon \in \mathbb{R}\backslash\{0\}$ is a sufficiently small scaling parameter so that $R_X^{(\epsilon)}$ is a valid pmf. The orthogonality constraints in (4.35) and (4.37) ensure that $R_X^{(\epsilon)}$ sums to unity, and the unit norm constraints in (4.36) and (4.38) are imposed without loss of generality. We note that instead of considering local perturbations of P_X, it is possible to define a local neighborhood, or more precisely, a χ^2-divergence ball with radius ϵ^2 around P_X and proceed with our analysis using such local neighborhoods. However, we omit an exposition of local neighborhoods from this chapter because they are not required to illustrate our main ideas.

Inspired by the spherical and multiplicative perturbations in (4.33) and (4.34), we define the ensuing notions of "information vectors" and "feature functions."

Definition 4.3 (Information Vector). *We refer to any vector $\psi \in \mathbb{R}^{|\mathcal{X}|}$ (with standard Euclidean inner product) as an* information vector *if it is orthogonal to $\sqrt{P_X}^T$ and unit norm, i.e. if it satisfies* (4.35) *and* (4.36).

Definition 4.4 (Feature Function). *We refer to any function $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ as a* feature function *if it has zero mean and unit variance, i.e. if it satisfies* (4.37) *and* (4.38).

While information vectors can be used to define spherical perturbations of P_X and feature functions can be used to define multiplicative perturbations of P_X, feature functions can more generally be viewed as a means of extracting relevant features from data or embedding categorical data into $\mathbb{R}$. From this standpoint, the zero mean and unit variance constraints in Definition 4.4 (and the equivalent constraints in Definition 4.3) are reasonable since they do not hinder us from extracting useful information from data. We note that the relations (4.33), (4.34), and (4.6) portray a *three-way correspondence* between a locally perturbed pmf $R_X^{(\epsilon)}$ of P_X, an information vector ψ, and a feature function f:

$$R_X^{(\epsilon)} \;\leftrightarrow\; \psi \;\leftrightarrow\; f \tag{4.39}$$

and we will often use the equivalent information vector or feature function representations of $R_X^{(\epsilon)}$ for convenience. We next present a version of the local approximation result in (2.25) in chapter 2 specialized to the setting of KL divergence, cf. [53], [230, Section 4.2].

Proposition 4.3 (Local Approximation of KL Divergence). *Consider any two pmfs $R_X^{(\epsilon)}, Q_X^{(\alpha\epsilon)} \in \mathcal{P}_{\mathcal{X}}^{\circ}$ that are local perturbations of the reference pmf $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$:*

$$\forall x \in \mathcal{X},\;\; R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}\, v_1(x) = P_X(x)\,(1 + \epsilon\, r_1(x)) \tag{4.40}$$

$$\forall x \in \mathcal{X},\ \ Q_X^{(\alpha\epsilon)}(x) = P_X(x) + \alpha\epsilon\sqrt{P_X(x)}\, v_2(x) = P_X(x)\left(1 + \alpha\epsilon\, r_2(x)\right) \tag{4.41}$$

where $v_1, v_2 \in \mathbb{R}^{|\mathcal{X}|}$ are information vectors, $r_1, r_2 \in \mathcal{L}^2(\mathcal{X}, P_X)$ are feature functions, $\alpha \in \mathbb{R}$ is a fixed ratio between the scaling parameters of $Q_X^{(\alpha\epsilon)}$ and $R_X^{(\epsilon)}$, and $\epsilon \in \mathbb{R}$ is sufficiently small so that $R_X^{(\epsilon)}$ and $Q_X^{(\alpha\epsilon)}$ are valid strictly positive pmfs. Then, the KL divergence between $R_X^{(\epsilon)}$ and $Q_X^{(\alpha\epsilon)}$ can be locally approximated as:[53]

$$\begin{aligned} D(R_X^{(\epsilon)}||Q_X^{(\alpha\epsilon)}) &= \frac{1}{2}\epsilon^2 \left\|v_1 - \alpha v_2\right\|_2^2 + o(\epsilon^2) \\ &= \frac{1}{2}\epsilon^2 \left\|r_1 - \alpha r_2\right\|_{P_X}^2 + o(\epsilon^2) = \frac{1}{2}\epsilon^2\, \mathbb{E}\Big[\left(r_1(X) - \alpha r_2(X)\right)^2\Big] + o(\epsilon^2)\,. \end{aligned}$$

Proof. Observe that for all $x \in \mathcal{X}$:

$$\begin{aligned} \log\left(\frac{R_X^{(\epsilon)}(x)}{Q_X^{(\alpha\epsilon)}(x)}\right) &= \log\left(\frac{R_X^{(\epsilon)}(x)}{P_X(x)}\right) - \log\left(\frac{Q_X^{(\alpha\epsilon)}(x)}{P_X(x)}\right) \\ &= \log(1 + \epsilon r_1(x)) - \log(1 + \alpha\epsilon r_2(x)) \\ &= \epsilon\left(r_1(x) - \alpha r_2(x)\right) - \frac{1}{2}\epsilon^2 r_1(x)^2 + \frac{1}{2}\epsilon^2\alpha^2 r_2(x)^2 + o(\epsilon^2) \end{aligned} \tag{4.42}$$

where second equality follows from (4.40) and (4.41), and the third equality follows from the second order Taylor approximation of $x \mapsto \log(1+x)$ for $|x| < 1$. Then, taking expectations with respect to $R_X^{(\epsilon)}$ yields:

$$\begin{aligned} D(R_X^{(\epsilon)}||Q_X^{(\alpha\epsilon)}) &= \epsilon\, \mathbb{E}_{R_X^{(\epsilon)}}[r_1(X) - \alpha r_2(X)] - \frac{1}{2}\epsilon^2\, \mathbb{E}_{R_X^{(\epsilon)}}\Big[r_1(X)^2 - \alpha^2 r_2(X)^2\Big] + o(\epsilon^2) \\ &= \epsilon\, \mathbb{E}_{P_X}[r_1(X) - \alpha r_2(X)] + \epsilon^2\, \mathbb{E}_{P_X}\Big[r_1(X)^2 - \alpha r_1(X) r_2(X)\Big] \\ &\quad - \frac{1}{2}\epsilon^2\, \mathbb{E}_{P_X}\Big[r_1(X)^2 - \alpha^2 r_2(X)^2\Big] + o(\epsilon^2) \\ &= \frac{1}{2}\epsilon^2\, \mathbb{E}_{P_X}\Big[r_1(X)^2 - 2\alpha r_1(X) r_2(X) + \alpha^2 r_2(X)^2\Big] + o(\epsilon^2) \\ &= \frac{1}{2}\epsilon^2\, \mathbb{E}_{P_X}\Big[\left(r_1(X) - \alpha r_2(X)\right)^2\Big] + o(\epsilon^2) \end{aligned}$$

where the second equality follows from (4.40), and the third equality holds because r_1 and r_2 have zero mean. This proves the second local approximation in the proposition statement. The first local approximation trivially follows from the second because of the equivalence between information vectors and feature functions (expressed in (4.40) and (4.41)). ■

Proposition 4.3 portrays that the squared Euclidean ℓ^2-norms of spherical perturbation vectors, or equivalently, squared $\mathcal{L}^2(\mathcal{X}, P_X)$-norms of multiplicative perturbation

[53] The ratio α is a constant with respect to the scaling parameter ϵ (which tends to 0).

functions are good approximations of KL divergence under local perturbation assumptions. Indeed, setting $\alpha = 0$ in Proposition 4.3 recovers the KL divergence case of (2.25) in chapter 2. This conveys why we refer to normalized spherical perturbation vectors as information vectors. (It is worth mentioning that squared Euclidean ℓ^2-norms of spherical perturbation vectors also determine the χ^2-divergence between a perturbed pmf and the reference pmf without any local approximations.)

We close this subsection by providing a useful interpretation of feature functions. Suppose $\alpha = 1$, and consider the locally perturbed pmfs $R_X^{(\epsilon)}, Q_X^{(\epsilon)} \in \mathcal{P}_{\mathcal{X}}^{\circ}$ as defined in (4.40) and (4.41). It is well-known that log-likelihood ratio functions serve as useful sufficient statistics in various inference problems. Using (4.42) (with $\alpha = 0$), we can show that the log-likelihood ratio function $L_R : \mathcal{X} \to \mathbb{R}$ between $R_X^{(\epsilon)}$ and P_X is locally proportional to the corresponding feature function $r_1 \in \mathcal{L}^2(\mathcal{X}, P_X)$:

$$\forall x \in \mathcal{X}, \;\; L_R(x) \triangleq \log\left(\frac{R_X^{(\epsilon)}(x)}{P_X(x)}\right) = \epsilon r_1(x) + o(\epsilon)\,. \tag{4.43}$$

Similarly, the log-likelihood ratio function $L_Q : \mathcal{X} \to \mathbb{R}$ between $Q_X^{(\epsilon)}$ and P_X satisfies:

$$\forall x \in \mathcal{X}, \;\; L_Q(x) \triangleq \log\left(\frac{Q_X^{(\epsilon)}(x)}{P_X(x)}\right) = \epsilon r_2(x) + o(\epsilon)\,. \tag{4.44}$$

Therefore, feature functions locally represent log-likelihood ratios of locally perturbed pmfs to the reference pmf. Furthermore, (4.42) also implies that the log-likelihood ratio between $R_X^{(\epsilon)}$ and $Q_X^{(\epsilon)}$ is locally proportional to the difference between the corresponding feature functions:

$$\forall x \in \mathcal{X}, \;\; \log\left(\frac{R_X^{(\epsilon)}(x)}{Q_X^{(\epsilon)}(x)}\right) = L_R(x) - L_Q(x) = \epsilon\left(r_1(x) - r_2(x)\right) + o(\epsilon) \tag{4.45}$$

where we use (4.43) and (4.44). This portrays that the function $r_1 - r_2 \in \mathcal{L}^2(\mathcal{X}, P_X)$ contains all the information required to distinguish between $R_X^{(\epsilon)}(x)$ and $Q_X^{(\epsilon)}(x)$ in a binary hypothesis testing scenario under local approximations (because the log-likelihood ratio in (4.45) is a sufficient statistic for this problem).

■ 4.3.2 Local Geometry of Binary Hypothesis Testing

As a (non-rigorous) example of how to exploit the local approximation structure introduced in subsection 4.3.1, consider a binary hypothesis testing problem with hypothesis random variable $U \sim \mathsf{Bernoulli}(\frac{1}{2})$ (i.e. uniform Bernoulli prior), and likelihoods $P_{X|U=0} = R_X^{(\epsilon)}$ and $P_{X|U=1} = Q_X^{(\epsilon)}$ given by the locally perturbed pmfs defined in (4.40) and (4.41) (where we assume that $\alpha = 1$). Suppose we observe $n \in \mathbb{N}$ samples X_1^n that are drawn conditionally i.i.d. given U from the likelihoods:

$$\text{Given } U = 0: \;\; X_1^n \overset{\text{i.i.d.}}{\sim} P_{X|U=0} = R_X^{(\epsilon)}\,, \tag{4.46}$$

$$\text{Given } U=1: \quad X_1^n \overset{\text{i.i.d.}}{\sim} P_{X|U=1} = Q_X^{(\epsilon)}. \tag{4.47}$$

Then, the decision rule that minimizes the probability of error in inferring the hypothesis U based on the samples X_1^n is the *maximum likelihood* (ML) decision rule $\hat{U}_{\mathsf{ML}}^n : \mathcal{X}^n \to \{0,1\}$:

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{R_X^{(\epsilon)}(X_i)}{Q_X^{(\epsilon)}(X_i)}\right) = \widehat{\mathbb{E}}_n[L_R(X) - L_Q(X)] \underset{\hat{U}_{\mathsf{ML}}^n(X_1^n)=1}{\overset{\hat{U}_{\mathsf{ML}}^n(X_1^n)=0}{\gtrless}} 0 \tag{4.48}$$

where we use (4.45), and $\widehat{\mathbb{E}}_n[\cdot]$ denotes the empirical expectation operator corresponding to the empirical distribution $\hat{P}_{X_1^n}$ of the observations X_1^n (see (C.13) and (C.14) in appendix C.2). We next illustrate the elegant geometry associated with the ML decision rule.

For sufficiently large sample size n, since $\hat{P}_{X_1^n}$ is restricted to a small neighborhood around the true distribution that generates the i.i.d. samples X_1^n with high probability (cf. Theorem C.2 in appendix C.2), and the true distribution is a local perturbation of P_X, we may assume that $\hat{P}_{X_1^n}$ is also a local perturbation of P_X:

$$\forall x \in \mathcal{X}, \quad \hat{P}_{X_1^n}(x) = P_X(x) + \gamma\epsilon\sqrt{P_X(x)}\,\hat{\psi}(x) \tag{4.49}$$

$$= P_X(x)\Big(1 + \gamma\epsilon\hat{f}(x)\Big) \tag{4.50}$$

with information vector $\hat{\psi} \in \mathbb{R}^{|\mathcal{X}|}$ and feature function $\hat{f} \in \mathcal{L}^2(\mathcal{X}, P_X)$, where $\gamma > 0$ is a fixed ratio between the scaling parameters of $\hat{P}_{X_1^n}$ and $R_X^{(\epsilon)}, Q_X^{(\epsilon)}$. Thus, we can write:

$$\begin{aligned}
\widehat{\mathbb{E}}_n[L_R(X) - L_Q(X)] &= \mathbb{E}_{P_X}[L_R(X) - L_Q(X)] + \gamma\epsilon\,\mathbb{E}_{P_X}\Big[\hat{f}(X)\,(L_R(X) - L_Q(X))\Big] \\
&= \gamma\epsilon\,\mathbb{E}_{P_X}\Big[\hat{f}(X)\,(L_R(X) - L_Q(X))\Big] + o(\epsilon^2) \\
&= \gamma\epsilon^2\,\mathbb{E}_{P_X}\Big[\hat{f}(X)\,(r_1(X) - r_2(X))\Big] + o(\epsilon^2) \\
&= \gamma\epsilon^2\left\langle \hat{f}, r_1 - r_2\right\rangle_{P_X} + o(\epsilon^2) \qquad (4.51) \\
&= \gamma\epsilon^2\,\hat{\psi}^T(v_1 - v_2) + o(\epsilon^2) \qquad (4.52)
\end{aligned}$$

where the first equality follows from (4.50), the second equality follows from (4.42) and the zero mean and unit variance constraints on feature functions, the third equality follows from (4.43) and (4.44), and the final equality follows from (4.40), (4.41), (4.49), and (4.50). This implies that the ML decision rule in (4.48) can be described as:

$$\hat{\psi}^T(v_1 - v_2) + o(1) \underset{\hat{U}_{\mathsf{ML}}^n(X_1^n)=1}{\overset{\hat{U}_{\mathsf{ML}}^n(X_1^n)=0}{\gtrless}} 0 \tag{4.53}$$

under appropriate local approximation assumptions (with high probability for sufficiently large n).

The characterization in (4.53) has a beautiful geometric interpretation. It portrays that under local approximations, the ML decision rule simply projects the information vector $\hat{\psi}$ corresponding to the empirical distribution $\hat{P}_{X_1^n}$ onto the direction $v_1 - v_2$. In other words, it suffices to only monitor the spherical perturbation of $\hat{P}_{X_1^n}$ from P_X along one specific direction that is relevant to making decisions between $R_X^{(\epsilon)}$ and $Q_X^{(\epsilon)}$.

For inference problems based on i.i.d. samples (and memoryless noise models), the order of the data samples is irrelevant in the decision making.[54] So, the information contained in the data is carried by the empirical distribution. Under local approximations, evaluating the empirical expectation of various feature functions can therefore be viewed as monitoring the components of the spherical perturbation of the empirical distribution along different directions, or equivalently, as extracting different kinds of partial information. As we discussed at the outset of this chapter, when the desired latent (hypothesis) variable U and the likelihoods of the data X given U are known, we can easily determine which part of the information in the data is "useful." For example, in the binary hypothesis testing scenario above, the projection $\hat{\psi}^T(v_1 - v_2)$ is a (local) sufficient statistic of the data X_1^n for making decisions on U, and all other orthogonal components of $\hat{\psi}$ can be discarded. Without this knowledge, we cannot deem any part of the information in the data as irrelevant. However, processing, storage, or communication constraints often compel us to discard some partial information in high-dimensional problems. Intelligently doing this without severely degrading performance in future inference tasks requires some geometric structure to decompose information into parts that can potentially be dissipated. In the next subsection, we explain how modal decompositions address such lossy information processing problems.

■ 4.3.3 Feature Extraction using Modal Decompositions

We now illustrate how to find relevant features for an unspecified inference task based on modal decompositions. Suppose we observe i.i.d. samples $X_1, \ldots, X_n \in \mathcal{X}$ from some distribution that is a local perturbation of a reference distribution $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$. Our discussion on binary hypothesis testing conveys that the problem of decomposing the information contained in $\hat{P}_{X_1^n}$ into parts reduces to decomposing the information vector $\hat{\psi} \in \mathbb{R}^{|\mathcal{X}|}$ (cf. (4.49)) in terms of an orthonormal basis of information vectors $\{v_1, \ldots, v_{|\mathcal{X}|-1}\} \subseteq \mathbb{R}^{|\mathcal{X}|}$, or equivalently, decomposing the feature function $\hat{f} \in \mathcal{L}^2(\mathcal{X}, P_X)$ (cf. (4.50)) in terms of the orthonormal (or pairwise uncorrelated) basis of feature functions $\{r_1, \ldots, r_{|\mathcal{X}|-1}\} \subseteq \mathcal{L}^2(\mathcal{X}, P_X)$, where each r_i is related to v_i via (4.30).[55] Specifically, this entails computing the inner products or projection statistics:

$$\hat{\psi}^T v_i = \mathbb{E}_{P_X}\left[\hat{f}(X) r_i(X)\right] \propto \widehat{\mathbb{E}}_n[r_i(X)] = \frac{1}{n}\sum_{j=1}^{n} r_i(X_j) \tag{4.54}$$

[54] More generally, this is also true for *exchangeable* sampling models.

[55] Recall that $v_1, \ldots, v_{|\mathcal{X}|-1}$ are all orthogonal to $\psi_1 = \sqrt{P_X}^T$, and $r_1, \ldots, r_{|\mathcal{X}|-1}$ are all zero mean i.e. orthogonal to $f_1 = \mathbf{1}$ in the $\mathcal{L}^2(\mathcal{X}, P_X)$-inner product sense.

for all $i \in \{1, \ldots, |\mathcal{X}| - 1\}$, where the proportionality follows from (4.50). By the completeness of the basis $\{v_1, \ldots, v_{|\mathcal{X}|-1}\}$, we can recover $\hat{\psi}$ from the set of all inner products $\{\hat{\psi}^T v_i : i \in \{1, \ldots, |\mathcal{X}| - 1\}\}$. Hence, the set of real-valued projection statistics in (4.54) for $i \in \{1, \ldots, |\mathcal{X}| - 1\}$ decomposes the information contained in the empirical distribution $\hat{P}_{X_1^n}$ under local approximations. At this point, there is no reason to believe any projection statistic is more valuable or informative than any other projection statistic (irrespective of our choice of orthonormal basis). However, the story is different when we observe i.i.d. data through a memoryless noise model.

Fix any joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ with marginal reference pmfs $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, and let the conditional distribution $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denote the memoryless noise model or channel from $\mathcal{X}$ to $\mathcal{Y}$.[56] Since the row stochastic matrix corresponding to the channel $P_{Y|X}$ maps distributions in $\mathcal{P}_{\mathcal{X}}$ to distributions in $\mathcal{P}_{\mathcal{Y}}$ via left multiplication, it also maps information vectors in $\mathbb{R}^{|\mathcal{X}|}$ and feature functions in $\mathcal{L}^2(\mathcal{X}, P_X)$ to information vectors in $\mathbb{R}^{|\mathcal{Y}|}$ and feature functions in $\mathcal{L}^2(\mathcal{Y}, P_Y)$, respectively. In particular, it is straightforward to verify that the channel transformation on information vectors is given by the DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ (or the CDM $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$—see Definitions 4.1 and 4.2) corresponding to $P_{X,Y}$, and the the channel transformation on feature functions is given by the associated conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ (see (4.5)). So, the channel $P_{Y|X}$ maps the locally perturbed input pmf $R_X^{(\epsilon)} \in \mathcal{P}_{\mathcal{X}}^{\circ}$ in (4.33) and (4.34):

$$\forall x \in \mathcal{X}, \;\; R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}\,\psi(x) = P_X(x)\,(1 + \epsilon f(x)) \tag{4.55}$$

with scaling parameter $\epsilon \in \mathbb{R}$, information vector $\psi \in \mathbb{R}^{|\mathcal{X}|}$, and feature function $f \in \mathcal{L}^2(\mathcal{X}, P_X)$, to the locally perturbed output pmf $R_Y^{(\epsilon\tau)} \in \mathcal{P}_{\mathcal{Y}}^{\circ}$:

$$\forall y \in \mathcal{Y}, \;\; R_Y^{(\epsilon\tau)}(y) = P_Y(y) + \epsilon\tau\sqrt{P_Y(y)}\,\phi(y) = P_Y(y)\,(1 + \epsilon\tau g(y)) \tag{4.56}$$

with scaling parameter $\epsilon\tau \in \mathbb{R}$, information vector $\phi \in \mathbb{R}^{|\mathcal{Y}|}$, and feature function $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$, if and only if we have:

$$B\psi = \tilde{B}\psi = \tau\phi \tag{4.57}$$

$$C(f) = \tau g \tag{4.58}$$

where $\tau = \|B\psi\|_2 \in [0, 1]$ ensures that ϕ and g are normalized. Note that parts 1 and 2 of Theorem 4.1 ensure that $\tau \leq 1$ and that ϕ and g satisfy the relevant orthogonality properties.

By Proposition 4.3, the KL divergence between $R_X^{(\epsilon)}$ and P_X is given by:

$$D(R_X^{(\epsilon)}||P_X) = \frac{1}{2}\epsilon^2\,\|\psi\|_2^2 + o(\epsilon^2) = \frac{1}{2}\epsilon^2\,\mathbb{E}\Big[f(X)^2\Big] + o(\epsilon^2) \tag{4.59}$$

[56] Note that our development of the local information geometry of information vectors and feature functions trivially carries over to $\mathcal{P}_{\mathcal{Y}}$.

and the KL divergence between $R_Y^{(\epsilon\tau)}$ and P_Y is given by:

$$D(R_Y^{(\epsilon\tau)}||P_Y) = \frac{1}{2}\epsilon^2 \left\|B\psi\right\|_2^2 + o(\epsilon^2) = \frac{1}{2}\epsilon^2 \mathbb{E}\Big[\mathbb{E}[f(X)|Y]^2\Big] + o(\epsilon^2)\,. \tag{4.60}$$

This shows that for any information vector $\psi \in \mathbb{R}^{|\mathcal{X}|}$ or feature function $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ with fixed input KL divergence given by $\|\psi\|_2^2 = \mathbb{E}[f(X)^2] = 1$, the output KL divergence is determined by $\|B\psi\|_2^2 = \mathbb{E}[\mathbb{E}[f(X)|Y]^2]$, which depends on the direction of ψ or f and the (common) SVD structure of B and C. (This is why we refer to $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ as the divergence transition matrix.) Hence, depending on the SVD of B and C, some information vectors and feature functions are corrupted severely by the channel $P_{Y|X}$, while others remain more observable at the output end.

Parts 3 and 4 of Theorem 4.1 portray that when the input information vector $\psi = \psi_2$ (the dominant right singular vector of $\tilde{B}$), the output KL divergence is locally maximized so that $\|B\psi\|_2^2 = \|B\psi_2\|_2^2 = \sigma_2^2 = \rho_{\mathsf{max}}(X;Y)^2$, and the output information vector $\phi = \phi_2$ (the dominant left singular vector of $\tilde{B}$). Equivalently, the output KL divergence is locally maximized when the input feature function f is the maximal correlation function $f_2 \in \mathcal{L}^2(\mathcal{X}, P_X)$ in the modal decomposition of $P_{X,Y}$ (see Theorem 4.3), and the corresponding output feature function g is the maximal correlation function $g_2 \in \mathcal{L}^2(\mathcal{Y}, P_Y)$. More precisely, the specialization of Theorem 2.1 from chapter 2 to KL divergence yields:

$$\sigma_2^2 = \lim_{\epsilon\to 0^+} \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}}:\\ 0<D(R_X||P_X)\le\frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \tag{4.61}$$

where $R_Y \in \mathcal{P}_{\mathcal{Y}}$ is the marginal pmf of Y induced by R_X after it passes through the channel $P_{Y|X}$, and the supremum in (4.61) is achieved by the trajectory of locally perturbed pmfs (see subsection 2.3.1):

$$\forall x \in \mathcal{X},\;\; \tilde{R}_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}\,\psi_2(x) = P_X(x)\,(1 + \epsilon f_2(x)) \tag{4.62}$$

as $\epsilon \to 0^+$. Hence, while the DPI for KL divergence states that $R_X^{(\epsilon)}$ and P_X become less distinguishable after passing through the channel $P_{Y|X}$:[57]

$$D(R_Y^{(\epsilon\tau)}||P_Y) \le D(R_X^{(\epsilon)}||P_X)\,, \tag{4.63}$$

(4.61) conveys that the reduction of KL divergence is locally minimized when $R_X^{(\epsilon)} = \tilde{R}_X^{(\epsilon)}$, i.e. when $\psi = \psi_2$ or $f = f_2$. This means that the maximal correlation functions f_2 and g_2 are the feature functions which correspond to multiplicative perturbation directions that are least corrupted by the channel $P_{Y|X}$. More generally, Proposition 4.1 illustrates that to locally minimize the largest reduction of output KL divergence over a k-dimensional input spherical or multiplicative perturbation subspace

[57] It is well-known that KL divergence characterizes the error exponent in *Stein's regime* of binary hypothesis testing, cf. [230, Section 13.1].

with $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$, we must use the k-dimensional subspace of information vectors spanned by $\psi_2, \ldots, \psi_{k+1} \in \mathbb{R}^{|\mathcal{X}|}$, or feature functions spanned by the maximal correlation functions $f_2, \ldots, f_{k+1} \in \mathcal{L}^2(\mathcal{X}, P_X)$ (see Theorem 4.3). Therefore, the maximal correlation functions $f_2, \ldots, f_{k+1}$ and $g_2, \ldots, g_{k+1} \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ are the feature functions which correspond to the set of k uncorrelated multiplicative perturbation directions that are least corrupted by the channel $P_{Y|X}$. (We remark the the contraction of KL divergence under local approximations along different maximal correlation function directions can be construed as a finer kind of SDPI.)

Let us now state our high-dimensional feature extraction problem from a local information geometric lens. Suppose we only observe i.i.d. noisy outputs $Y_1, \ldots, Y_n$ of a *known* memoryless channel $P_{Y|X}$ with hidden i.i.d. inputs $X_1, \ldots, X_n$ from an unknown distribution that is a local perturbation of P_X. Since $|\mathcal{Y}|$ is very large in the high-dimensional regime, we want to discard parts of the information in the empirical distribution $\hat{P}_{Y_1^n} \in \mathcal{P}_{\mathcal{Y}}$:

$$\forall y \in \mathcal{Y}, \;\; \hat{P}_{Y_1^n}(y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i = y\}\,, \tag{4.64}$$

which we can assume is a local perturbation of P_Y with corresponding information vector $\hat{\phi} \in \mathbb{R}^{|\mathcal{Y}|}$ and feature function $\hat{g} \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ (with high probability for sufficiently large n, as explained in the previous subsection). So, instead of keeping all $|\mathcal{Y}| - 1$ real numbers that define $\hat{P}_{Y_1^n}$, we would like to store $k \in \{1, \ldots, |\mathcal{Y}| - 1\}$ real-valued features (where typically $k \ll |\mathcal{Y}|$). Unlike earlier, we now have a canonical choice of orthonormal basis of information vectors or feature functions to project on. In particular, our development so far portrays that we should compute the projection statistics (see Proposition 4.1):

$$\forall i \in \{2, \ldots, k+1\}, \;\; \hat{\phi}^T \phi_i = \mathbb{E}_{P_Y}[\hat{g}(Y) g_i(Y)] \propto \frac{1}{n} \sum_{j=1}^{n} g_i(Y_j) \tag{4.65}$$

corresponding to the k dominant left singular vectors of $\tilde{B}$, or the maximal correlation functions $g_2, \ldots, g_{k+1}$ in the modal decomposition of $P_{X,Y}$ (see Theorem 4.3). Furthermore, these statistics are defined by feature functions $g_2, \ldots, g_{k+1}$ of Y that are maximally correlated with corresponding feature functions $f_2, \ldots, f_{k+1}$ of X.

Returning to our discussion at the outset of this chapter, suppose we truly have a Markov model $U \to X_1^n \to Y_1^n$, where the latent variable U and its probability distribution P_U are unknown, $X_1, \ldots, X_n$ are conditionally i.i.d. given U with unknown conditional distribution $P_{X|U}$, and Y_1^n are noisy outputs of the data X_1^n from the known memoryless channel $P_{Y|X}$. We observe i.i.d. samples Y_1^n conditioned on some value of U, but the inference task of decoding U is unspecified because U, P_U, and $P_{X|U}$ are unknown. Under the assumption that all conditional pmfs in $P_{X|U}$ are local perturbations of P_X (which is known), if we are compelled to store only k real-valued features of Y_1^n for the purposes of inferring U (which may be revealed in the future), then it is

reasonable to compute the k projection statistics in (4.65) corresponding to the k dominant maximal correlation functions $g_2, \ldots, g_{k+1}$. Indeed, these projection directions are the most observable at the output end of the channel $P_{Y|X}$. This illustrates the utility of modal decompositions and maximal correlation functions for extracting useful features in high-dimensional scenarios. Moreover, in practice, it is convenient to use maximal correlation feature functions without considering how well our local approximation assumptions hold, because maximal correlation functions can be easily learned using structured and efficient algorithms (as the next section shows).

Since we will not examine local approximations in the remainder of this chapter, it is worth mentioning another elegant observation based on local approximations before we proceed to the next section. The ensuing proposition uses Proposition 4.3 to obtain a modal decomposition of mutual information, which parallels part 2 of Theorem 4.3 and (4.32), under a "weak dependence" assumption.

Proposition 4.4 (Modal Decomposition of Mutual Information). *Suppose the bivariate distribution $P_{X,Y}^{(\epsilon)} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}^{\circ}$ satisfies the* weak dependence *condition:*

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; P_{X,Y}^{(\epsilon)}(x,y) = P_X(x)P_Y(y) + \epsilon\sqrt{P_X(x)P_Y(y)}\,\zeta(x,y) \tag{4.66}$$

where $\epsilon \in \mathbb{R}$ is a sufficiently small scaling parameter, $\zeta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a fixed spherical perturbation such that:

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \sqrt{P_X(x)P_Y(y)}\,\zeta(x,y) = 0\,,$$

and $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ are the fixed marginal distributions of $P_{X,Y}^{(\epsilon)}$ (i.e. $P_{X,Y}^{(\epsilon)}$ is a local perturbation of the product distribution $P_X P_Y \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}^{\circ}$). Then, the mutual information between X and Y can be locally approximated as:

$$I_\epsilon(X;Y) \triangleq D(P_{X,Y}^{(\epsilon)}||P_X P_Y) = \frac{1}{2}\left\|\tilde{B}_\epsilon\right\|_{\mathsf{Fro}}^2 + o(\epsilon^2) = \frac{1}{2}\sum_{i=2}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \sigma_i(B_\epsilon)^2 + o(\epsilon^2)$$

where $\tilde{B}_\epsilon \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ and $B_\epsilon \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ denote the CDM and DTM associated with $P_{X,Y}^{(\epsilon)}$, respectively.

Proof. Using Proposition 4.3 with $\alpha = 0$, we get:

$$\begin{aligned} D(P_{X,Y}^{(\epsilon)}||P_X P_Y) &= \frac{1}{2}\epsilon^2 \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \zeta(x,y)^2 + o(\epsilon^2) \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \left(\frac{P_{X,Y}^{(\epsilon)}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}\right)^2 + o(\epsilon^2) \\ &= \frac{1}{2}\left\|\tilde{B}_\epsilon\right\|_{\mathsf{Fro}}^2 + o(\epsilon^2) \end{aligned}$$

where the second equality follows from (4.66), the third equality follows from Definition 4.2, and $\|\tilde{B}_\epsilon\|_{\mathsf{Fro}}^2$ scales likes ϵ^2 (when X and Y are not independent). ■

■ 4.4 Algorithm for Information Decomposition and Feature Extraction

We begin this section by recalling a complementary perspective that also demonstrates the utility of modal decompositions for the purposes of feature extraction. Given a bivariate distribution $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, it is often desirable to cluster the elements of $\mathcal{X}$ or $\mathcal{Y}$ in a manner that captures the important dependencies between X and Y. For example, in the "Netflix problem" [23], where $\mathcal{X}$ is the set of subscriber indices and $\mathcal{Y}$ is the set of movie indices, clustering the subscribers in $\mathcal{X}$ according to what movies they watch can help in recommendation systems since subscribers in the same cluster probably have similar tastes in movies. Likewise, clustering the movies in $\mathcal{Y}$ according to which subscribers watch them can potentially help identify movies in the same genre (without any genre labels). However, since our random variables are categorical, in order to utilize simple clustering algorithms such as K-means clustering, we must first embed elements of $\mathcal{X}$ or $\mathcal{Y}$ into points in k-dimensional Euclidean spaces with $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$. As we discussed in subsection 4.2.2, a natural choice of k real-valued features of X and Y that summarize the salient dependencies between X and Y for an unspecified inference task is to use the maximal correlation functions $\{f_2(X), \ldots, f_{k+1}(X)\} \subseteq \mathcal{L}^2(\mathcal{X}, P_X)$ and $\{g_2(Y), \ldots, g_{k+1}(Y)\} \subseteq \mathcal{L}^2(\mathcal{Y}, P_Y)$. These feature functions carry orthogonal modes of information, and each $f_i(X)$ is maximally correlated with $g_i(Y)$ subject to being orthogonal to all previous maximal correlation functions. Furthermore, these functions yield the following *embeddings* of $\mathcal{X}$ and $\mathcal{Y}$ into the Euclidean space $\mathbb{R}^k$:

$$\mathcal{X} \ni x \;\; \mapsto \;\; [f_2(x) \cdots f_{k+1}(x)]^T \in \mathbb{R}^k \,, \tag{4.67}$$

$$\mathcal{Y} \ni y \;\; \mapsto \;\; [g_2(y) \cdots g_{k+1}(y)]^T \in \mathbb{R}^k \,, \tag{4.68}$$

which permit us to cluster the elements of $\mathcal{X}$ and $\mathcal{Y}$ by clustering the corresponding embedded points in $\mathbb{R}^k$. Therefore, both the local information geometric perspective in subsection 4.3.3 and the categorical data embedding perspective here identify maximal correlation functions as the feature functions that are particularly suitable for decomposing information and for future use in unspecified inference tasks.

In practical settings, we rarely have knowledge of the true distribution $P_{X,Y}$. Instead, we usually have access to $n \in \mathbb{N}$ samples of training data $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in \{1, \ldots, n\}\}$ that (we assume) are drawn i.i.d. from the unknown distribution $P_{X,Y}$. For instance, in the "Netflix problem," each sample (X_i, Y_i) conveys that subscriber X_i has streamed movie Y_i. To solve the *unsupervised learning* problem of finding k real-valued features of X and Y that summarize the salient dependencies between X and Y for an unspecified inference task, we need to develop an algorithm that efficiently estimates the first few dominant maximal correlation functions in the modal decomposition of $P_{X,Y}$ from training data. Fortunately, since maximal correlation functions are singular vectors of a conditional expectation operator, we can draw on existing techniques from the numerical linear algebra literature.

■ 4.4.1 Orthogonal Iteration Method

We assume for the time being that $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ is known, and hence, its corresponding DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ and CDM $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ are also known. Due to the equivalence between the DTM and the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ shown in subsection 4.2.1, we consider the problem of computing dominant singular vectors of $\tilde{B}$, which correspond to maximal correlation functions via (4.13) and (4.14). One of the earliest and most well-known algorithms for computing the principal singular value and its corresponding singular vectors is the *power iteration method* from numerical linear algebra, which simply repeatedly multiplies $\tilde{B}^T\tilde{B}$ to an arbitrary initial vector $\psi \in \mathbb{R}^{|\mathcal{X}|}$ (with intermediate re-normalization steps), cf. [106, Section 7.3.1], [66, Section 4.4.1]. If ψ has a component in the direction of the principal right singular vector $\psi_2 \in \mathbb{R}^{|\mathcal{X}|}$ of $\tilde{B}$, and the principal singular value σ_2 of $\tilde{B}$ is well-separated from the second largest singular value σ_3, then it can be shown that the power iteration method converges exponentially fast with rate σ_3^2/σ_2^2 to ψ_2. Furthermore, the leading $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$ singular vectors of $\tilde{B}$ can be computed sequentially by repeatedly running the power iteration method with initial vectors that are orthogonal to all previously computed leading singular vectors.

However, it is preferable to compute the leading k singular vectors of $\tilde{B}$ in parallel. The most basic algorithm that achieves this is the *orthogonal iteration method*, cf. [106, Section 7.3.2], [66, Section 4.4.3], which is presented as Algorithm 1. We note that the termination condition of Algorithm 1 is derived from the variational characterization in Proposition 4.2, which portrays that $\mathrm{tr}\big((\hat{U}_k^{(i)})^T\tilde{B}\hat{V}_k^{(i)}\big)$ achieves its maximum possible value of $\|\tilde{B}\|_{(1,k)}$ when $\hat{V}_k^{(i)}$ and $\hat{U}_k^{(i)}$ have columns equal to the leading k right and left singular vectors of $\tilde{B}$, respectively. Moreover, the *thin QR decomposition* steps in the algorithm can be computed using the classical *Gram-Schmidt process* or using more sophisticated techniques like *Householder transformations* or *Givens rotations* [106, Section 5.2].

The convergence properties of Algorithm 1 are well-established in the literature, cf. [106, Section 7.3.2], [66, Section 4.4.3]. For example, if the singular values of $\tilde{B}$ are well-separated, and the initialization matrix $V_k^{(1)} \in \mathbb{R}^{|\mathcal{X}|\times k}$ has columns $v_2, \ldots, v_{k+1} \in \mathbb{R}^{|\mathcal{X}|}$ that satisfy the condition:

$$\exists \text{ distinct } j_1, \ldots, j_k \in \{2, \ldots, k+1\},\ \forall i \in \{1, \ldots, k\},\ \ \psi_{i+1}^T v_{j_i} \neq 0\,, \tag{4.69}$$

then as $i \to \infty$, $V_k^{(i)}$ and $U_k^{(i)}$ converge exponentially fast with rate $\sigma_{k+2}^2/\sigma_{k+1}^2$ to the true orthonormal k-frames of right and left singular vectors $[\psi_2 \cdots \psi_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ and $[\phi_2 \cdots \phi_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|})$, respectively, up to permutations of the columns, and $\mathrm{tr}\big((\hat{U}_k^{(i)})^T\tilde{B}\hat{V}_k^{(i)}\big)$ converges to the true Ky Fan k-norm $\|\tilde{B}\|_{(1,k)}$ (see Proposition 4.2). Appropriate generalizations of these convergence results hold when the singular values of $\tilde{B}$ are not distinct or the condition in (4.69) is not satisfied, but we omit an exposition of such generalizations for brevity. Although we only present the orthogonal iteration method for SVD computation in this chapter, we remark that there are several other

Require: CDM $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$, number of modes $k \in \{1, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$.

1. Initialization: Randomly choose $V_k^{(1)} \in \mathbb{R}^{|\mathcal{X}|\times k}$, and set iteration index $i = 0$.

repeat

2. Increment the iteration index: $i \leftarrow i + 1$.
3. Orthonormalize $V_k^{(i)} \in \mathbb{R}^{|\mathcal{X}|\times k}$ using the thin QR decomposition to obtain $\hat{V}_k^{(i)} \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$:
$$V_k^{(i)} = \hat{V}_k^{(i)} R_1^{(i)}$$
where $R_1^{(i)} \in \mathbb{R}^{k\times k}$ is an upper triangular matrix.
4. Compute the update $U_k^{(i)} \in \mathbb{R}^{|\mathcal{Y}|\times k}$:
$$U_k^{(i)} = \tilde{B}\hat{V}_k^{(i)}\,.$$
5. Orthonormalize $U_k^{(i)}$ using the thin QR decomposition to obtain $\hat{U}_k^{(i)} \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{Y}|})$:
$$U_k^{(i)} = \hat{U}_k^{(i)} R_2^{(i)}$$
where $R_2^{(i)} \in \mathbb{R}^{k\times k}$ is an upper triangular matrix.
6. Compute the update $V_k^{(i+1)} \in \mathbb{R}^{|\mathcal{X}|\times k}$:
$$V_k^{(i+1)} = \tilde{B}^T\hat{U}_k^{(i)}\,.$$

until $\mathrm{tr}\left(\left(\hat{U}_k^{(i)}\right)^T \tilde{B}\,\hat{V}_k^{(i)}\right)$ stops increasing.

Algorithm 1. Orthogonal Iteration Method.

algorithms in the numerical linear algebra literature that compute SVDs with better numerical stability and faster convergence rate, e.g. the *QR iteration* algorithm and its numerically enhanced variants, *Krylov subspace* based methods such as the *Lanczos algorithm*, etc. We refer readers to [66,106] for further details regarding such algorithms.

■ 4.4.2 Extended Alternating Conditional Expectations Algorithm

While the orthogonal iteration method for computing singular vectors of $\tilde{B}$ from subsection 4.4.1 is well-known, we now present an equivalent statistical version of this algorithm that directly computes maximal correlation functions (or singular vectors of C). In particular, the equivalence between the DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ and the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ shown in section 4.2 allows us to transform Algorithm 1 into the equivalent Algorithm 2. We note that steps 6 and 10 in Algorithm 2 correspond to steps 4 and 6 in Algorithm 1, respectively, where we use the fact that $B^T \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ is equivalent to the adjoint operator $C^* : \mathcal{L}^2(\mathcal{Y}, P_Y) \to \mathcal{L}^2(\mathcal{X}, P_X)$ of C (in a sense similar to (4.9) mutatis mutandis), and C^* maps any function $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ to the function $C^*(g) \in \mathcal{L}^2(\mathcal{X}, P_X)$ given by (cf. appendix C.3):

$$\forall x \in \mathcal{X},\ \ (C^*(g))(x) = \mathbb{E}[g(Y)|X = x]\,. \tag{4.70}$$

Require: Joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, number of modes $k \in \{1, \dots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$.

1. Initialization: Randomly choose $\underline{r}_k^{(1)} : \mathcal{X} \to \mathbb{R}^k$, and set iteration index $i = 0$.

repeat

2. Increment the iteration index: $i \leftarrow i + 1$.
3. Center the function $\underline{r}_k^{(i)} : \mathcal{X} \to \mathbb{R}^k$ to obtain the function $\underline{r}_{k,0}^{(i)} : \mathcal{X} \to \mathbb{R}^k$:
$$\forall x \in \mathcal{X},\ \ \underline{r}_{k,0}^{(i)}(x) = \underline{r}_k^{(i)}(x) - \mathbb{E}\Big[\underline{r}_k^{(i)}(X)\Big] .$$
4. Compute the Cholesky decomposition of the covariance matrix of $\underline{r}_{k,0}^{(i)}(X)$ to obtain the upper triangular matrix $R_1^{(i)} \in \mathbb{R}^{k\times k}$:
$$\mathbb{E}\Big[\underline{r}_{k,0}^{(i)}(X)\, \underline{r}_{k,0}^{(i)}(X)^T\Big] = \Big(R_1^{(i)}\Big)^T R_1^{(i)} .$$
5. Whiten $\underline{r}_{k,0}^{(i)}(X)$ using $R_1^{(i)}$ to obtain the function $\hat{r}_k^{(i)} : \mathcal{X} \to \mathbb{R}^k$:
$$\forall x \in \mathcal{X},\ \ \hat{r}_k^{(i)}(x) = \Big(R_1^{(i)}\Big)^{-T} \underline{r}_{k,0}^{(i)}(x) .$$
6. Compute the updated function $\underline{s}_k^{(i)} : \mathcal{Y} \to \mathbb{R}^k$:
$$\forall y \in \mathcal{Y},\ \ \underline{s}_k^{(i)}(y) = \mathbb{E}\Big[\hat{r}_k^{(i)}(X)\Big|Y = y\Big] .$$
7. Center the function $\underline{s}_k^{(i)}$ to obtain the function $\underline{s}_{k,0}^{(i)} : \mathcal{Y} \to \mathbb{R}^k$:
$$\forall y \in \mathcal{Y},\ \ \underline{s}_{k,0}^{(i)}(y) = \underline{s}_k^{(i)}(y) - \mathbb{E}\Big[\underline{s}_k^{(i)}(Y)\Big] .$$
8. Compute the Cholesky decomposition of the covariance matrix of $\underline{s}_{k,0}^{(i)}(Y)$ to obtain the upper triangular matrix $R_2^{(i)} \in \mathbb{R}^{k\times k}$:
$$\mathbb{E}\Big[\underline{s}_{k,0}^{(i)}(Y)\, \underline{s}_{k,0}^{(i)}(Y)^T\Big] = \Big(R_2^{(i)}\Big)^T R_2^{(i)} .$$
9. Whiten $\underline{s}_{k,0}^{(i)}(Y)$ using $R_2^{(i)}$ to obtain the function $\hat{s}_k^{(i)} : \mathcal{Y} \to \mathbb{R}^k$:
$$\forall y \in \mathcal{Y},\ \ \hat{s}_k^{(i)}(y) = \Big(R_2^{(i)}\Big)^{-T} \underline{s}_{k,0}^{(i)}(y) .$$
10. Compute the updated function $\underline{r}_k^{(i+1)} : \mathcal{X} \to \mathbb{R}^k$:
$$\forall x \in \mathcal{X},\ \ \underline{r}_k^{(i+1)}(x) = \mathbb{E}\Big[\hat{s}_k^{(i)}(Y)\Big|X = x\Big] .$$

until $\mathbb{E}\Big[\hat{r}_k^{(i)}(X)^T \hat{s}_k^{(i)}(Y)\Big]$ stops increasing.

Algorithm 2. Extended ACE Algorithm.

Steps 4 and 5 in Algorithm 2 correspond to step 3 in Algorithm 1. To verify this, consider a matrix $V_k = [v_2 \cdots v_{k+1}] \in \mathbb{R}^{|\mathcal{X}|\times k}$ (similar to $V_k^{(i)}$ in Algorithm 1) where $v_2, \dots, v_{k+1} \in \mathbb{R}^{|\mathcal{X}|}$, and a corresponding vector-valued function $\underline{r}_k : \mathcal{X} \to \mathbb{R}^k$, $\underline{r}_k(x) = [r_2(x) \cdots r_{k+1}(x)]^T$ (similar to $\underline{r}_k^{(i)}$ in Algorithm 2) where $r_2, \dots, r_{k+1} \in \mathcal{L}^2(\mathcal{X}, P_X)$,

such that (4.30) holds. Then, if V_k has thin QR decomposition $V_k = \hat{V}_k R_1$, where $\hat{V}_k \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ and $R_1 \in \mathbb{R}^{k \times k}$ is upper triangular, we have using (4.30) that:

$$\mathbb{E}\Big[\underline{r}_k(X)\,\underline{r}_k(X)^T\Big] = V_k^T V_k = R_1^T \hat{V}_k^T \hat{V}_k R_1 = R_1^T R_1 \tag{4.71}$$

where R_1 is the Cholesky factor in the *Cholesky decomposition* of the covariance matrix of $\underline{r}_k(X)$ [106, Theorem 4.2.5], [129, Corollary 7.2.9]. This establishes the desired correspondence. Similarly, steps 8 and 9 in Algorithm 2 correspond to step 5 in Algorithm 1.[58] The additional centering steps 3 and 7 in Algorithm 2 are required because steps 6 and 10 in Algorithm 2 perform updates using the operators C and C^*, which correspond to B and B^T, respectively, rather than the matrices $\tilde{B}$ and $\tilde{B}^T$ used in steps 4 and 6 in Algorithm 1.[59] Finally, we note that the termination condition of Algorithm 2 is also derived from the variational characterization in Proposition 4.2.

Due to the equivalence between Algorithms 1 and 2, the convergence properties of Algorithm 2 trivially follow from the convergence properties of Algorithm 1. For example, as before, if the singular values of C are well-separated, and the initialization function $\underline{r}_k^{(1)} : \mathcal{X} \to \mathbb{R}^k$ has coordinate functions $r_2^{(1)}, \ldots, r_{k+1}^{(1)} \in \mathcal{L}^2(\mathcal{X}, P_X)$ such that $\underline{r}_k^{(1)}(x) = [r_2^{(1)}(x) \cdots r_{k+1}^{(1)}(x)]^T$ for all $x \in \mathcal{X}$ and the coordinate functions satisfy a condition akin to (4.69):

$$\exists \text{ distinct } j_1, \ldots, j_k \in \{2, \ldots, k+1\},\ \forall i \in \{1, \ldots, k\},\ \ \mathbb{E}\Big[f_{i+1}(X)\, r_{j_i}^{(1)}(X)\Big] \neq 0\,, \tag{4.72}$$

then as $i \to \infty$, $\underline{r}_k^{(i)}$ and $\underline{s}_k^{(i)}$ converge exponentially fast to the k leading maximal correlation functions (stacked into vectors) $\mathcal{X} \ni x \mapsto [f_2(x) \cdots f_{k+1}(x)]^T$ and $\mathcal{Y} \ni y \mapsto [g_2(y) \cdots g_{k+1}(y)]^T$, respectively, up to permutations of the entries.[60]

In the case $k = 1$, the simplified Algorithm 2 is known as the *alternating conditional expectations (ACE) algorithm* in the *non-parametric regression* literature [35]. So, we briefly elucidate the connection between maximal correlation and regression. In [35], Breiman and Friedman study a generalization of the following idealized regression problem:

$$\min_{\substack{f \in \mathcal{L}^2(\mathcal{X}, P_X),\, g \in \mathcal{L}^2(\mathcal{Y}, P_Y): \\ \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0, \\ \mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1}} \mathbb{E}\Big[(f(X) - g(Y))^2\Big] \tag{4.73}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are general sets and $P_{X,Y}$ is a general joint probability measure such that $\mathcal{L}^2(\mathcal{X}, P_X)$ and $\mathcal{L}^2(\mathcal{Y}, P_Y)$ are separable Hilbert spaces, and we assume that the

[58] We remark that the Cholesky decomposition and whitening steps of Algorithm 2 can be executed using more efficient approaches if desired.

[59] Technically, we only need to center $\underline{r}_k^{(1)}$ once at the beginning of Algorithm 2, and $\underline{r}_k^{(i)}$ should remain zero mean over iterations. However, repeatedly centering in Algorithm 2 makes it more stable (particularly when empirical expectations are used instead of true ones).

[60] We remark that the Cholesky decompositions in Algorithm 2 are unique, and the inverses in steps 5 and 9 exist, when the aforementioned conditions for convergence are satisfied.

minimum exists. Since $\mathbb{E}[(f(X)-g(Y))^2] = \mathbb{E}[f(X)^2] - 2\mathbb{E}[f(X)g(Y)] + \mathbb{E}[g(Y)^2] = 2 - 2\mathbb{E}[f(X)g(Y)]$ for any normalized functions $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$, the minimizing functions of (4.73) are precisely the maximal correlation functions that extremize (4.1). Hence, the non-parametric regression problem in (4.73) is equivalent to the maximal correlation problem in (4.1), and both can be solved using the ACE algorithm. Furthermore, the SVD computation view of the ACE algorithm is already (tacitly) present in [35]. So, from a statistical perspective, our main contribution in this section is the extension of the standard ACE algorithm in Algorithm 2 so that it computes multiple maximal correlation functions in parallel. This is why we refer to Algorithm 2 as the *extended ACE algorithm*.

We next return to the setting where the true distribution $P_{X,Y}$ is unknown, but we are given i.i.d. training data $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in \{1, \ldots, n\}\}$ from $P_{X,Y}$. We cannot directly use the extended ACE algorithm to estimate $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$ leading pairs of maximal correlation functions from this data, because Algorithm 2 requires knowledge of $P_{X,Y}$. However, a natural modification is to use the empirical joint distribution $\hat{P}_{X_1^n, Y_1^n}$ of the data in Algorithm 2 in place of $P_{X,Y}$, where the empirical joint distribution is defined as:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; \hat{P}_{X_1^n, Y_1^n}(x,y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = x, Y_i = y\}\,. \tag{4.74}$$

We refer to this modified version of Algorithm 2 as the *sample extended ACE algorithm*, where we replace the expectations in steps 3, 4, 7, and 8, and the termination condition of Algorithm 2 with empirical expectations, and the conditional expectations in steps 6 and 10 with empirical conditional expectations. For example, step 6 is modified to:

$$\forall y \in \mathcal{Y}, \;\; \underline{s}_k^{(i)}(y) = \frac{1}{n \hat{P}_{Y_1^n}(y)} \sum_{j=1}^{n} \hat{r}_k^{(i)}(X_j) \mathbb{1}\{Y_j = y\}\,. \tag{4.75}$$

In practice, the different empirical expectations can be computed with (possibly overlapping) subsets of the n samples to improve computational complexity. Furthermore, Algorithm 1 can also be modified to use training data by replacing the true CDM in steps 4 and 6 and the termination condition with the CDM corresponding to the empirical joint distribution $\hat{P}_{X_1^n, Y_1^n}$, and we refer to the resulting algorithm as the *sample orthogonal iteration method*.

On the computational front, well-known results from numerical linear algebra (discussed earlier) suggest that under mild regularity conditions, the sample extended ACE algorithm and sample orthogonal iteration method are both guaranteed to converge exponentially fast to estimates of the true k leading pairs of maximal correlation functions or singular vectors of $\tilde{B}$. Hence, these algorithms are known to be computationally efficient, and we do not delve into their computational complexity in this thesis.

On the statistical front, we will perform *sample complexity analysis* of the sample orthogonal iteration method in section 4.6. This analysis will also hold mutatis mutandis

for the sample extended ACE algorithm due to the equivalence between Algorithms 1 and 2. Specifically, we will focus on the high-dimensional regime where the sample size n is large enough to accurately estimate the true marginals $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, but not large enough to accurately estimate the true joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. So, in our analysis, we will assume that P_X and P_Y are known, but $P_{X,Y}$ is not known. This assumption is reasonable even if the sample size n is not large enough to accurately estimate the marginals P_X and P_Y, because additional *unlabeled* training samples from P_X and P_Y are often very cheaply available, and can be exploited to estimate P_X and P_Y precisely.[61] It is straightforward to appropriately modify the sample orthogonal iteration method to incorporate this information by defining the "empirical CDM" $\hat{B}_n \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ entry-wise as:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \quad \left[\hat{B}_n\right]_{y,x} = \frac{\hat{P}_{X_1^n,Y_1^n}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}} \tag{4.76}$$

where we use both the empirical joint distribution $\hat{P}_{X_1^n,Y_1^n}$ of the bivariate training data and the knowledge of P_X and P_Y. The sample extended ACE algorithm can also be modified accordingly.[62] For example, the equivalent of $\hat{B}_n$ is the operator $\hat{C}_n : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$, which maps any $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to $\hat{C}_n(f) \in \mathcal{L}^2(\mathcal{Y}, P_Y)$:

$$(\hat{C}_n(f))(y) = \frac{1}{nP_Y(y)} \sum_{i=1}^{n} f(X_i)\mathbb{1}\{Y_i = y\} \; - \; \mathbb{E}[f(X)]\,. \tag{4.77}$$

Most of our sample complexity analysis in section 4.6 pertains to the sample orthogonal iteration method corresponding to the "empirical CDM" in (4.76). In particular, Propositions 4.5 and 4.6 and Theorems 4.5 and 4.6 will illustrate that the sample versions of Algorithms 1 and 2 are *consistent* under appropriate scaling conditions of k relative to n, i.e. these algorithms produce estimates of leading singular vectors and Ky Fan k-norms of $\tilde{B}$ that "converge" to the true quantities in probability and with respect to appropriate loss functions as $n \to \infty$.

We close this section with some pertinent remarks about the sample version of Breiman and Friedman's ACE algorithm (or the sample extended ACE algorithm with $k = 1$) for real-valued data and its relation to maximal correlation. To this end, recall that according to Definition 2.3 in chapter 2 or (4.1), maximal correlation is clearly well-defined for any bivariate distribution on $\mathbb{R}^2$, i.e. "in the population." However, it turns out to be quite subtle to define "in the sample" when only real-valued training data is available [272, Lectures 11 and 12].

To understand this, consider the traditional notion of Pearson correlation coefficient. In the population setting, the Pearson correlation coefficient of two jointly distributed random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ is given by $\mathbb{E}[XY]$. Likewise, in the

[61] Although we consider our learning setting to be unsupervised because the latent variable U is unspecified, we can construe each sample $(X_i, Y_i) \sim P_{X,Y}$ as *labeled* data.

[62] Since we assume that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ are known and strictly positive entry-wise, the "empirical CDM" in (4.76) and the modified conditional expectations, e.g. (4.77), in the sample extended ACE algorithm are well-defined.

sample setting, the correlation coefficient for centered and normalized bivariate data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$, with $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$ and $y = [y_1 \cdots y_n]^T \in \mathbb{R}^n$ satisfying $x^T\mathbf{1} = y^T\mathbf{1} = 0$ and $\|x\|_2 = \|y\|_2 = 1$, is given by the projection $x^T y$. Furthermore, if the bivariate data is drawn i.i.d. from $P_{X,Y}$ and then centered and normalized, then the two notions of correlation coefficient coincide by the *strong law of large numbers* (SLLN). Therefore, Pearson correlation coefficient is both well-defined in the population and in the sample.

Since population maximal correlation is defined as a maximum of population Pearson correlation coefficients, a seemingly sound way to define sample maximal correlation is via sample correlation coefficients. So, for bivariate data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ (which is not necessarily centered or normalized) with $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$ and $y = [y_1 \cdots y_n]^T \in \mathbb{R}^n$, we define the quantity, cf. [272, Lecture 11]:

$$\tilde{\rho}_{\mathsf{max}}(x; y) = \sup_{\substack{f:\mathbb{R}\to\mathbb{R},\, g:\mathbb{R}\to\mathbb{R}:\\ f(x)^T\mathbf{1}=g(y)^T\mathbf{1}=0\\ \|f(x)\|_2=\|g(y)\|_2=1}} f(x)^T g(y) \tag{4.78}$$

where we optimize over all functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, and we apply functions on $\mathbb{R}$ to vectors in $\mathbb{R}^n$ entry-wise, i.e. $f(x) = [f(x_1) \cdots f(x_n)]^T$ and $g(y) = [g(y_1) \cdots g(y_n)]^T$. Unfortunately, $\tilde{\rho}_{\mathsf{max}}(x; y) = 1$ for every pair of vectors $x, y \in \mathbb{R}^n$, and this definition of sample maximal correlation is ineffectual. Thus, statisticians define the output of the sample version of Breiman and Friedman's ACE algorithm as the sample maximal correlation, cf. [272, Lecture 11].

However, this does not entirely circumvent the difficulty in defining maximal correlation in the sample, because the conditional expectations in steps 6 and 10 of Algorithm 2 are nontrivial to approximate using training samples when the underlying random variables are continuous. Breiman and Friedman propose the use of various *data smoothers* (e.g. histogram smoother, nearest neighbor smoother) to approximate these conditional expectations in [35]. This renders a unified analysis of the sample ACE algorithm for all data smoothers almost impossible, and separate convergence and consistency analyses are required for the sample ACE algorithm with different data smoothers. Breiman and Friedman prove some sufficient conditions on data smoothers that ensure convergence and consistency, and in particular, establish consistency for the nearest neighbor smoother [35]. We note that there is no canonical choice of data smoother, and the definition of sample maximal correlation evidently varies based on the choice of data smoother. On the other hand, in the finite alphabet setting of this chapter, we can easily approximate conditional expectations using empirical conditional distributions (which correspond to histogram smoothers), and the corresponding sample ACE algorithm can be used to canonically define maximal correlation in the sample.

■ 4.5 Comparison to Related Statistical Techniques

Until now, we have discussed the close connections between our approach of learning maximal correlation functions from training data as a means of feature extraction for

unspecified inference tasks and the following notions in the literature:

1. correspondence analysis [24, 125], which also exploits modal decompositions for data visualization,

2. the theory of Lancaster distributions [165, 166], which studies modal decompositions of bivariate distributions over general spaces for their own sake (also see subsection 2.2.2 in chapter 2),

3. the ACE algorithm for non-parametric regression [35], which we extend into Algorithm 2 in the finite alphabet setting.

A wonderful and unified exposition of these ideas can be found in [37]. In the ensuing subsections, we compare our approach to some other related techniques in the statistics literature.

■ 4.5.1 Principal Component Analysis

Principal component analysis is one of the most popular and well-known dimensionality reduction techniques in statistics.[63] It was developed by Pearson in [223], and independently by Hotelling in [130]. It is instructive to compare the sample extended ACE algorithm to principal component analysis, because the two approaches have a clear resemblance in that both utilize the SVD. In principal component analysis, the observed data are real-valued vectors $y_1, \ldots, y_n \in \mathbb{R}^m$ (with $m, n \in \mathbb{N}$). We stack these vectors together to form a matrix $\tilde{Y} = [y_1 \cdots y_n] \in \mathbb{R}^{m \times n}$, and compute the SVD of $\tilde{Y}$. Then, we can project each observed m-dimensional vector onto the subspace spanned by the $k \in \{1, \ldots, m\}$ (where typically $k < m$) leading left singular vectors of $\tilde{Y}$ to form a reduced k-dimensional representation.

In the sample extended ACE algorithm, we observe the samples $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in \{1, \ldots, n\}\}$. We then compute the vectors $\{\hat{b}_x \in \mathbb{R}^{|\mathcal{Y}|} : x \in \mathcal{X}\}$ which are shifted and scaled versions of the empirical conditional distributions of Y given $X = x$ for each $x \in \mathcal{X}$:

$$\forall y \in \mathcal{Y}, \;\; \hat{b}_x(y) = \sqrt{\frac{P_X(x)}{P_Y(y)}} \hat{P}^n_{Y|X}(y|x) - \sqrt{P_X(x) P_Y(y)} \tag{4.79}$$

where we assume that $\hat{P}_{X_1^n} = P_X$ and $\hat{P}_{Y_1^n} = P_Y$ for simplicity, and for every $x \in \mathcal{X}$, the empirical conditional distribution $\hat{P}^n_{Y|X=x} \in \mathcal{P}_{\mathcal{Y}}$ is defined as:

$$\forall y \in \mathcal{Y}, \;\; \hat{P}^n_{Y|X}(y|x) \triangleq \frac{1}{n \hat{P}_{X_1^n}(x)} \sum_{i=1}^{n} \mathbb{1}\{X_i = x, Y_i = y\} \,. \tag{4.80}$$

These vectors are stacked together to form the "empirical CDM" $\hat{B}_n = [b_1 \cdots b_{|\mathcal{X}|}] \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ in (4.76), and we compute the SVD of $\hat{B}_n$. The sample analogs of the statistics

[63] This technique is also known as the *Karhunen-Loève transform* in signal processing.

computed in (4.65) are projections of $\hat{P}_{Y_1^n}$ (properly shifted and scaled) onto the leading left singular vectors of $\hat{B}_n$.

Therefore, the two approaches are almost identical. The key difference is that in the sample extended ACE algorithm, we operate in the space of probability distributions rather than data. Consequently, "a strong advantage of the [extended] ACE procedure is the ability to incorporate variables of quite different type in terms of the set of values they can assume" [35].

■ 4.5.2 Canonical Correlation Analysis

In the optimization defining maximal correlation in (4.1), we look for general functions $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ such that $f(X)$ and $g(Y)$ are highly correlated. If we further restrict these functions to lie in linear subspaces (or sub-Hilbert spaces) of the functional spaces $\mathcal{L}^2(\mathcal{X}, P_X)$ and $\mathcal{L}^2(\mathcal{Y}, P_Y)$ in the optimization, we can still define the conditional expectation operator C as a linear map from a subspace of functions on $\mathcal{X}$ to a subspace of functions on $\mathcal{Y}$. Finding such constrained functions that are highly correlated corresponds to computing the SVD of C composed with appropriate projection operators. Thus, our entire discussion regarding the SVD structure of C and iterative algorithms to compute maximal correlation functions holds in this scenario.

A particular case of interest is when we have jointly distributed random vectors $\underline{X} \in \mathbb{R}^{m_1}$ and $\underline{Y} \in \mathbb{R}^{m_2}$ (with $m_1, m_2 \in \mathbb{N}$), where $\underline{X}$ and $\underline{Y}$ have zero mean (for simplicity), full rank covariance matrices $K_X = \mathbb{E}[\underline{X}\underline{X}^T] \in \mathbb{R}_{\succeq 0}^{m_1 \times m_1}$ and $K_Y = \mathbb{E}[\underline{Y}\underline{Y}^T] \in \mathbb{R}_{\succeq 0}^{m_2 \times m_2}$, respectively, and cross-covariance matrix $K_{X,Y} = \mathbb{E}[\underline{X}\underline{Y}^T] \in \mathbb{R}^{m_1 \times m_2}$, and we constrain the functions in the optimization defining maximal correlation to be linear functions. With a little abuse of notation, we parametrize the linear functions $f : \mathbb{R}^{m_1} \to \mathbb{R}$ and $g : \mathbb{R}^{m_2} \to \mathbb{R}$ using the vectors $\underline{f} \in \mathbb{R}^{m_1}$ and $\underline{g} \in \mathbb{R}^{m_2}$, respectively, so that $f(\underline{x}) = \underline{f}^T \underline{x}$ for all $\underline{x} \in \mathbb{R}^{m_1}$ and $g(\underline{y}) = \underline{g}^T \underline{y}$ for all $\underline{y} \in \mathbb{R}^{m_2}$. Then, we can specialize the definition of maximal correlation in (4.1) into the *canonical correlation coefficient*:

$$\max_{\substack{f:\mathbb{R}^{m_1}\to\mathbb{R},\, g:\mathbb{R}^{m_2}\to\mathbb{R}:\\ f,g \text{ linear functions}\\ \mathbb{E}[f(\underline{X})^2]=\mathbb{E}[g(\underline{Y})^2]=1}} \mathbb{E}[f(\underline{X})g(\underline{Y})] = \max_{\substack{\underline{f}\in\mathbb{R}^{m_1},\, \underline{g}\in\mathbb{R}^{m_2}:\\ \underline{f}^T K_X \underline{f} = \underline{g}^T K_Y \underline{g} = 1}} \underline{f}^T K_{X,Y} \underline{g}\,. \tag{4.81}$$

This is the setup of Hotelling's *canonical correlation analysis* (CCA) [131]. The optimizing arguments of (4.81):

$$\underline{f}^* = K_X^{-\frac{1}{2}} \underline{v} \quad \text{and} \quad \underline{g}^* = K_Y^{-\frac{1}{2}} \underline{u} \tag{4.82}$$

define the first pair of canonical variables $(\underline{f}^*)^T \underline{X}$ and $(\underline{g}^*)^T \underline{Y}$, where $\underline{v} \in \mathbb{R}^{m_1}$ and $\underline{u} \in \mathbb{R}^{m_2}$ are the left and right singular vectors, respectively, corresponding to the largest singular value of the matrix, cf. [119]:

$$\tilde{K} \triangleq K_X^{-\frac{1}{2}} K_{X,Y} K_Y^{-\frac{1}{2}}\,. \tag{4.83}$$

Furthermore, successive pairs of singular vectors of $\tilde{K} \in \mathbb{R}^{m_1 \times m_2}$ determine ensuing pairs of canonical variables. When CCA is used in practice, the covariance and cross-covariance matrices K_X, K_Y, and $K_{X,Y}$ must be estimated from data samples.

The matrix $\tilde{K}$ has a strong resemblance to the adjoint of the DTM in Definition 4.1. While we can directly compute the SVD of $\tilde{K}$ to solve the CCA problem, we note that a modified version of the sample extended ACE algorithm can also be used to compute pairs of canonical variables. Indeed, it suffices to incorporate the linear function constraints in (4.81) into Algorithm 2. Notice that the only stages of Algorithm 2 where we may get non-linear functions is after the updates in steps 6 and 10. So, we need to project the updated functions obtained from these steps onto the corresponding subspaces of linear functions. With this modification, the sample extended ACE algorithm solves the CCA problem. (Of course, for continuous random vectors $\underline{X}$ and $\underline{Y}$, we also need to employ data smoothers to approximate the conditional expectations in steps 6 and 10 of Algorithm 2 as discussed earlier.)

We conclude our discussion of CCA with some additional remarks. Firstly, since the CCA problem only requires knowledge of the first and second moments $\underline{X}$ and $\underline{Y}$, we can treat $\underline{X}$ and $\underline{Y}$ as though they are jointly Gaussian distributed (as is commonly done in *linear least squares estimation*). Secondly, a further special case of CCA is when $m_1 = m_2$ and the noise model is actually AWGN, i.e. $K_Y = K_X + \nu^2 I$ and $K_{X,Y} = K_X$, where $\nu^2 > 0$ is some noise variance. This simplifies the CCA problem as the covariance matrices defining $\tilde{K}$ commute, and are hence, jointly diagonalizable. Consequently, it is straightforward to argue that the ordering of the eigenvectors of K_Y (according to the ordering of its eigenvalues) is consistent with the ordering of the eigenvectors of $\tilde{K}$. Hence, in this special setting, the (ordered) canonical variables obtained from CCA correspond to the (ordered) principal components in principal component analysis. Lastly, since the canonical correlation coefficient in (4.81) is a constrained version of maximal correlation in (4.1), one might wonder whether maximal correlation can be perceived as a canonical correlation coefficient. To answer this question in the affirmative, consider the jointly distributed discrete random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with finite alphabets (that have been the subject of this chapter). Corresponding to these random variables, define the zero mean random vectors of shifted indicators:

$$\underline{X} = \left([\mathbb{1}\{X = 1\} \cdots \mathbb{1}\{X = |\mathcal{X}|\}] - P_X\right)^T \in \mathbb{R}^{|\mathcal{X}|}, \tag{4.84}$$

$$\underline{Y} = \left([\mathbb{1}\{Y = 1\} \cdots \mathbb{1}\{Y = |\mathcal{Y}|\}] - P_Y\right)^T \in \mathbb{R}^{|\mathcal{Y}|}. \tag{4.85}$$

Then, it is easy to verify that the canonical correlation coefficient of $\underline{X}$ and $\underline{Y}$ is given by:

$$\max_{\substack{\underline{f} \in \mathbb{R}^{|\mathcal{X}|},\, \underline{g} \in \mathbb{R}^{|\mathcal{Y}|}: \\ \mathbb{E}\left[\left(\underline{f}^T \underline{X}\right)^2\right] = \mathbb{E}\left[\left(\underline{g}^T \underline{Y}\right)^2\right] = 1}} \mathbb{E}\left[\left(\underline{f}^T \underline{X}\right)\left(\underline{g}^T \underline{Y}\right)\right] = \rho_{\mathsf{max}}(X;Y) \tag{4.86}$$

which is equal to the maximal correlation of X and Y. Therefore, maximal correlation of random variables with finite ranges can be viewed as a special case of CCA.

■ 4.5.3 Diffusion Maps

Diffusion maps were proposed in [51] (and other papers by the authors of [51] and their collaborators) as a general conceptual framework for understanding so called "kernel eigenmap methods" such as *Laplacian eigenmaps* [22]. They have been utilized in several machine learning problems such as *manifold learning* and *spectral clustering* (see e.g. [20, Section 2]). As explained in [51], "the remarkable idea [behind this approach] is that eigenvectors of Markov matrices can be thought of as coordinates on the data set. Therefore, the data ... can be represented (embedded) as a cloud of points in a Euclidean space." We briefly delineate the basic idea of diffusion maps from the perspective of the modal decompositions in this chapter.

The discussion in subsection 4.5.1 shows that we can construe the columns $\{\hat{b}_x \in \mathbb{R}^{|\mathcal{Y}|} : x \in \mathcal{X}\}$ of the "empirical CDM" $\hat{B}_n$ as representations of the elements of $\mathcal{X}$ as data points in $\mathbb{R}^{|\mathcal{Y}|}$, and moreover, performing principal component analysis on these points corresponds to our modal decomposition approach using the sample extended ACE algorithm. For simplicity, let us proceed with the assumption that $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ is known instead of working with data samples. Rather than associating each $x \in \mathcal{X}$ with a column of the CDM, let us (equivalently) associate it with the conditional distribution $P_{Y|X=x} \in \mathcal{P}_{\mathcal{Y}}$. This association also embeds each $x \in \mathcal{X}$ into $\mathbb{R}^{|\mathcal{Y}|-1}$ (after transposing $P_{Y|X=x}$ and exploiting the fact that it sums to unity). In order to reduce the dimension $|\mathcal{Y}|-1$ of this embedding, we recast the modal decomposition in part 1 of Theorem 4.3 as:

$$\forall x \in \mathcal{X},\ \ P_{Y|X=x} = P_Y + \sum_{i=2}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \sigma_i\, f_i(x)\, g_i^T \mathsf{diag}(P_Y) \tag{4.87}$$

where we abuse notation and treat $g_i \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ as a vector $g_i = [g_i(1) \cdots g_i(|\mathcal{Y}|)]^T \in \mathbb{R}^{|\mathcal{Y}|}$ for every $i \in \{2, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$. Evidently, each conditional distribution $P_{Y|X=x}$, and hence, each element x, can be equivalently represented using the column vector:

$$\mathcal{X} \ni x \ \ \mapsto \ \ \Big[\sigma_2 f_2(x) \cdots \sigma_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} f_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}(x)\Big]^T \in \mathbb{R}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}-1}\,. \tag{4.88}$$

It is straightforward to verify that the standard Euclidean ℓ^2-distance between representations of the form (4.88) for any two elements $x, x' \in \mathcal{X}$ precisely captures a χ^2-like distance, known as the squared *diffusion distance* [51], between the corresponding conditional distributions:

$$D_{\mathsf{diff}}(P_{Y|X=x}, P_{Y|X=x'}) \triangleq \sum_{y\in\mathcal{Y}} \frac{\big(P_{Y|X}(y|x) - P_{Y|X}(y|x')\big)^2}{P_Y(y)} \tag{4.89}$$

$$= \sum_{i=2}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}} \sigma_i^2 \,\big(f_i(x) - f_i(x')\big)^2 \tag{4.90}$$

where $D_{\mathsf{diff}}(\cdot,\cdot)$ is parametrized by P_Y. (Thus, clustering embeddings of $\mathcal{X}$ given by (4.88) using ℓ^2-distance corresponds to clustering embeddings of $\mathcal{X}$ given by conditional

distributions using diffusion distance.) Furthermore, we can truncate the isometric embedding in (4.88) and only retain the leading $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$ entries since the remaining singular values are often very small, where typically $k \ll \min\{|\mathcal{X}|, |\mathcal{Y}|\}$.[64] This yields the following lower dimensional embedding of $\mathcal{X}$ into $\mathbb{R}^k$:

$$\mathcal{X} \ni x \;\; \mapsto \;\; [\sigma_2 f_2(x) \cdots \sigma_{k+1} f_{k+1}(x)]^T \in \mathbb{R}^k \,, \tag{4.91}$$

which we note is very closely related to our proposed embedding in (4.67) (which does not have the scaling by singular values). It is also straightforward to see using (4.90) that when $\sigma_{k+2}, \ldots, \sigma_{\min\{|\mathcal{X}|,|\mathcal{Y}|\}}$ are very small, the standard Euclidean ℓ^2-distance between two representations of the form (4.91) is a "good" approximation of the diffusion distance between the corresponding conditional distributions. (So, clustering embeddings of $\mathcal{X}$ given by (4.91) using ℓ^2-distance is roughly equivalent to clustering embeddings of $\mathcal{X}$ given by conditional distributions using diffusion distance.)

The diffusion map is an embedding analogous to that in (4.91). Suppose we are given a weighted undirected graph with vertex set $\mathcal{X}$, and we seek to embed the vertices of this graph into $\mathbb{R}^k$. As expounded in [20, Section 2], one way to do this is to consider the Markov transition probability matrix $W = P_{Y|X} \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ corresponding to the *random walk* on this graph, where we let $\mathcal{Y} = \mathcal{X}$ and we construe Y as a one-step transition of X for this Markov chain. The t-step Markov matrix W^t for some $t \in \mathbb{N} \cup \{0\}$ and its invariant distribution $P_X = P_Y \in \mathcal{P}_{\mathcal{X}}$ define a joint pmf on $\mathcal{X} \times \mathcal{X}$. Following our earlier discussion, we immediately obtain the embedding in (4.91), which represents vertices of our graph as vectors in $\mathbb{R}^k$ with $k \in \{1, \ldots, |\mathcal{X}| - 1\}$. However, since the Markov chain defined by W is reversible, cf. [170, Section 9.1], the corresponding conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ is self-adjoint and its singular vectors and eigenvectors coincide. Hence, we obtain the following version of the modal decomposition in (4.87):

$$\forall x \in \mathcal{X}, \;\; W_x^t = P_X + \sum_{i=2}^{|\mathcal{X}|} \lambda_i^t \, \tilde{f}_i(x) \, \tilde{f}_i^T \mathsf{diag}(P_X) \tag{4.92}$$

where $W_x^t \in \mathcal{P}_{\mathcal{X}}$ denotes the xth row of W^t, $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{|\mathcal{X}|} \geq -1$ are the ordered eigenvalues of C (whose absolute values yield the singular values $\sigma_1, \ldots, \sigma_{|\mathcal{X}|}$), and $\tilde{f}_1 = \mathbf{1}, \tilde{f}_2, \ldots, \tilde{f}_{|\mathcal{X}|} \in \mathbb{R}^{|\mathcal{X}|}$ are the corresponding eigenvectors such that $\tilde{f}_1 = f_1$ and $\tilde{f}_2, \ldots, \tilde{f}_{|\mathcal{X}|}$ are a reordering of the maximal correlation functions $f_2, \ldots, f_{|\mathcal{X}|} \in \mathcal{L}^2(\mathcal{X}, P_X)$ construed as vectors in $\mathbb{R}^{|\mathcal{X}|}$. This decomposition gives rise to the diffusion map from $\mathcal{X}$ to $\mathbb{R}^k$:

$$\mathcal{X} \ni x \;\; \mapsto \;\; \left[\lambda_2^t \tilde{f}_2(x) \cdots \lambda_{k+1}^t \tilde{f}_{k+1}(x)\right]^T \in \mathbb{R}^k \tag{4.93}$$

which is parametrized by $t \in \mathbb{N} \cup \{0\}$, and orders the maximal correlation functions using eigenvalues of C rather than singular values. Much like (4.90), for any $x, x' \in \mathcal{X}$,

[64] In practice, appropriate values of k are determined using techniques like identifying "elbows" in *scree plots*, cf. [20, Section 1.1.4].

the diffusion distance between the rows $W_x^t, W_{x'}^t \in \mathcal{P}_{\mathcal{X}}$ (as defined in (4.89)) is given by the standard Euclidean ℓ^2-distance between the representations of x, x' produced by (4.93) when $k = |\mathcal{X}| - 1$. Moreover, the diffusion map parallels the embedding in (4.91) when $t = 1$, and the embedding in (4.67) when $t = 0$.

These associations illustrate the relationship between our approach towards feature extraction and diffusion maps. We close this section by remarking that there are further parallels between the theory in this chapter and aspects of *spectral graph theory*. For example, the symmetric normalized *Laplacian matrix* of a graph has precisely the same structure as the DTM matrix in Definition 4.1—see e.g. [233, Section II-D], [246, Section 2.2].

■ 4.6 Sample Complexity Analysis

We mainly analyze various aspects of the sample orthogonal iteration method's sample complexity in this section. As mentioned in subsection 4.4.2, our results also hold for the equivalent sample extended ACE algorithm. Recall that we are given $n \in \mathbb{N}$ samples of training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are drawn i.i.d. from a fixed but unknown joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. The sample orthogonal iteration method computes estimates of the leading $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$ (typically with $k \ll \min\{|\mathcal{X}|, |\mathcal{Y}|\}$) singular vectors of the CDM $\tilde{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$. In the first two subsections of this section, we will operate in the high-dimensional regime where $|\mathcal{X}| = o(n)$, $|\mathcal{Y}| = o(n)$, and $|\mathcal{X}||\mathcal{Y}| = \omega(n)$. For example, we might have $|\mathcal{X}| = \Theta(n^{2/3})$, $|\mathcal{Y}| = \Theta(n^{2/3})$, and $|\mathcal{X}||\mathcal{Y}| = \Theta(n^{4/3})$. Hence, we will assume that the marginal pmfs $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ are known since they can be estimated accurately from training data, but the joint pmf $P_{X,Y}$ is still unknown because it cannot be consistently estimated from data. This assumption can be intuitively justified by the following theorem which presents known results that characterize the minimax rates of estimating discrete distributions in TV distance, cf. [118, 148].

Theorem 4.4 (Minimax Estimation of Discrete Distributions). *Consider the probability simplex $\mathcal{P}_{\mathcal{X}}$ on a finite alphabet $\mathcal{X}$ with $2 \le |\mathcal{X}| < \infty$.*

1. (Classical regime [148, Section 4]) *If $|\mathcal{X}| = O(1)$ is fixed, and X_1^n are i.i.d. samples from some unknown probability distribution in $\mathcal{P}_{\mathcal{X}}$, then we have:*

$$\sqrt{\frac{|\mathcal{X}| - 1}{2\pi n}} + o\Big(\frac{1}{\sqrt{n}}\Big) \le \inf_{Q_X^n(\cdot)} \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} \mathbb{E}_{P_X}[\|Q_X^n(X_1^n) - P_X\|_{\mathsf{TV}}]$$
$$\le \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} \mathbb{E}_{P_X}\Big[\Big\|\hat{P}_{X_1^n} - P_X\Big\|_{\mathsf{TV}}\Big] \le \sqrt{\frac{|\mathcal{X}| - 1}{2\pi n}} + o\Big(\frac{1}{\sqrt{n}}\Big)$$

where the infimum is over all estimators $Q_X^n : \mathcal{X}^n \to \mathcal{P}_{\mathcal{X}}$ of P_X based on X_1^n, the suprema are over all pmfs in $\mathcal{P}_{\mathcal{X}}$, and the expectations are with respect to the true product distribution of X_1^n.

2. (Critical high-dimensional regime [118, Section I-A]) *If $|\mathcal{X}| = n/\alpha$ for some constant $\alpha > 0$, and X_1^n are i.i.d. samples from some unknown probability distribution in $\mathcal{P}_{\mathcal{X}}$, then we have:*

$$\Omega\left(\frac{1}{\sqrt{\alpha}}\right) = \liminf_{n\to\infty} \inf_{Q_X^n(\cdot)} \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} \mathbb{E}_{P_X}[\|Q_X^n(X_1^n) - P_X\|_{\mathsf{TV}}]$$
$$\leq \limsup_{n\to\infty} \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} \mathbb{E}_{P_X}\left[\left\|\hat{P}_{X_1^n} - P_X\right\|_{\mathsf{TV}}\right] = O\left(\frac{1}{\sqrt{\alpha}}\right).$$

Furthermore, if $|\mathcal{X}| = \omega(n)$, then no estimator $Q_X^n : \mathcal{X}^n \to \mathcal{P}_{\mathcal{X}}$ for P_X based on X_1^n is consistent under the TV distance loss, which shows that $|\mathcal{X}| = n/\alpha$ is the critical scaling of $|\mathcal{X}|$ with respect to n.

Part 2 of Theorem 4.4 conveys that in our aforementioned high-dimensional regime of $|\mathcal{X}|$, $|\mathcal{Y}|$, and n, the joint pmf $P_{X,Y}$ with underlying alphabet size $|\mathcal{X}||\mathcal{Y}|$ cannot be estimated consistently under the TV distance loss, but the marginal pmfs P_X and P_Y with alphabet sizes $|\mathcal{X}|$ and $|\mathcal{Y}|$, respectively, can be consistently estimated. In the next two subsections, we will let $\delta > 0$ be a universal constant such that the marginal pmfs satisfy:

$$\forall x \in \mathcal{X}, \;\; P_X(x) \geq \delta\,, \tag{4.94}$$
$$\forall y \in \mathcal{Y}, \;\; P_Y(y) \geq \delta\,, \tag{4.95}$$

where e.g. we may define δ to be the minimum probability mass among all $P_X(x)$ and $P_Y(y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. (Note that when we prove minimax upper bounds in the sequel, we will assume that the marginals P_X and P_Y are fixed although the joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ can vary. Hence, δ is "universal" in the sense that it does not vary with our choice of $P_{X,Y}$.) Since P_X and P_Y are known, we will define the "empirical CDM" $\hat{B}_n \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ corresponding to the training data, which is used in the sample orthogonal iteration method, according to (4.76). We remark that the known marginal pmfs P_X and P_Y are not necessarily equal to the empirical marginal distributions $\hat{P}_{X_1^n}$ and $\hat{P}_{Y_1^n}$, because they may have been estimated using additional unlabeled training samples and possibly using estimators other than the empirical distribution.

Since we cannot accurately estimate the large dimensional distribution $P_{X,Y}$ due to constraints on the sample size n, we can perceive the sample orthogonal iteration method (or the sample extended ACE algorithm) as an effort to circumvent this impediment. Indeed, this algorithm only estimates parts of the bivariate distribution $P_{X,Y}$ with the hope that this partial knowledge is useful for the purposes of future inference tasks. So, we intuitively expect the sample orthogonal iteration method to require fewer training samples than algorithms that attempt a full estimation of $P_{X,Y}$.

■ 4.6.1 Estimation of Ky Fan k-Norms of CDMs

Since the sample version of Algorithm 1 terminates when the quantity in its termination condition converges to $\|\hat{B}_n\|_{(1,k)}$, we can think of the sample orthogonal iteration

method as a means of computing the "plug-in" estimator $\|\hat{B}_n\|_{(1,k)}$ of the true Ky Fan k-norm $\|\tilde{B}\|_{(1,k)}$ of the CDM $\tilde{B}$. So, we begin by providing a minimax upper bound on the problem of estimating $\|\tilde{B}\|_{(1,k)}$ by analyzing the sample complexity of the "plug-in" estimator determined by the sample orthogonal iteration method. Our analysis requires two auxiliary results. The first is a useful singular value stability result that upper bounds the Ky Fan k-norm difference between two matrices (see Lemma C.1 in appendix C.1), and the second is a vector generalization of Bernstein's inequality (see Lemma C.6 in appendix C.2). In order to prove an upper bound on the *mean square error* (MSE) of the plug-in estimator $\|\hat{B}_n\|_{(1,k)}$, we next derive an exponential concentration of measure inequality for $\|\hat{B}_n\|_{(1,k)}$ using these lemmata.

Proposition 4.5 (Ky Fan k-Norm Estimation Tail Bound). *For every* $0 \leq t \leq \frac{1}{\delta}\sqrt{\frac{k}{2}}$*:*

$$\mathbb{P}\left(\left|\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right| \geq t\right) \leq \exp\left(\frac{1}{4} - \frac{n\delta^2 t^2}{8k}\right).$$

Proof. For any $t \geq 0$, observe that Lemma C.1 in appendix C.1 implies:

$$\mathbb{P}\left(\left|\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right| \geq t\right) \leq \mathbb{P}\left(\left\|\hat{B}_n - \tilde{B}\right\|_{\mathsf{Fro}} \geq \frac{t}{\sqrt{k}}\right). \tag{4.96}$$

To upper bound the right hand side, consider the random matrix $V_i \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ corresponding to each sample (X_i, Y_i) for $i \in \{1, \ldots, n\}$, which is defined entry-wise as:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; [V_i]_{y,x} = \frac{\mathbb{1}\{X_i = x, Y_i = y\} - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}. \tag{4.97}$$

The random matrices $V_1, \ldots, V_n$ are i.i.d. with mean $\mathbb{E}[V_i] = \tilde{B}$, and we will construe them as vectors with ℓ^2-norm given by the Frobenius norm. Let $C = \sqrt{2}/\delta$ and $\nu = 1/\delta^2$. Then, each V_i satisfies:

$$\begin{aligned}
\left\|V_i - \tilde{B}\right\|_{\mathsf{Fro}}^2 &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{\left(\mathbb{1}\{X_i = x, Y_i = y\} - P_{X,Y}(x,y)\right)^2}{P_X(x)P_Y(y)} \\
&\leq \frac{1}{\delta^2} \max_{a\in\mathcal{X},\, b\in\mathcal{Y}} \sum_{x\neq a}\sum_{y\neq b} P_{X,Y}(x,y)^2 + (1 - P_{X,Y}(a,b))^2 \\
&\leq \frac{2}{\delta^2} = C^2 \;\; a.s.
\end{aligned} \tag{4.98}$$

where the second inequality uses (4.94) and (4.95), and the final inequality holds because $\sum_{x,y} P_{X,Y}(x,y)^2 \leq \sum_{x,y} P_{X,Y}(x,y) = 1$. Moreover, we have:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left\|V_i - \tilde{B}\right\|_{\mathsf{Fro}}^2\right] = \mathbb{E}\left[\left\|V_1 - \tilde{B}\right\|_{\mathsf{Fro}}^2\right]$$

$$\leq \frac{1}{\delta^2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbb{VAR}(\mathbb{1}\{X_1 = x, Y_1 = y\})$$
$$\leq \frac{1}{\delta^2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y)$$
$$= \nu$$

where the first equality holds because $V_1, \ldots, V_n$ are i.i.d., the second inequality uses (4.98), (4.94), and (4.95), and the third inequality uses the fact that $\mathbb{VAR}(\mathbb{1}\{X_1 = x, Y_1 = y\}) = P_{X,Y}(x,y)(1 - P_{X,Y}(x,y))$. Now notice that $\hat{B}_n = \frac{1}{n}\sum_{i=1}^n V_i$. Hence, applying the vector Bernstein inequality in Lemma C.6 of appendix C.2 to the right hand side of (4.96), we get:

$$\mathbb{P}\left(\left|\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right| \geq t\right) \leq \exp\left(\frac{1}{4} - \frac{n\delta^2 t^2}{8k}\right)$$

for every $0 \leq t \leq \frac{1}{\delta}\sqrt{\frac{k}{2}}$. This completes the proof. ■

This bound illustrates that estimating $\|\tilde{B}\|_{(1,k)}$ using $\|\hat{B}_n\|_{(1,k)}$ to within for a small error of $t > 0$, and with a confidence level (or probability) of at least $1 - \epsilon$ for some small $\epsilon > 0$, requires n to grow linearly with k. Hence, Proposition 4.5 characterizes the sample complexity of the plug-in estimator $\|\hat{B}_n\|_{(1,k)}$. We now use this tail bound to establish an upper bound on the MSE of this plug-in estimator.

Theorem 4.5 (Ky Fan k-Norm Estimation MSE Bound). *For every sufficiently large $n \geq 4$ such that $\log(4kn) \leq n/16$, the following minimax upper bound holds:*

$$\inf_{\hat{f}_n(\cdot)} \sup_{\substack{P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}: \\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\hat{f}_n(X_1^n, Y_1^n) - \left\|\tilde{B}\right\|_{(1,k)}\right)^2\right]$$
$$\leq \sup_{\substack{P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}: \\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right)^2\right]$$
$$\leq \frac{6k + 8k\log(nk)}{n\delta^2}$$

where the infimum is over all estimators $\hat{f}_n : \mathcal{X}^n \times \mathcal{Y}^n \to [0, \infty)$ of $\|\tilde{B}\|_{(1,k)}$ based on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with knowledge of P_X and P_Y, the suprema are over all couplings $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ with fixed marginals P_X and P_Y that satisfy (4.94) and (4.95), and the expectations are with respect to the product distribution of the i.i.d. data samples.

Proof. The first inequality is trivially true since $\|\hat{B}_n\|_{(1,k)}$ is a valid estimator of $\|\tilde{B}\|_{(1,k)}$. The second inequality turns out to be an immediate consequence of Proposition 4.5. To prove it, fix any coupling $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ with marginal pmfs P_X and P_Y

satisfying (4.94) and (4.95). Observe that:[65]

$$\begin{aligned}\left|\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right| &\leq \left\|\hat{B}_n\right\|_{(1,k)} + \left\|\tilde{B}\right\|_{(1,k)} \\ &\leq k\left(1 + \left\|\hat{B}_n\right\|_{\mathsf{op}}\right) \\ &\leq k\left(1 + \sqrt{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{\left(\hat{P}^n_{X,Y}(x,y) - P_X(x)P_Y(y)\right)^2}{P_X(x)P_Y(y)}}\right) \\ &\leq k\left(1 + \frac{1}{\delta}\sqrt{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \hat{P}^n_{X,Y}(x,y)^2 + P_X(x)^2P_Y(y)^2}\right) \\ &\leq k\left(1 + \frac{\sqrt{2}}{\delta}\right) \quad a.s. \end{aligned} \tag{4.99}$$

where the second inequality follows from part 1 of Theorem 4.1, the third inequality follows from $\|\hat{B}_n\|_{\mathsf{op}} \leq \|\hat{B}_n\|_{\mathsf{Fro}}$, the fourth inequality uses (4.94), (4.95), and the fact that $(a-b)^2 \leq a^2 + b^2$ for $a, b \geq 0$, and the final inequality holds because $\sum_{x,y} Q_{X,Y}(x,y)^2 \leq \sum_{x,y} Q_{X,Y}(x,y) = 1$ for any joint pmf $Q_{X,Y}$. Next, define the event $E = \{|\|\hat{B}_n\|_{(1,k)} - \|\tilde{B}\|_{(1,k)}| \geq t\}$ for any $0 \leq t \leq \sqrt{k}/(\delta\sqrt{2})$. Using the law of total expectation, we have:

$$\begin{aligned}\mathbb{E}\left[\left(\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right)^2\right] &= \mathbb{E}\left[\left(\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right)^2 \middle| E^c\right]\mathbb{P}(E^c) \\ &\quad + \mathbb{E}\left[\left(\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right)^2 \middle| E\right]\mathbb{P}(E) \\ &\leq t^2 + k^2\left(1 + \frac{\sqrt{2}}{\delta}\right)^2 \mathbb{P}(E)\end{aligned}$$

where the second inequality holds due to (4.99) and the bound $\mathbb{P}(E^c) \leq 1$. Then, we can employ Proposition 4.5 and optimize over t to produce the bound:

$$\begin{aligned}&\sup_{\substack{P_{X,Y}\in\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}:\\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right)^2\right] \\ &\qquad \leq \min_{0\leq s\leq \frac{k}{2\delta^2}} s + k^2\left(1 + \frac{\sqrt{2}}{\delta}\right)^2 \exp\left(\frac{1}{4} - \frac{n\delta^2 s}{8k}\right)\end{aligned}$$

[65] Since $\hat{P}_{X_1^n} \approx P_X$ and $\hat{P}_{Y_1^n} \approx P_Y$ in the regime of interest, $\|\hat{B}_n\|_{\mathsf{op}} \lesssim 1$ and we intuitively expect that the bound in (4.99) can be improved to $|\|\hat{B}_n\|_{(1,k)} - \|\tilde{B}\|_{(1,k)}| \lesssim 2k$ *a.s.* However, we rigorously obtain the weaker bound in (4.99) because we define $\hat{B}_n$ according to (4.76) (instead of as the true CDM corresponding to $\hat{P}_{X_1^n,Y_1^n}$) in order to enable straightforward applications of well-known exponential concentration of measure inequalities.

where we use the change of variables $s = t^2$.

Consider the function $F : \mathbb{R} \to \mathbb{R}$, $F(s) = s + \alpha \exp(-\beta s)$, where $\alpha = k^2(1 + (\sqrt{2}/\delta))^2 \exp\left(\frac{1}{4}\right)$ and $\beta = n\delta^2/(8k)$. A straightforward calculus argument shows that the global minimum of F occurs at $s^\star = \log(\alpha\beta)/\beta$ and $F(s^\star) = (1+\log(\alpha\beta))/\beta$. Hence, we have:

$$\begin{aligned}\sup_{\substack{P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}: \\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\left\|\hat{B}_n\right\|_{(1,k)} - \left\|\tilde{B}\right\|_{(1,k)}\right)^2\right] & \\ \leq \frac{1 + \log\left(\frac{1}{8}\exp\left(\frac{1}{4}\right) nk\delta^2 \left(1 + \frac{\sqrt{2}}{\delta}\right)^2\right)}{\frac{n\delta^2}{8k}} & \\ \leq \frac{6k + 8k\log(nk)}{n\delta^2} &\end{aligned}$$

where the second inequality uses the inequalities $\delta \leq \frac{1}{2}$ (since $|\mathcal{X}|, |\mathcal{Y}| \geq 2$), $\delta + \sqrt{2} \leq 2$, and $\log(2) \geq \frac{1}{2}$, and we assume that $0 \leq s^\star \leq \frac{k}{2\delta^2}$, or equivalently, that:

$$0 \leq \frac{\frac{1}{4} - \log(8) + \log\left(nk\left(\delta + \sqrt{2}\right)^2\right)}{n} \leq \frac{1}{16}.$$

To satisfy the left hand side inequality, it suffices to take $n \geq 4$ since $k \geq 1$, $(\delta + \sqrt{2})^2 \geq 2$, and $\exp\left(-\frac{1}{4}\right) \leq 1$. On the other hand, to satisfy the right hand side inequality, it suffices to take n sufficiently large so that $\log(4kn) \leq \frac{n}{16}$ since $\frac{1}{4} < \log(8)$ and $\delta + \sqrt{2} \leq 2$. This completes the proof. ■

Similar to Proposition 4.5, Theorem 4.5 also portrays the relationship between n and k to achieve a given minimax value of MSE. In particular, larger values k require more samples n to achieve the same MSE value. Hence, estimating the entire nuclear norm of $\tilde{B}$ (which corresponds to estimating all principal modes of $\tilde{B}$, i.e. $k = \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$) requires far more samples than estimating the first few modes of $\tilde{B}$. More generally, the structure of the Lidskii inequality in Proposition C.2 of appendix C.1 implies that the results in Proposition 4.5 and Theorem 4.5 hold for estimating the sum of any k singular values of $\tilde{B}$ (rather than just $\|\tilde{B}\|_{(1,k)}$) using the corresponding plug-in estimator.

We do not prove a minimax lower bound in this chapter to characterize the precise minimax rate for this problem, because we are only interested in analyzing the sample orthogonal iteration method. In fact, evidence from the closely related matrix estimation literature suggests that the plug-in estimator may not be minimax optimal, and even if it is minimax optimal, establishing this optimality is likely to be challenging, cf. [45, Section 2.1]. So, we leave the development of "good" minimax lower bounds as a future research endeavor.

■ 4.6.2 Estimation of Dominant Information Vectors

Since we use the sample orthogonal iteration method to find the dominant k singular vectors, or information vectors, of the "empirical CDM" $\hat{B}_n$ defined in (4.76), we analyze the sample complexity of estimating the dominant k singular vectors of the CDM $\tilde{B}$ in this subsection. (This arguably captures the consistency of the sample orthogonal iteration method more meaningful than the analysis in subsection 4.6.1.) For simplicity, we only consider estimation of the dominant k right singular vectors $\{\psi_2, \dots, \psi_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{X}|}$ of $\tilde{B}$. We let $\{\hat{\psi}_2, \dots, \hat{\psi}_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{X}|}$ be the dominant k right singular vectors of $\hat{B}_n$ produced by the sample orthogonal iteration method, which are plug-in estimators of the true information vectors $\{\psi_2, \dots, \psi_{k+1}\}$. Unfortunately, despite the existence of singular subspace stability results like *Wedin's theorem*, cf. [263, Theorem 4], and the *Davis-Kahan* $\sin(\Theta)$ *theorem*, cf. [294, Theorems 1 and 3], the individual singular vectors of a matrix can vary greatly even under small matrix perturbations. So, instead of analyzing the convergence of each $\hat{\psi}_i$ to ψ_i (which could be done by imposing additional separation conditions on the singular values of $\tilde{B}$), we will analyze the convergence of $\|\tilde{B}\hat{\Psi}_{(k)}\|_{\mathsf{Fro}}^2$ to $\|\tilde{B}\Psi_{(k)}\|_{\mathsf{Fro}}^2$, where we define:

$$\Psi_{(k)} \triangleq [\psi_2 \cdots \psi_{k+1}] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|}) \tag{4.100}$$

$$\hat{\Psi}_{(k)} \triangleq \left[\hat{\psi}_2 \cdots \hat{\psi}_{k+1}\right] \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|}) \tag{4.101}$$

and we recognize $\|B\Psi_{(k)}\|_{\mathsf{Fro}}^2$ as the squared $(2,k)$-norm of $\tilde{B}$ (see (C.3) in appendix C.1):

$$\left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2 = \left\|\tilde{B}\right\|_{(2,k)}^2 = \sum_{i=2}^{k+1} \sigma_i^2 \,. \tag{4.102}$$

In the ensuing analysis, we will prove a minimax upper bound on the MSE between $\|\tilde{B}\hat{\Psi}_{(k)}\|_{\mathsf{Fro}}^2$ and $\|\tilde{B}\Psi_{(k)}\|_{\mathsf{Fro}}^2$. This formulation is agnostic to the instability of individual singular vectors under perturbations because it uses subspaces of singular vectors. Moreover, it can be construed as utilizing a loss function that is more suited to our setting. Indeed, $\|\tilde{B}\Psi_{(k)}\|_{\mathsf{Fro}}^2$ can be perceived as a "rank k approximation" of mutual χ^2-information (see part 2 of Theorem 4.3) or standard mutual information under a weak dependence assumption (see Proposition 4.4), and $\|\tilde{B}\hat{\Psi}_{(k)}\|_{\mathsf{Fro}}^2$ estimates this "rank k approximation."

As before, our analysis requires two auxiliary results. The first is a certain stability result for the squared $(2,k)$-norm of a matrix (see Lemma C.2 in appendix C.1), and the second is a matrix version of Bernstein's inequality (see Lemma C.7 in appendix C.2). In order to prove our MSE upper bound, we first prove an exponential concentration of measure inequality for $\|\tilde{B}\hat{\Psi}_{(k)}\|_{\mathsf{Fro}}^2$.

Proposition 4.6 (Information Vector Estimation Tail Bound). *For every* $0 \leq t \leq 4k$*:*

$$\mathbb{P}\left(\left|\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right| \geq t\right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(-\frac{n\delta t^2}{64k^2}\right).$$

Proof. For any $t \geq 0$, observe that Lemma C.2 in appendix C.1 implies:

$$\mathbb{P}\left(\left|\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right| \geq t\right) \leq \mathbb{P}\left(\left\|\hat{B}_n - \tilde{B}\right\|_{\mathsf{op}} \geq \frac{t}{4k}\right) \tag{4.103}$$

where we use the fact that $\|\tilde{B}\|_{\mathsf{op}} \leq 1$ (see part 1 of Theorem 4.1). Next, as in (4.97) in the proof of Proposition 4.5, define the i.i.d. random matrices $V_1, \ldots, V_n \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ with mean $\mathbb{E}[V_i] = \tilde{B}$ corresponding to the samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, respectively. Furthermore, define the random matrix $Z_i \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ corresponding to each sample (X_i, Y_i) for $i \in \{1, \ldots, n\}$ entry-wise as:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; [Z_i]_{y,x} = \frac{\mathbb{1}\{X_i = x, Y_i = y\}}{\sqrt{P_X(x)P_Y(y)}}\,.$$

It is straightforward to verify that $V_i - \tilde{B} = Z_i - B$ *a.s.* for every $i \in \{1, \ldots, n\}$, where $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ is the true DTM corresponding to $P_{X,Y}$. Let $C = \nu = 1 + \frac{1}{\delta}$. Now notice that each V_i satisfies:

$$\begin{aligned}
\left\|V_i - \tilde{B}\right\|_{\mathsf{op}} &= \|Z_i - B\|_{\mathsf{op}} \\
&\leq \|Z_i\|_{\mathsf{op}} + \|B\|_{\mathsf{op}} \\
&\leq 1 + \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{1}{\sqrt{P_X(x)P_Y(y)}} \\
&\leq 1 + \frac{1}{\delta} = C \;\; a.s.
\end{aligned}$$

where the second inequality follows from the triangle inequality, the third inequality uses part 1 of Theorem 4.1 and the easily verifiable fact that $\|Z_i\|_{\mathsf{op}} = 1/\sqrt{P_X(x)P_Y(y)}$ with probability $P_{X,Y}(x, y)$, and the final inequality holds due to (4.94) and (4.95). Moreover, we have:

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbb{COV}(V_i)\right\|_{\mathsf{op}} &= \|\mathbb{COV}(V_1)\|_{\mathsf{op}} \\
&= \left\|\mathbb{E}\left[(Z_1 - B)(Z_1 - B)^T\right]\right\|_{\mathsf{op}} \\
&= \left\|\mathbb{E}\left[Z_1 Z_1^T\right] - BB^T\right\|_{\mathsf{op}} \\
&\leq \left\|\mathbb{E}\left[Z_1 Z_1^T\right]\right\|_{\mathsf{op}} + \left\|BB^T\right\|_{\mathsf{op}} \\
&= 1 + \left\|\mathbb{E}\left[Z_1 Z_1^T\right]\right\|_{\mathsf{op}} \\
&= 1 + \max_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \frac{P_{X|Y}(x|y)}{P_X(x)} \\
&\leq 1 + \frac{1}{\delta} \max_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)
\end{aligned}$$

$$= 1 + \frac{1}{\delta} = \nu$$

where the first equality holds because $V_1, \ldots, V_n$ are i.i.d., the second equality follows from the definition of covariance of a random matrix in Lemma C.7 in appendix C.2, the fourth inequality uses the triangle inequality, the fifth equality holds because $\|BB^T\|_{\mathsf{op}} = 1$ (see part 1 of Theorem 4.1), the sixth equality holds because a straightforward calculation yields that $\mathbb{E}[Z_1 Z_1^T]$ is a $|\mathcal{Y}| \times |\mathcal{Y}|$ diagonal matrix with diagonal entries:

$$\forall y \in \mathcal{Y}, \quad \left[\mathbb{E}\left[Z_1 Z_1^T\right]\right]_{y,y} = \sum_{x \in \mathcal{X}} \frac{P_{X|Y}(x|y)}{P_X(x)},$$

and the seventh inequality holds due to (4.94). Likewise, we also have:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{COV}\left(V_i^T\right) \right\|_{\mathsf{op}} \leq 1 + \frac{1}{\delta} = \nu \,.$$

Now notice that $\hat{B}_n = \frac{1}{n}\sum_{i=1}^n V_i$. Hence, applying the matrix Bernstein inequality in Lemma C.7 of appendix C.2 to the right hand side of (4.103), we get for every $0 \leq t \leq 4k$:

$$\begin{aligned}
\mathbb{P}\left(\left| \left\| \tilde{B}\hat{\Psi}_{(k)} \right\|_{\mathsf{Fro}}^2 - \left\| \tilde{B}\Psi_{(k)} \right\|_{\mathsf{Fro}}^2 \right| \geq t \right) &\leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(-\frac{3n\delta t^2}{128k^2(1+\delta)} \right) \\
&\leq (|\mathcal{X}| + |\mathcal{Y}|) \exp\left(-\frac{n\delta t^2}{64k^2} \right)
\end{aligned}$$

where we use the inequality $\delta \leq \frac{1}{2}$ (since $|\mathcal{X}|, |\mathcal{Y}| \geq 2$) in the second inequality. This completes the proof. ■

The bound in Proposition 4.6 illustrates that estimating the right singular vectors $\{\psi_2, \ldots, \psi_{k+1}\} \subseteq \mathbb{R}^{|\mathcal{X}|}$ of $\tilde{B}$ corresponding to its k largest singular values to within for a small error of $t > 0$, and with a confidence level (or probability) of at least $1 - \epsilon$ for some small $\epsilon > 0$, requires n to grow quadratically with k. Thus, Proposition 4.6 characterizes the sample complexity of estimating the dominant k information vectors of $\tilde{B}$. We next utilize this tail bound to establish an upper bound on the MSE between $\|\tilde{B}\hat{\Psi}_{(k)}\|_{\mathsf{Fro}}^2$ and $\|\tilde{B}\Psi_{(k)}\|_{\mathsf{Fro}}^2$.

Theorem 4.6 (Information Vector Estimation MSE Bound). *For every sufficiently large $n \geq \frac{4}{\delta(|\mathcal{X}|+|\mathcal{Y}|)}$ such that $\log(n\delta(|\mathcal{X}| + |\mathcal{Y}|)) \leq \frac{n\delta}{4}$, the following minimax upper bound holds:*

$$\inf_{\hat{f}_n(\cdot)} \sup_{\substack{P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}: \\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}} \left[\left(\left\| \tilde{B}\hat{f}_n(X_1^n, Y_1^n) \right\|_{\mathsf{Fro}}^2 - \left\| \tilde{B}\Psi_{(k)} \right\|_{\mathsf{Fro}}^2 \right)^2 \right]$$

$$\leq \sup_{\substack{P_{X,Y}\in\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}:\\ P_X,P_Y\geq\delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right)^2\right]$$
$$\leq \frac{64k^2\log(n\delta(|\mathcal{X}|+|\mathcal{Y}|)) - 16k^2}{n\delta}$$

where the infimum is over all estimators $\hat{f}_n : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ *of* $\Psi_{(k)}$ *based on the data* $(X_1,Y_1),\ldots,(X_n,Y_n)$ *with knowledge of* P_X *and* P_Y*, the suprema are over all couplings* $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ *with fixed marginals* P_X *and* P_Y *that satisfy* (4.94) *and* (4.95)*, and the expectations are with respect to the product distribution of the i.i.d. data samples.*

Proof. As in subsection 4.6.1, the first inequality is trivially true since $\hat{\Psi}_{(k)}$ is a valid estimator of $\Psi_{(k)}$, and the second inequality turns out to be an immediate consequence of Proposition 4.6. Indeed, first observe using (C.12) (in the proof of Lemma C.2 in appendix C.1) and the bound $\|\tilde{B}\|_{\mathsf{op}} \leq 1$ (see part 1 of Theorem 4.1) that:

$$\begin{aligned}
\left|\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right| &\leq 2\sum_{i=2}^{k+1}\left|\left\|\tilde{B}\psi_i\right\|_2 - \left\|\tilde{B}\hat{\psi}_i\right\|_2\right| \\
&\leq 2\sum_{i=2}^{k+1}\left\|\tilde{B}\left(\psi_i - \hat{\psi}_i\right)\right\|_2 \\
&\leq 2\sum_{i=2}^{k+1}\left\|\psi_i - \hat{\psi}_i\right\|_2 \\
&\leq 2\sum_{i=2}^{k+1}\|\psi_i\|_2 + \left\|\hat{\psi}_i\right\|_2 \\
&= 4k \;\; a.s. \qquad (4.104)
\end{aligned}$$

where the second inequality follows from the reverse triangle inequality, the third inequality holds because $\|\tilde{B}\|_{\mathsf{op}} \leq 1$, the fourth inequality follows from the triangle inequality, and the final equality holds because $\|\psi_i\|_2 = \|\hat{\psi}_i\|_2 = 1$. Next, define the event $E = \{|\|\tilde{B}\hat{\Psi}_{(k)}\|_{\mathsf{Fro}}^2 - \|\tilde{B}\Psi_{(k)}\|_{\mathsf{Fro}}^2| \geq t\}$ for any $0 \leq t \leq 4k$. Using the law of total expectation, we have:

$$\begin{aligned}
\mathbb{E}\left[\left(\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right)^2\right] &= \mathbb{E}\left[\left(\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right)^2 \middle| E^c\right]\mathbb{P}(E^c) \\
&\quad + \mathbb{E}\left[\left(\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right)^2 \middle| E\right]\mathbb{P}(E) \\
&\leq t^2 + 16k^2(|\mathcal{X}|+|\mathcal{Y}|)\exp\left(-\frac{n\delta t^2}{64k^2}\right)
\end{aligned}$$

where the second inequality holds due to Proposition 4.6, (4.104), and the bound $\mathbb{P}(E^c) \leq 1$. Then, we can use the change of variables $s = t^2$ and optimize over s to produce the bound:

$$\sup_{\substack{P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}: \\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right)^2\right] \leq \min_{0\leq s\leq 16k^2} s + 16k^2(|\mathcal{X}|+|\mathcal{Y}|)\exp\left(-\frac{n\delta s}{64k^2}\right).$$

Finally, using the straightforward calculus argument from the proof of Theorem 4.5, we have:

$$\sup_{\substack{P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}: \\ P_X, P_Y \geq \delta \text{ entry-wise}}} \mathbb{E}_{P_{X,Y}}\left[\left(\left\|\tilde{B}\hat{\Psi}_{(k)}\right\|_{\mathsf{Fro}}^2 - \left\|\tilde{B}\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right)^2\right] \leq \frac{1+\log\left(\frac{1}{4}(|\mathcal{X}|+|\mathcal{Y}|)n\delta\right)}{\frac{n\delta}{64k^2}} \leq \frac{64k^2\log(n\delta(|\mathcal{X}|+|\mathcal{Y}|)) - 16k^2}{n\delta}$$

where the second inequality uses the bound $\log(4) \geq \frac{5}{4}$, and we must assume that:

$$0 \leq \frac{\log\left(16k^2(|\mathcal{X}|+|\mathcal{Y}|)\frac{n\delta}{64k^2}\right)}{\frac{n\delta}{64k^2}} \leq 16k^2$$

or equivalently that:

$$0 \leq \frac{\log(n\delta(|\mathcal{X}|+|\mathcal{Y}|)) - \log(4)}{n\delta} \leq \frac{1}{4}.$$

To satisfy the left hand side inequality, it suffices to take:

$$n \geq \frac{4}{\delta(|\mathcal{X}|+|\mathcal{Y}|)}.$$

On the other hand, to satisfy the right hand side inequality, it suffices to take n sufficiently large so that:

$$\log(n\delta(|\mathcal{X}|+|\mathcal{Y}|)) \leq \frac{n\delta}{4}$$

since $\log(4) > 0$. This completes the proof. ■

Much like Proposition 4.6, Theorem 4.6 also illustrates that larger values of k require more samples n to achieve the same MSE value. Thus, estimating the full right singular vector basis (i.e. all $k = \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1$ information vectors) of $\tilde{B}$ requires far more samples than estimating the first few dominant information vectors of $\tilde{B}$. More generally, Proposition 4.6 and Theorem 4.6 continue to hold for estimating any arbitrary k right singular vectors of $\tilde{B}$ rather than the top k singular vectors. Finally, as before, we leave the development of corresponding minimax lower bounds as a future research endeavor.

■ 4.6.3 Comparison of Sanov Exponents

While subsections 4.6.1 and 4.6.2 contain the bulk of our main sample complexity analysis, we will present some auxiliary observations in this subsection and subsection 4.6.4. We assume for simplicity that $|\mathcal{X}| = O(1)$ and $|\mathcal{Y}| = O(1)$ with respect to the sample size n in this subsection. In this (classical) regime, the entire unknown joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ can be estimated accurately from the i.i.d. training samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. So, we do not specifically assume as before that the marginal pmfs $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ are known. This means that instead of using the "empirical CDM" in (4.76), the sample orthogonal iteration method uses the CDM:

$$\tilde{B}_n \triangleq \beta\Big(\hat{P}_{X_1^n,Y_1^n}\Big) - \sqrt{\hat{P}_{Y_1^n}}^T \sqrt{\hat{P}_{X_1^n}} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} \tag{4.105}$$

corresponding to the empirical joint distribution $\hat{P}_{X_1^n,Y_1^n}$ in steps 4 and 6 and the termination condition in Algorithm 1, where the map $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}$ is defined in (4.19).[66] Furthermore, we also assume for convenience that $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}^{\circ}$, which implies that the corresponding DTM $B \in \mathcal{B}^{\circ}$, cf. (4.21).

We now illustrate that estimating the dominant $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$ (where $k = O(1)$) right singular vectors of the CDM $\tilde{B}$ using the sample orthogonal iteration method is more efficient than estimating the entire DTM B using the plug-in estimator $\beta(\hat{P}_{X_1^n,Y_1^n})$ in a Sanov exponent sense. To proceed with our analysis, for any $\tau > 0$, let us define the sets:

$$R_\tau \triangleq \Big\{Q_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} : \|\beta(Q_{X,Y}) - B\|_{\mathsf{op}} \leq \tau\Big\} \tag{4.106}$$

$$S_\tau \triangleq \left\{Q_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} : \left|\left\|B\Psi_{(k)}^{\beta(Q_{X,Y}) - \sqrt{Q_Y}^T\sqrt{Q_X}}\right\|_{\mathsf{Fro}}^2 - \left\|B\Psi_{(k)}\right\|_{\mathsf{Fro}}^2\right| \leq \tau\right\} \tag{4.107}$$

where $\beta(Q_{X,Y}) - \sqrt{Q_Y}^T\sqrt{Q_X} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ denotes the CDM corresponding to $Q_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ (and $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $Q_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ are the associated marginal pmfs), $\Psi_{(k)}^A \in \mathcal{V}_k(\mathbb{R}^{|\mathcal{X}|})$ denotes the orthonormal k-frame that collects the leading k right singular vectors of a matrix $A \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ (see the definition in Lemma C.2 in appendix C.1), and $\Psi_{(k)} = \Psi_{(k)}^{\tilde{B}}$ is defined in (4.100). Definition (4.106) conveys that the estimation error between the plug-in estimator $\beta(\hat{P}_{X_1^n,Y_1^n})$ and the true DTM B will be measured using the operator norm. Likewise, definition (4.107) conveys that the estimation error between plug-in estimator $\Psi_{(k)}^{\tilde{B}_n}$ (produced by the sample orthogonal iteration method) and $\Psi_{(k)}$ will be measured in terms of the absolute deviation between the squared Frobenius norms $\|B\Psi_{(k)}^{\tilde{B}_n}\|_{\mathsf{Fro}}^2$ and $\|B\Psi_{(k)}\|_{\mathsf{Fro}}^2$; this is motivated by the discussion at the outset of subsection 4.6.2. Clearly, the sets R_τ and S_τ are non-empty as they contain the true

[66] Note that the definition in (4.19) can be extended to hold for all joint pmfs rather than just those in $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. Indeed, if $\hat{P}_{X_1^n}(x) = 0$ for some $x \in \mathcal{X}$, then the xth column of $\beta\big(\hat{P}_{X_1^n,Y_1^n}\big)$ is defined to be zero. Likewise, if $\hat{P}_{Y_1^n}(y) = 0$ for some $y \in \mathcal{Y}$, then the yth row of $\beta\big(\hat{P}_{X_1^n,Y_1^n}\big)$ is defined to be zero.

distribution $P_{X,Y}$, and bounded since $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ is bounded in its ambient Euclidean space. The next lemma identifies some other properties of these sets. (We note that part 2 of Lemma 4.1 is analogous to (4.103) in the proof of Proposition 4.6.)

Lemma 4.1 (Properties of R_τ and S_τ). *Let $b_{\mathsf{min}} > 0$ denote the minimum entry of the DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$:*

$$b_{\mathsf{min}} \triangleq \min_{x\in\mathcal{X},\, y\in\mathcal{Y}} [B]_{y,x}\,.$$

Then, we have:

1. *For every $0 < \tau < b_{\mathsf{min}}$, $R_\tau \subseteq \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}}$ is a compact set.*
2. *For every $\tau > 0$, $R_\tau \subseteq S_{4k\tau}$.*

Proof. See appendix C.4. ■

Since the sequence of empirical pmfs $\hat{P}_{X_1^n,Y_1^n}$ converges to $P_{X,Y}$ *a.s.* by the SLLN as $n \to \infty$, and the map $\beta : \mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}$ is continuous (as shown in part 3 of Theorem 4.2), $\beta(\hat{P}_{X_1^n,Y_1^n})$ converges to B *a.s.* as $n \to \infty$ (by the *Mann-Wald continuous mapping theorem*). Hence, $\beta(\hat{P}_{X_1^n,Y_1^n})$ is a consistent estimator of B, and we have convergence in probability:[67]

$$\forall \tau > 0,\quad \lim_{n\to\infty} \mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in R_\tau^c\Big) = 0\,. \tag{4.108}$$

We expect the probability in (4.108) to decay exponentially, and the rate of its decay determines the efficiency of estimating B using $\beta(\hat{P}_{X_1^n,Y_1^n})$. Similarly, we also expect:

$$\forall \tau > 0,\quad \lim_{n\to\infty} \mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in S_\tau^c\Big) = 0 \tag{4.109}$$

since $|\|B\Psi_{(k)}^{\tilde{B}_n}\|_{\mathsf{Fro}}^2 - \|B\Psi_{(k)}\|_{\mathsf{Fro}}^2| \le 4k\|\beta(\hat{P}_{X_1^n,Y_1^n}) - B\|_{\mathsf{op}} \to 0$ *a.s.* as $n \to \infty$ using (C.25) in appendix C.4, and the exponential rate of decay here determines the efficiency of estimating $\|\tilde{B}\|_{(2,k)}^2$, cf. (4.102), via the sample orthogonal iteration method.

We next define the *information projection* problems, cf. [59, Section 3]:[68]

$$I_{\mathsf{DTM}}(\tau) \triangleq \inf_{Q_{X,Y}\in R_\tau^c} D(Q_{X,Y}||P_{X,Y}) \tag{4.110}$$

$$I_{\mathsf{ACE}}(\tau) \triangleq \inf_{Q_{X,Y}\in S_\tau^c} D(Q_{X,Y}||P_{X,Y}) \tag{4.111}$$

for any $\tau > 0$. Using Sanov's theorem from large deviations theory (see Theorem C.2 in appendix C.2), we have for any $0 < \tau < b_{\mathsf{min}}$:

$$I_{\mathsf{DTM}}(\tau) = \lim_{n\to\infty} -\frac{1}{n}\log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in R_\tau^c\Big)\Big) \tag{4.112}$$

[67] Set theoretic complements of R_τ and S_τ are taken with respect to the probability simplex of all joint pmfs on $\mathcal{X} \times \mathcal{Y}$.

[68] We refer to the problem in (4.111) as $I_{\mathsf{ACE}}(\tau)$ because the sample orthogonal iteration method and the sample extended ACE algorithm are essentially equivalent.

where the limit is well-defined and the equality holds because R_τ^c is an open set since R_τ is closed according to part 1 of Lemma 4.1. On the other hand, Sanov's theorem (see Theorem C.2) also gives for any $\tau > 0$:

$$I_{\mathsf{ACE}}(\tau) \leq \liminf_{n\to\infty} -\frac{1}{n}\log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in S_\tau^c\Big)\Big) \tag{4.113}$$

where we have equality and the inferior limit becomes a limit if S_τ is closed, or the interior of the closure of S_τ is contained in S_τ. According to (4.112) and (4.113), (4.110) and (4.111) capture the exponential rates of decay of the probabilities in (4.108) and (4.109), respectively. The next proposition presents the relationship between (4.110) and (4.111).

Proposition 4.7 (Bound between Sanov Exponents). *For any error tolerance $0 < \tau < b_{\mathsf{min}}$, we have:*

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{n}\log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in S_{4k\tau}^c\Big)\Big) &\geq I_{\mathsf{ACE}}(4k\tau) \\ &\geq I_{\mathsf{DTM}}(\tau) = \lim_{n\to\infty} -\frac{1}{n}\log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in R_\tau^c\Big)\Big)\,.\end{aligned}$$

Proof. This is immediate from (4.112), (4.113), and part 2 of Lemma 4.1. Indeed, since $S_{4k\tau}^c \subseteq R_\tau^c$, we have $I_{\mathsf{ACE}}(4k\tau) \geq I_{\mathsf{DTM}}(\tau)$ using (4.110) and (4.111). This completes the proof. ■

Proposition 4.7 portrays that the exponential rate of estimating B using $\beta(\hat{P}_{X_1^n,Y_1^n})$ to within a fidelity of $0 < \tau < b_{\mathsf{min}}$ is upper bounded by the exponential rate of estimating $\|\tilde{B}\|_{(2,k)}^2$ using $\|B\Psi_{(k)}^{\tilde{B}_n}\|_{\mathsf{Fro}}^2$ (where $\Psi_{(k)}^{\tilde{B}_n}$ is produced by the sample orthogonal iteration method) to within a fidelity of $4k\tau$. Therefore, given a fixed confidence level of $1-\epsilon$ for some $\epsilon \in (0,1)$, estimating the DTM to within an error tolerance of $0 < \tau < b_{\mathsf{min}}$:

$$\mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in R_\tau^c\Big) \leq \epsilon \tag{4.114}$$

requires more data samples, i.e. larger n, in general than estimating $\|\tilde{B}\|_{(2,k)}^2$ to within an error tolerance of $4k\tau$:

$$\mathbb{P}\Big(\hat{P}_{X_1^n,Y_1^n} \in S_{4k\tau}^c\Big) \leq \epsilon\,. \tag{4.115}$$

Therefore, when $k \ll \min\{|\mathcal{X}|,|\mathcal{Y}|\}$, the fidelities τ and $4k\tau$ are comparable, and it is indeed more efficient to estimate the dominant information vectors of the CDM $\tilde{B}$ compared to estimating the entire DTM B.

■ 4.6.4 Heuristic Comparison of Local Chernoff Exponents

In this subsection, we perform some heuristic analysis to illustrate that there is often a sample complexity gain in estimating maximal correlation, cf. (4.1), over estimating a single value of the DTM. The bulk of our analysis in subsections 4.6.1 and 4.6.2 has

focused on obtaining a trade-off between the sample size n and the number of modes k, while the trade-off with alphabet sizes $|\mathcal{X}|$ and $|\mathcal{Y}|$ has only been implicitly obtained (through the constant δ). In contrast, we consider estimating only *one* principal mode (i.e. $k = 1$) here, and focus on sample complexity with respect to $|\mathcal{X}|$ and $|\mathcal{Y}|$.

Much like subsection 4.6.3, we assume for convenience in this subsection that $|\mathcal{X}| = |\mathcal{Y}| = K$ and $K = O(1)$ with respect to the sample size n, and that the unknown joint distribution $P_{X,Y} \in \mathcal{P}^{\circ}_{\mathcal{X}\times\mathcal{Y}}$. Furthermore, we also assume that the marginal pmfs $P_X \in \mathcal{P}^{\circ}_{\mathcal{X}}$ and $P_Y \in \mathcal{P}^{\circ}_{\mathcal{Y}}$ are known as in subsections 4.6.1 and 4.6.2. Now suppose we observe the i.i.d. training samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from $P_{X,Y}$ as before. Consider the problem of estimating the correlation $\mathbb{E}[f(X)g(Y)]$ between a given pair of functions $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ with the constraints:

$$\mathbb{E}\Big[f(X)^2\Big] = \mathbb{E}\Big[g(Y)^2\Big] = 1 \tag{4.116}$$

from these training samples. (Note that we do not restrict the functions to have zero mean.) In order to gauge the sample complexity of estimating $\mathbb{E}[f(X)g(Y)]$, we seek to determine the rate at which the plug-in estimator:

$$\widehat{\mathbb{E}}_n[f(X)g(Y)] = \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(Y_i) \tag{4.117}$$

where $\widehat{\mathbb{E}}_n[\cdot]$ denotes the empirical expectation operator corresponding to the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ (see (C.14) in appendix C.2), converges to $\mathbb{E}[f(X)g(Y)]$ in probability as $n \to \infty$. To facilitate "back-of-the-envelope" calculations of such rates (or error exponents), we present a tight characterization of the relevant Chernoff exponent in the vanishing precision level limit.

Proposition 4.8 (Local Approximation of Chernoff Exponent for Bivariate Distributions). *Given the bivariate distribution $P_{X,Y} \in \mathcal{P}^{\circ}_{\mathcal{X}\times\mathcal{Y}}$, for any functions $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ satisfying (4.116) such that $\mathbb{E}[f(X)g(Y)] \neq 0$ and $\mathbb{VAR}(f(X)g(Y)) > 0$, we have:*

$$-\lim_{\Delta\to 0^+}\frac{1}{\Delta^2}\lim_{n\to\infty}\frac{1}{n}\log\left(\mathbb{P}\left(\left|\frac{\widehat{\mathbb{E}}_n[f(X)g(Y)]}{\mathbb{E}[f(X)g(Y)]} - 1\right| \geq \Delta\right)\right) = \frac{\mathbb{E}[f(X)g(Y)]^2}{2\,\mathbb{VAR}(f(X)g(Y))}\,.$$

Proof. This follows from Lemma C.3 in appendix C.2 by considering the pair of discrete random variables (X, Y) as a single discrete random variable, and letting $t(x, y) = f(x)g(y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ in Lemma C.3. ■

Proposition 4.8 illustrates that the large deviations rate of decay of the probability that $\widehat{\mathbb{E}}_n[f(X)g(Y)]$ has relative error greater than or equal to $\Delta > 0$ in estimating $\mathbb{E}[f(X)g(Y)]$ is inversely proportional to the squared *coefficient of variation* of $f(X)g(Y)$ as $\Delta \to 0^+$. Since the squared coefficient of variation is more tractable than general Sanov or Chernoff exponents, we will compare estimation of maximal correlation with estimation of a particular entry of the DTM in terms of local Chernoff exponents.

Specifically, we consider the first (leading) pair of maximal correlation functions $f_2 \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $g_2 \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ whose correlation is $\rho_{\mathsf{max}}(X;Y) = \mathbb{E}[f_2(X)g_2(Y)]$ (see part 4 of Theorem 4.1), and the pair of functions $\check{f}_{x_0} \in \mathcal{L}^2(\mathcal{X}, P_X)$ and $\check{g}_{y_0} \in \mathcal{L}^2(\mathcal{Y}, P_Y)$:

$$\forall x \in \mathcal{X}, \quad \check{f}_{x_0}(x) \triangleq \frac{\mathbb{1}\{x = x_0\}}{\sqrt{P_X(x_0)}} \tag{4.118}$$

$$\forall y \in \mathcal{Y}, \quad \check{g}_{y_0}(y) \triangleq \frac{\mathbb{1}\{y = y_0\}}{\sqrt{P_Y(y_0)}} \tag{4.119}$$

for arbitrary choices of $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$, whose correlation is:

$$\mathbb{E}\Big[\check{f}_{x_0}(X)\check{g}_{y_0}(Y)\Big] = \frac{P_{X,Y}(x_0, y_0)}{\sqrt{P_X(x_0)P_Y(y_0)}} = [B]_{y_0,x_0} \tag{4.120}$$

where $B \in \mathbb{R}^{K\times K}$ is the DTM corresponding to $P_{X,Y}$. It can be checked that $\check{f}_{x_0}$ and $\check{g}_{y_0}$ satisfy (4.116) (and f_2 and g_2 clearly satisfy (4.116) by definition). As both P_X and P_Y are precisely given, we treat the estimation of (4.120) as the same as the estimation of the entry $P_{X,Y}(x_0, y_0)$ of the joint distribution. The correlation between f_2 and g_2 is generally high due to (4.1). Hence, $\check{f}_{x_0}$ and $\check{g}_{y_0}$ typically have smaller correlation than f_2 and g_2. When K is large, we will argue that this correlation gap is particularly large. This results in the estimation of $\rho_{\mathsf{max}}(X;Y)$ requiring a significantly smaller number of samples than the estimation of $P_{X,Y}(x_0, y_0)$.

Suppose X and Y are not independent so that $\rho_{\mathsf{max}}(X;Y) = \mathbb{E}[f_2(X)g_2(Y)] > 0$.[69] Then, using Proposition 4.8, observe that the ratio between the local Chernoff exponents corresponding to the estimation of $P_{X,Y}(x_0, y_0)$ and the estimation of $\rho_{\mathsf{max}}(X;Y)$ is:

$$G(P_{X,Y}) \triangleq \frac{\mathbb{E}\Big[\check{f}_{x_0}(X)\check{g}_{y_0}(Y)\Big]^2 \mathbb{VAR}(f_2(X)g_2(Y))}{\mathbb{E}[f_2(X)g_2(Y)]^2 \mathbb{VAR}\Big(\check{f}_{x_0}(X)\check{g}_{y_0}(Y)\Big)} = \frac{P_{X,Y}(x_0, y_0)}{\rho_{\mathsf{max}}(X;Y)^2} \frac{\mathbb{VAR}(f_2(X)g_2(Y))}{(1 - P_{X,Y}(x_0, y_0))} \tag{4.121}$$

where the ratio $G(P_{X,Y})$ is a function of $P_{X,Y}$. As indicated earlier, we expect the first ratio term $P_{X,Y}(x_0, y_0)/\rho_{\mathsf{max}}(X;Y)^2$ in (4.121) to decay as K grows in most cases of interest. On the other hand, when we square f_2 and g_2 element-wise, the property of "maximal correlation" is intuitively lost. Hence, we expect the second ratio term in (4.121) to be insignificant (or constant with respect to K). In principle, this intuition does not hold for all joint pmfs $P_{X,Y} \in \mathcal{P}^{\circ}_{\mathcal{X}\times\mathcal{Y}}$, but we believe that it holds for "most" joint pmfs when K is large.

To heuristically demonstrate this and establish how $G(P_{X,Y})$ scales with K, suppose $P_{X,Y}$ is randomly generated in the following way. We generate a $K \times K$ random matrix $Z \in \mathbb{R}^{K\times K}$ whose entries are i.i.d. exponential random variables with rate 1, and let $P_{X,Y}(x, y) = [Z]_{y,x}/K^2$ for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. (Strictly speaking, the resulting random $P_{X,Y}$ is not normalized, but suffices as an approximation of a valid pmf because

[69] It is straightforward to show that $\mathbb{VAR}(f_2(X)g_2(Y)) > 0$.

$\sum_{x,y} P_{X,Y}(x,y) \to 1$ *a.s.* as $K \to \infty$ by the SLLN.[70] Furthermore, $P_{X,Y}$ satisfies all the regularity conditions needed to define (4.121) *a.s.*, e.g. $P_{X,Y} \in \mathcal{P}^{\circ}_{\mathcal{X}\times\mathcal{Y}}$ *a.s.*) Then, standard calculations using exponential random variables yield:

$$\mathbb{E}[P_{X,Y}(x_0,y_0)] = \frac{1}{K^2} \tag{4.122}$$

$$\mathbb{E}\Big[\min_{x\in\mathcal{X},\, y\in\mathcal{Y}} P_{X,Y}(x,y)\Big] = \frac{1}{K^4} \tag{4.123}$$

where the expectations are with respect to the law of Z, $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$ are fixed in (4.122), and (4.123) uses the *semigroup property of exponential distributions* which shows that $\min_{x,y}[Z]_{y,x}$ has exponential distribution with rate K^2. Moreover, since $P_X(x) = (1/K^2)\sum_y Z_{y,x}$ for every $x \in \mathcal{X}$ and $KP_X(x) \to 1$ *a.s.* as $K \to \infty$ by the SLLN, we will approximate P_X with a uniform distribution. Similarly, we will approximate P_Y with a uniform distribution. Hence, an approximation to the CDM is the zero mean random matrix:

$$\tilde{B} = \frac{1}{K}\Big(Z - \mathbf{1}\mathbf{1}^T\Big) \in \mathbb{R}^{K\times K}\,. \tag{4.124}$$

Notice that the expected value of $\|\tilde{B}\|_{\mathsf{op}}^2$ satisfies the lower bound:

$$\begin{aligned}
\mathbb{E}\Big[\big\|\tilde{B}\big\|_{\mathsf{op}}^2\Big] &= \mathbb{E}\Big[\big\|\tilde{B}\tilde{B}^T\big\|_{\mathsf{op}}\Big] \\
&= \frac{1}{K^2}\,\mathbb{E}\Big[\big\|ZZ^T - Z\mathbf{1}\mathbf{1}^T - \mathbf{1}\mathbf{1}^T Z^T + K\mathbf{1}\mathbf{1}^T\big\|_{\mathsf{op}}\Big] \\
&\geq \frac{1}{K^2}\,\big\|\mathbb{E}\big[ZZ^T\big] - K\mathbf{1}\mathbf{1}^T\big\|_{\mathsf{op}} \\
&= \frac{1}{K}
\end{aligned} \tag{4.125}$$

where the third inequality follows from applying Jensen's inequality and simplifying the result, and the final equality follows from direct calculation. Intuitively, the *Marčenko-Pastur law* applied to $K\tilde{B}\tilde{B}^T$ suggests that $\mathbb{E}[\|\tilde{B}\tilde{B}^T\|_{\mathsf{op}}] = O(1/K)$, cf. [79, Theorem 4.1], [18, Theorem 3.6]. This upper bound can be deduced using (the proof techniques of) results like [18, Theorem 5.8], or [270, Theorem 2.3.8] which states that $\mathbb{E}[\|\tilde{B}\|_{\mathsf{op}}]^2 = O(1/K)$. Thus, $\mathbb{E}[\|\tilde{B}\|_{\mathsf{op}}^2]$, which represents the expected value of squared maximal correlation (see part 3 of Theorem 4.1), has the following scaling with K:

$$\mathbb{E}\Big[\big\|\tilde{B}\big\|_{\mathsf{op}}^2\Big] = \Theta\Big(\frac{1}{K}\Big) \tag{4.126}$$

which we note is a factor of $\Theta(K)$ larger than (4.122) as predicted. Therefore, for arbitrary $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$, (4.122) and (4.126) (non-rigorously) portray that the

[70] If we seek to rigorize this model, then we must set $P_{X,Y}(x,y) = [Z]_{y,x}/N$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $N = \sum_{x,y}[Z]_{y,x}$ is the normalization random variable. The resulting random pmf $P_{X,Y}$ has a *flat Dirichlet distribution* that is uniform on the probability simplex of all pmfs on $\mathcal{X}\times\mathcal{Y}$.

ratio (4.121) scales like $\Theta(1/K)$ on average, where we assume that $\mathbb{VAR}(f_2(X)g_2(Y))$ has constant scaling $\Theta(1)$ on average (as mentioned earlier). This scaling law is also illustrated in [182, Figure 1] via numerical simulations. Additionally, when $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$ are the elements with the minimum joint probability mass $P_{X,Y}(x_0, y_0) = \min_{x,y} P_{X,Y}(x, y)$, (4.123) and (4.126) (non-rigorously) portray that the ratio (4.121) scales like $\Theta(1/K^3)$ on average.

These heuristic calculations show that the maximal correlation functions f_2 and g_2 are not only good information carriers, but the correlation between them is also easier to estimate compared to other pairs of functions like $\check{f}_{x_0}$ and $\check{g}_{y_0}$. Indeed, $G(P_{X,Y})$ in (4.121) is the ratio between the number of samples required to achieve an asymptotically small precision level $\Delta > 0$ and confidence level $1 - \epsilon$, for some small $\epsilon > 0$, in the estimation of $\mathbb{E}[\check{f}_{x_0}(X)\check{g}_{y_0}(Y)]$ and the number of samples required to achieve the same precision and confidence levels in the estimation of $\mathbb{E}[f_2(X)g_2(Y)]$. Since we argue above that $G(P_{X,Y})$ is $\Theta(1/K)$ on average, the sample size n required to estimate $\rho_{\mathsf{max}}(X;Y)$ is a factor of $\Theta(K)$ smaller than that required to estimate $P_{X,Y}(x_0, y_0)$. Consequently, the sample ACE algorithm, i.e. the sample version of Algorithm 2 with $k = 1$ mode, which computes an estimate of maximal correlation akin to $\widehat{E}_n[f_2(X)g_2(Y)]$ in its termination condition, also (intuitively) benefits from this saving in the sample complexity compared to plug-in estimation of the entries of the joint pmf $P_{X,Y}$.

■ 4.7 Conclusion and Future Directions

In this chapter, we developed modal decompositions of bivariate distributions, which are well-known in the theory of correspondence analysis and Lancaster distributions (see Theorem 4.3), from the perspective of feature extraction for the purposes of performing unspecified inference tasks. We now briefly delineate the four main contributions in this development. Our first main contribution was to illustrate that maximal correlation functions, which can be obtained from the formulation of maximal correlation in (4.1) and its natural generalizations in Propositions 4.1 and 4.2, can be used as informative feature functions of data. In particular, we argued the utility of maximal correlation functions by:

1. using local information geometric analysis to reveal how maximal correlation functions can meaningfully decompose information contained in categorical data that is observed through a memoryless noise model,

2. expounding how maximal correlation functions can be used as embeddings of categorical bivariate data that capture the salient dependencies of the underlying joint distribution.

The aforementioned local information geometric analysis unveiled a trinity of equivalent representations of a probability distribution under local approximations, namely the distribution itself, an associated information vector, and an associated feature function. Furthermore, we showed that the operation of a channel $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ (which

acts on probability distributions) under local approximations can be equivalently described by the corresponding DTM $B \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ (which acts on information vectors) and the corresponding conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ (which acts on feature functions). Our second main contribution was to characterize the important defining properties of DTMs and conditional expectation operators in Theorems 4.1 and 4.2, as well as in Proposition C.4 in appendix C.3.

Since Propositions 4.1 and 4.2 demonstrated that maximal correlation functions are singular vectors of C, we next considered the problem of estimating maximal correlation functions from training data by adapting known techniques from numerical linear algebra. Specifically, our third main contribution was to reinterpret the well-known orthogonal iteration method for computing singular vectors of the DTM B (see Algorithm 1) from a statistical standpoint. This engendered an iterative procedure for computing maximal correlation functions from data that we called the sample extended ACE algorithm (see Algorithm 2). This algorithm turned out to be a generalization of Breiman and Friedman's ACE algorithm for regression, and we illustrated how it serves as an important feature extraction and dimensionality reduction tool for categorical bivariate data. (We remark that several applications of the extended ACE algorithm and our broader development here to softmax regression, neural networks, and image classification are depicted in [133].)

In order to better understand the sample extended ACE algorithm, we elucidated its close connection to several well-known techniques in the literature such as principal component analysis, canonical correlation analysis, and diffusion maps. Furthermore, our fourth main contribution was our sample complexity analysis for the sample orthogonal iteration method (which is equivalent to the sample extended ACE algorithm). In particular, Propositions 4.5 and 4.6 and Theorems 4.5 and 4.6 portrayed the relationship between sample size and number of modes being estimated, and Propositions 4.7 and 4.8 conveyed some complementary perspectives.

We conclude our discussion on modal decompositions by proposing some avenues for future research. As we mentioned at the ends of subsections 4.6.1 and 4.6.2, proving minimax lower bounds for the estimation problems in these subsections is a viable direction of future work. We additionally suggest some approaches to extend the theory of modal decompositions for bivariate distributions developed in this chapter to univariate or general multivariate distribution settings. Such scenarios are clearly of practical interest. For example, continuing with our running narrative of the "Netflix problem," we may only have access to the frequency at which different movies are streamed by subscribers (without any information regarding which subscribers watched a particular movie) due to privacy concerns. This is obviously a univariate setting. Alternatively, in addition to collecting data about subscribers and movies, we may also have information about a third variable, e.g. whether or not a subscriber watched a movie in its entirety. In this setting, our data samples are 3-tuples since we have three categorical variables of interest. In both these scenarios, we are interested in the unsupervised learning problem of finding a small number of real-valued features of the categorical variables for an

unspecified inference task. Clearly, the discussion in this chapter suggests that modal decompositions of the distributions generating the data are particularly useful for feature extraction.

In the univariate case, there is no canonical definition of "informative" feature functions when the inference task is unknown. (In fact, if there are no dependencies to capture, the term "modal decomposition" is perhaps a misnomer in this scenario.) However, one approach to constructing feature functions in this scenario is to first associate a self-adjoint operator on the Hilbert space of real-valued functions on the given categorical alphabet to the empirical distribution of the data, and then use the orthonormal basis of eigenvectors of this operator as the set of feature functions. There are many avenues to explore within this general framework. For instance, we can construe the empirical distribution of the categorical data as an invariant measure of a reversible Markov chain, which defines a self-adjoint conditional expectation operator. Since there are several reversible Markov chains with the same invariant measure, we can impose additional constraints to yield a unique canonical reversible Markov chain (e.g. choose the random walk on the sparsest weighted undirected graph, cf. [170, Section 9.1], where the vertex set is the categorical alphabet, or assume that the ordered data samples are drawn from a reversible Markov chain so that it can be learned). The spectral decomposition of this Markov chain yields the desired real-valued feature functions.

In the multivariate case with three or more categorical variables, it is reasonable to require the learned feature functions to summarize the salient dependencies between the variables (much like the bivariate case). Since the joint distribution can be written as a higher-order tensor, one approach to obtaining a modal decomposition for the purposes of feature extraction is to decompose some higher-order tensor associated with the joint distribution by employing a generalization of the SVD. There are several such generalizations in multilinear algebra, e.g. the *canonical polyadic decomposition* or the *Tucker decomposition*, and such tensor decompositions have been widely applied in signal processing and machine learning contexts, cf. [254]. Exploring some of the aforementioned ideas could lead to other exciting approaches for feature extraction and dimensionality reduction.

■ 4.8 Digression: The Permutation Channel

In this final section, we study the problem of reliable communication through *permutation channels*. Specifically, we define and analyze a pertinent notion of information capacity for permutation channels. Permutation channels refer to *discrete memoryless channels* (DMCs) followed by random permutation transformations that are applied to the entire blocklength of the output codeword. Such channels can be perceived as models of communication links where packets are not delivered in sequence, and hence, the ordering of the packets does not contain any information. Since all information embedded within the ordering of symbols in a codeword is lost in such channels, information must be transmitted by varying the *types* (empirical distributions or compositions) of

the codewords.

At first glance, the problem of judiciously selecting codeword types for a "good" encoder appears to be closely related to our discussion of modal decompositions. Indeed, a naïve intuition suggests that under a local approximation lens, the "most detectable" codeword types at the receiver are spherical perturbations of some fixed source distribution (e.g. the capacity achieving input distribution of the DMC) along dominant singular vector directions of the DTM defined by the DMC and the source distribution, cf. [132]. Although this intuition originally propelled us to study permutation channels, the analysis in this section demonstrates that the intuition is erroneous, and that the information capacity of permutation channels is achieved through different coding techniques.

Therefore, the analysis in this section digresses from our overarching theme of information contraction. However, it turns out to be peripherally related to the next chapter on broadcasting. Indeed, the achievability proof in subsection 4.8.3 employs the so called *second moment method for TV distance*, which is precisely the technique used to prove that reconstruction is possible on trees when the BSC noise level is below the critical (Kesten-Stigum) threshold [83]. Hence, this section bridges our discussion of modal decompositions with our discussion of broadcasting in chapter 5 (since it is inspired by the former and its techniques are relevant to the latter). The ensuing subsections 4.8.1 and 4.8.2 provide some background literature to motivate our analysis, a formal description of the model, and an outline of our main results.

■ 4.8.1 Related Literature and Motivation

The setting of channel coding with transpositions, where the output codeword undergoes some reordering of its symbols, has been widely studied in both the coding theory and the communication networks communities. For example, in the coding theory literature, one earlier line of work concerned the construction of error-correcting codes that achieve capacity of the *random deletion channel*. The random deletion channel operates on the codeword space by deleting each input codeword symbol independently with some probability $p \in (0, 1)$, and copying it otherwise. As explained in [204, Section I], with sufficiently large alphabet size $q = 2^b$ (where each symbol can be construed as a packet with b bits), "embedding sequence numbers into the transmitted symbols [turns] the deletion channel [into a memoryless] erasure channel." Since the erasure channel setting was well-understood, the intriguing question became to construct (nearly) capacity achieving codes for the deletion channel using sufficiently large packet length b, but without embedding sequence numbers (see [71, 204], and the references therein). In particular, the author of [204] demonstrated that *low density parity check (LDPC) codes* with verification-based decoding formed a computationally tractable coding scheme with these properties. This scheme also tolerated transpositions of packets that were not deleted in the process of transmission. Therefore, it was equivalently a coding scheme for a memoryless erasure channel followed by a random permutation block (albeit with very large alphabet size). Several other coding schemes for erasure permutation channels with

sufficiently large alphabet size have also been developed in the literature; see e.g. [201], which builds upon the key conceptual ideas in [204], and the references therein.

This discussion regarding the random deletion channel has a patent counterpart in the communication networks literature. Indeed, in the context of the well-known *store-and-forward* transmission scheme for networks, packet losses (or deletions) were typically corrected using *Reed-Solomon codes* which assumed that each packet carried a header with a sequence number—see e.g. [292], [94, Section I], and the references therein. Much like in the deletion channel setting, this simplified the error correction problem since packet losses could be treated as erasures. However, "motivated by networks whose topologies change over time, or where several routes with unequal delays are available to transmit the data," the author of [94] illustrated that packet errors and losses could also be corrected using binary codes under a channel model where the impaired or lost packets were randomly permuted, and the packets were not indexed.

Several other aspects of permutation channels have also been investigated in the communication networks literature. For instance, the permutation channel was a useful model for point-to-point communication between a source and a receiver that used a lower level *multipath routed network*. Indeed, since packets (or symbols) could take different paths to the receiver in such a network, they would arrive at the destination out of order due to different delay profiles in the different paths. The authors of [287] established rate-delay tradeoffs for such communication networks, although they neglected to account for packet impairments such as deletions in their analysis for simplicity.

More recently, inspired by packet networks such as mobile ad hoc networks (where the network topology changes over time), and heavily loaded datagram-based networks (where packets are often re-routed for load balancing purposes), the authors of [160], [161], and [159] have considered the general problem of coding in channels where the codeword undergoes a random permutation and is subjected to impairments such as insertions, deletions, substitutions, and erasures. As stated in [160, Section I], the general strategy to communicate across a channel that applies a transformation to its codewords is to "encode the information in an object that is invariant under the given transformation." In the case of the permutation channel, the appropriate codes are the so called *multiset codes* (where the codewords are characterized by their empirical distribution over the underlying alphabet). The existence of certain perfect multiset codes is established in [161], and several other multiset code constructions based on lattices and *Sidon sets* are analyzed in [159].

An alternative motivation for analyzing permutation channels stems from the study of *DNA based storage systems*, cf. [123] and the references therein. The authors of [123] examined the storage capacity of systems where the source is encoded via DNA molecules. These molecules are cached in an unordered fashion akin to the effect of a permutation channel. However, as stated in [159, Section I-B], the model for such systems also differs from our model since the receiver samples the stored codewords with replacement and without errors. We refer readers to the comprehensive bibliography in [159] for other related literature on permutation channels.

As the discussion heretofore reveals, the majority of the literature on permutation channels analyzes its coding theoretic aspects. In contrast, we approach these channels from a purely information theoretic perspective. To our knowledge, there are no known results on the information capacity of the permutation channel model described in the next subsection. (Indeed, while the aforementioned references [71] and [123] have a more information theoretic focus, they analyze different models to ours.) In this section, we will prove some initial results towards a complete understanding of the information capacity of permutation channels. Rather interestingly, our achievability proofs will automatically yield computationally tractable codes for communication through certain permutation channels, thereby rendering the need for developing sophisticated coding schemes for these channels futile when achieving capacity is the sole objective.

■ 4.8.2 Permutation Channel Model

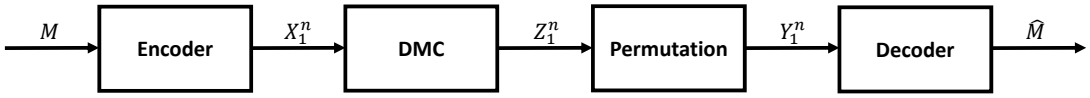


Figure 4.1. Illustration of a communication system with a DMC followed by a random permutation.

We define the point-to-point permutation channel model in analogy with standard information theoretic definitions, cf. [53, Section 7.5]. Let $M \in \mathcal{M} \triangleq \{1, \ldots, |\mathcal{M}|\}$ be a *message* random variable that is drawn uniformly from $\mathcal{M}$, $f_n : \mathcal{M} \to \mathcal{X}^n$ be a (possibly randomized) *encoder*, where $\mathcal{X}$ is the finite input alphabet of the channel and $n \in \mathbb{N}$ is the blocklength, and $g_n : \mathcal{Y}^n \to \mathcal{M}$ be a (possibly randomized) *decoder*, where $\mathcal{Y}$ is the finite output alphabet of the channel. The message M is first encoded into a codeword $X_1^n = f_n(M)$, where each $X_i \in \mathcal{X}$. This codeword is transmitted through a DMC defined by the conditional probability distributions $\{P_{Z|X=x} \in \mathcal{P}_{\mathcal{Y}} : x \in \mathcal{X}\}$ to produce $Z_1^n \in \mathcal{Y}^n$, where $Z_i \in \mathcal{Y}$. The memorylessness property of the DMC implies that:

$$\forall x_1^n \in \mathcal{X}^n, \forall z_1^n \in \mathcal{Y}^n, \;\; P_{Z_1^n|X_1^n}(z_1^n|x_1^n) = \prod_{i=1}^{n} P_{Z|X}(z_i|x_i)\,. \tag{4.127}$$

The noisy codeword Z_1^n is then passed through a *random permutation* transformation to generate $Y_1^n \in \mathcal{Y}^n$, i.e. for a randomly and uniformly chosen permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ (which is independent of everything else), each $Y_i = Z_{\pi(i)}$ for all $i \in \{1, \ldots, n\}$. Finally, the received codeword Y_1^n is decoded to produce an estimate $\hat{M} = g_n(Y_1^n)$ of M. Figure 4.1 illustrates this communication system.

Let the *average probability of error* in this model be:

$$P_{\mathsf{error}}^n \triangleq \mathbb{P}(M \neq \hat{M})\,, \tag{4.128}$$

and the "*rate*" of the encoder-decoder pair (f_n, g_n) be defined as:

$$R \triangleq \frac{\log(|\mathcal{M}|)}{\log(n)} \,. \tag{4.129}$$

So, we can also write $|\mathcal{M}| = n^R$. (Strictly speaking, n^R should be an integer, but we will neglect this detail since it will not affect our results.) We will say that a rate $R \geq 0$ is *achievable* if there is a sequence of encoder-decoder pairs $\{(f_n, g_n) : n \in \mathbb{N}\}$ such that $\lim_{n\to\infty} P_{\mathsf{error}}^n = 0$. Lastly, we operationally define the *permutation channel capacity* as:

$$C_{\mathsf{perm}}(P_{Z|X}) \triangleq \sup\{R \geq 0 : R \text{ is achievable}\} \,. \tag{4.130}$$

It is straightforward to verify that the scaling in (4.129) is indeed $\log(n)$ rather than the standard n. As mentioned earlier, due the random permutation in the model, all information embedded in the ordering within codewords is lost. (In fact, canonical fixed composition codes cannot carry more than one message in this setting.) So, the maximum number of decodable messages is upper bounded by the number of possible empirical distributions of Y_1^n:

$$n^R = |\mathcal{M}| \leq \binom{n + |\mathcal{Y}| - 1}{|\mathcal{Y}| - 1} \leq (n+1)^{|\mathcal{Y}|-1} \tag{4.131}$$

where taking log's and letting $n \to \infty$ yields $C_{\mathsf{perm}}(P_{Z|X}) \leq |\mathcal{Y}| - 1$ (at least non-rigorously). This justifies that $\log(n)$ is the correct scaling in (4.129), i.e. the maximum number of messages that can be reliably communicated is polynomial in the blocklength (rather than exponential).

In the remainder of this section, we will consider two canonical specializations of the aforementioned permutation channel model: the case where the DMC is a BSC, and the case where it is a *binary erasure channel* (BEC). In the context of [160], [161], and [159], the former case corresponds to permutation channels with substitution errors, and the latter case corresponds to permutation channels with erasures (or equivalently, deletions—see [159, Remark 1]). We will establish the permutation channel capacity of the BSC exactly in subsection 4.8.3. In particular, our achievability proof will follow from a binary hypothesis testing result that will be derived using the *second moment method*. Then, we will prove bounds on the permutation channel capacity of the BEC in subsection 4.8.4. Finally, we will conclude this digression into permutation channels and propose future research directions in subsection 4.8.5.

■ 4.8.3 Permutation Channel Capacity of BSC

In this subsection, we let $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$ within the formalism of subsection 4.8.2. Moreover, we let the DMC be a $\mathsf{BSC}(p)$, where $p \in [0, 1]$ is the crossover probability. To derive the permutation channel capacity of BSCs, we first prove a useful auxiliary lemma.

Lemma 4.2 (Testing between Converging Hypotheses). *Fix any $n \in \mathbb{N}$, and any constants $\epsilon_n \in \left(0, \frac{1}{2}\right)$ and $p_n \in \left(0, 1 - \left(1/n^{\frac{1}{2}-\epsilon_n}\right)\right)$ that can depend on n. Consider a binary hypothesis testing problem with hypothesis random variable $H \sim \mathsf{Bernoulli}\left(\frac{1}{2}\right)$ (i.e. uniform Bernoulli prior), and likelihoods $P_{X|H=0} = \mathsf{Bernoulli}(p_n)$ and $P_{X|H=1} = \mathsf{Bernoulli}\left(p_n + \left(1/n^{\frac{1}{2}-\epsilon_n}\right)\right)$ on the alphabet $\mathcal{X} = \{0,1\}$, such that we observe n samples X_1^n that are drawn conditionally i.i.d. given H from the likelihoods:*

$$\begin{aligned} &\textit{Given } H = 0: \quad X_1^n \overset{i.i.d.}{\sim} P_{X|H=0} = \mathsf{Bernoulli}(p_n)\,, \\ &\textit{Given } H = 1: \quad X_1^n \overset{i.i.d.}{\sim} P_{X|H=1} = \mathsf{Bernoulli}\left(p_n + \frac{1}{n^{\frac{1}{2}-\epsilon_n}}\right). \end{aligned}$$

Then, the minimum probability of error corresponding to the ML decoder $\hat{H}^n_{\mathsf{ML}} : \{0,1\}^n \to \{0,1\}$ for H based on X_1^n, $\hat{H}^n_{\mathsf{ML}}(X_1^n)$, satisfies:

$$P^n_{\mathsf{ML}} \triangleq \mathbb{P}\Big(\hat{H}^n_{\mathsf{ML}}(X_1^n) \neq H\Big) \leq \frac{3}{2n^{2\epsilon_n}}\,.$$

This implies that $\lim_{n\to\infty} P^n_{\mathsf{ML}} = 0$ when $\lim_{n\to\infty} n^{\epsilon_n} = +\infty$.

Proof. Let $T_n \in \mathcal{T}_n = \left\{\frac{k}{n} - c_n : k \in [n+1]\right\}$ be the translated arithmetic mean of X_1^n:

$$T_n \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i - c_n$$

where the constant c_n (that can depend on n) will be chosen later. Moreover, for ease of exposition, let T_n^- and T_n^+ be random variables with probability distributions given by the likelihoods $P_{T_n^-} = P_{T_n|H=0}$ and $P_{T_n^+} = P_{T_n|H=1}$, respectively, such that:

$$P_{T_n} = \frac{1}{2}P_{T_n^-} + \frac{1}{2}P_{T_n^+}\,.$$

It is straightforward to verify that T_n is a sufficient statistic of X_1^n for performing inference about H. So, the ML decoder for H based on X_1^n, $\hat{H}^n_{\mathsf{ML}}(X_1^n)$, is a function of T_n without loss of generality, and we denote it as $\hat{H}^n_{\mathsf{ML}} : \mathcal{T}_n \to \{0,1\}$, $\hat{H}^n_{\mathsf{ML}}(T_n)$ (with abuse of notation); in particular, the ML decoder simply thresholds the statistic T_n to detect H.

To upper bound $P^n_{\mathsf{ML}} = \mathbb{P}(\hat{H}^n_{\mathsf{ML}}(T_n) \neq H)$, recall *Le Cam's relation* for the ML decoding probability of error, cf. [278, proof of Theorem 2.2(i)]:

$$P^n_{\mathsf{ML}} = \frac{1}{2}\left(1 - \left\|P_{T_n^+} - P_{T_n^-}\right\|_{\mathsf{TV}}\right). \tag{4.132}$$

Furthermore, recall the so called second moment method lower bound on TV distance—see e.g. [83, Lemma 4.2(iii)]:

$$\left\|P_{T_n^+} - P_{T_n^-}\right\|_{\mathsf{TV}} = \frac{1}{2}\sum_{t\in\mathcal{T}_n}\left|P_{T_n|H}(t|1) - P_{T_n|H}(t|0)\right|$$

$$\geq \frac{1}{4}\sum_{t\in\mathcal{T}_n}\frac{\left(P_{T_n|H}(t|1)-P_{T_n|H}(t|0)\right)^2}{P_{T_n}(t)}$$

$$= \mathsf{LC}_{\frac{1}{2}}(P^+_{T_n}||P^-_{T_n}) \tag{4.133}$$

$$\geq \frac{\left(\mathbb{E}[T_n^+]-\mathbb{E}[T_n^-]\right)^2}{4\,\mathbb{E}[T_n^2]} \tag{4.134}$$

where the first equality follows from (2.4) in chapter 2, the second inequality holds because $|P_{T_n|H}(t|1)-P_{T_n|H}(t|0)| \leq P_{T_n|H}(t|1)+P_{T_n|H}(t|0)$ for all $t\in\mathcal{T}_n$, (4.133) follows from (2.11) and shows that $\|P^+_{T_n}-P^-_{T_n}\|_{\mathsf{TV}}$ is lower bounded by the Vincze-Le Cam distance $\mathsf{LC}_{1/2}(P^+_{T_n}||P^-_{T_n})$, and (4.134) follows from the Cauchy-Schwarz(-Bunyakovsky) inequality.

We now select the constant c_n. Since both $\|P^+_{T_n}-P^-_{T_n}\|_{\mathsf{TV}}$ and the numerator of (4.134) are invariant to the value of c_n, the best bound of the form (4.134) is obtained by minimizing the second moment $\mathbb{E}[T_n^2]$. Thus, c_n is given by:

$$c_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p_n + \frac{1}{2n^{\frac{1}{2}-\epsilon_n}} \tag{4.135}$$

using the facts that $X_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p_n)$ given $H=0$, and $X_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p_n + (1/n^{\frac{1}{2}-\epsilon_n}))$ given $H=1$. This ensures that $\mathbb{E}[T_n]=0$, and the denominator in (4.134) is $\mathbb{E}[T_n^2]=\mathbb{VAR}(T_n)$. We remark that with this choice of c_n, (4.134) can be perceived as a *Hammersley-Chapman-Robbins bound* [44,116], where the Vincze-Le Cam distance in (4.133) replaces the usual χ^2-divergence.

Combining (4.132) and (4.134) yields the following upper bound on the ML decoding probability of error P^n_{ML}:

$$P^n_{\mathsf{ML}} \leq \frac{1}{2}\left(1-\frac{\left(\mathbb{E}[T_n^+]-\mathbb{E}[T_n^-]\right)^2}{4\,\mathbb{E}[T_n^2]}\right) \tag{4.136}$$

which we now compute explicitly. Observe using (4.135) that:

$$\mathbb{E}\left[T_n^+\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i - c_n \middle| H=1\right] = \frac{1}{2n^{\frac{1}{2}-\epsilon_n}},$$

$$\mathbb{E}\left[T_n^-\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i - c_n \middle| H=0\right] = \frac{-1}{2n^{\frac{1}{2}-\epsilon_n}}.$$

Furthermore, using (4.135), we also get:

$$\mathbb{E}\left[T_n^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}X_i-\mathbb{E}[X_i]\right)\left(\sum_{k=1}^{n}X_k-\mathbb{E}[X_k]\right)\right]$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{VAR}(X_1) + \left(\frac{n-1}{n}\right)\mathbb{COV}(X_1, X_2) \\
&= \frac{1}{n}\left(p_n + \frac{1}{2n^{\frac{1}{2}-\epsilon_n}}\right)\left(1 - p_n - \frac{1}{2n^{\frac{1}{2}-\epsilon_n}}\right) \\
&\quad + \left(\frac{n-1}{n}\right)\left(\frac{1}{2}p_n^2 + \frac{1}{2}\left(p_n + \frac{1}{n^{\frac{1}{2}-\epsilon_n}}\right)^2 - \left(p_n + \frac{1}{2n^{\frac{1}{2}-\epsilon_n}}\right)^2\right) \\
&= \frac{p_n(1-p_n)}{n} + \frac{1-2p_n}{2n^{\frac{3}{2}-\epsilon_n}} + \frac{n-2}{4n^{2-2\epsilon_n}} \\
&\leq \frac{1}{4n} + \frac{1}{2n^{\frac{3}{2}-\epsilon_n}} + \frac{1}{4n^{1-2\epsilon_n}}
\end{aligned}
$$

where the final inequality follows from the bounds $p_n(1-p_n) \leq \frac{1}{4}$, $1 - 2p_n \leq 1$, and $n - 2 \leq n$. Plugging in these expressions into (4.136), we have:

$$
\begin{aligned}
P_{\mathsf{ML}}^n &\leq \frac{1}{2}\left(1 - \frac{\left(\frac{1}{n^{\frac{1}{2}-\epsilon_n}}\right)^2}{4\left(\frac{1}{4n} + \frac{1}{2n^{\frac{3}{2}-\epsilon_n}} + \frac{1}{4n^{1-2\epsilon_n}}\right)}\right) \\
&= \frac{1}{2}\left(1 - \frac{1}{1 + \frac{1}{n^{2\epsilon_n}} + \frac{2}{n^{\frac{1}{2}+\epsilon_n}}}\right) \\
&\leq \frac{1}{2}\left(1 - \frac{1}{1 + \frac{3}{n^{2\epsilon_n}}}\right) \\
&\leq \frac{3}{2n^{2\epsilon_n}}
\end{aligned} \tag{4.137}
$$

where the third inequality holds because $\epsilon_n < \frac{1}{2}$. This completes the proof. ■

This lemma illustrates that as long as the difference between the parameters that define $P_{X|H=0}$ and $P_{X|H=1}$ vanishes slower than $\Theta(1/\sqrt{n})$, we can decode the hypothesis H with vanishing probability of error as $n \to \infty$. Intuitively, this holds because the standard deviation of the sufficient statistic T_n is $O(1/\sqrt{n})$ (neglecting ϵ_n). So, as long as the difference between the two parameters is $\omega(1/\sqrt{n})$, it is possible to distinguish between the two hypotheses. We also remark that tighter upper bounds on P_{ML}^n can be obtained using standard exponential concentration of measure inequalities. However, the simpler second moment method approach suffices for our purposes.

We will also require the following useful estimate of the entropy of a binomial distribution from the literature in our converse proof of the permutation channel capacity of BSCs.

Lemma 4.3 (Approximation of Binomial Entropy [4, Equation (7)]). *Given a binomial random variable $X \sim \mathsf{binomial}(n,p)$ with $n \in \mathbb{N}$ and $p \in (0,1)$, we have:*

$$\left| H(X) - \frac{1}{2}\log(2\pi enp(1-p)) \right| \leq \frac{c(p)}{n}$$

for some constant $c(p) \geq 0$, where $H(\cdot)$ denotes the Shannon entropy function.

We next present our first main result of this section, which exploits Lemmata 4.2 and 4.3 to derive a *coding theorem* for BSCs.

Theorem 4.7 (Permutation Channel Capacity of BSC).

$$C_{\mathsf{perm}}(\mathsf{BSC}(p)) = \begin{cases} 1 & , \quad p = 0,1 \\ \frac{1}{2} & , \quad p \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right) \\ 0 & , \quad p = \frac{1}{2} \end{cases} .$$

Proof.

Converse for $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$: Suppose we are given a sequence of encoder-decoder pairs $\{(f_n, g_n) : n \in \mathbb{N}\}$ on message sets of cardinality $|\mathcal{M}| = n^R$ such that $\lim_{n\to\infty} P^n_{\mathsf{error}} = 0$. Consider the Markov chain $M \to X_1^n \to Z_1^n \to Y_1^n \to S_n \triangleq \sum_{i=1}^n Y_i$. Observe that for every $y_1^n \in \{0,1\}^n$ and $m \in \mathcal{M}$:

$$P_{Y_1^n|M}(y_1^n|m) = \binom{n}{|y_1^n|_{\mathsf{H}}}^{-1} \mathbb{P}(|Z_1^n|_{\mathsf{H}} = |y_1^n|_{\mathsf{H}} \,|M = m)$$

where $|\cdot|_{\mathsf{H}}$ denotes the Hamming weight of a binary string. Since $P_{Y_1^n|M}(y_1^n|m)$ depends on y_1^n through $|y_1^n|_{\mathsf{H}}$, the *Fisher-Neyman factorization theorem* implies that S_n is a sufficient statistic of Y_1^n for performing inference about M [150, Theorem 3.6]. Then, following the standard argument from [53, Section 7.9], we have:

$$\begin{aligned} R\log(n) &= H(M) \\ &= H(M|\hat{M}) + I(M;\hat{M}) \\ &\leq \log(2) + P^n_{\mathsf{error}} R\log(n) + I(M;Y_1^n) \\ &= \log(2) + P^n_{\mathsf{error}} R\log(n) + I(M;S_n) \\ &\leq \log(2) + P^n_{\mathsf{error}} R\log(n) + I(X_1^n;S_n) \end{aligned} \tag{4.138}$$

where the first equality holds because M is uniformly distributed, the third line follows from *Fano's inequality* and the DPI [53, Theorems 2.10.1 and 2.8.1], the fourth line holds because S_n is a sufficient statistic, cf. [53, Section 2.9], and the last line also follows from the DPI [53, Theorem 2.8.1] (cf. (3.1) in chapter 3).

We now upper bound $I(X_1^n; S_n)$. Notice that:

$$I(X_1^n;S_n) = H(S_n) - H(S_n|X_1^n)$$

$$\leq \log(n+1) - \sum_{x_1^n \in \{0,1\}^n} P_{X_1^n}(x_1^n) H(S_n | X_1^n = x_1^n) \tag{4.139}$$

where we use the fact that $S_n \in [n+1]$. Given $X_1^n = x_1^n$ for any fixed $x_1^n \in \{0,1\}^n$, $\{Z_i \sim \mathsf{Bernoulli}(p^{1-x_i}(1-p)^{x_i}) : i \in \{1,\ldots,n\}\}$ are mutually independent and $\sum_{i=1}^n Z_i = S_n$ *a.s.* Hence, we have:

$$\begin{aligned} H(S_n | X_1^n = x_1^n) &= H\left(\sum_{i=1}^k A_i + \sum_{j=k+1}^n B_j\right) \\ &\geq \max\left\{ H\left(\sum_{i=1}^k A_i\right), H\left(\sum_{j=k+1}^n B_j\right) \right\} \end{aligned}$$

where $k = |x_1^n|_{\mathsf{H}}$, $A_1^k \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(1-p)$ and $B_{k+1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p)$ are independent, and the inequality follows from [53, Problem 2.14]. (Note that if $k \in \{0, n\}$, then one of the summations above is trivially 0, and its entropy is also 0.) Since $\max\{k, n-k\} \geq \frac{n}{2}$, we can use Lemma 4.3 to get:

$$H(S_n | X_1^n = x_1^n) \geq \frac{1}{2}\log(\pi e p(1-p)n) - \frac{2c(p)}{n}$$

which we can substitute into (4.139) and obtain:

$$I(X_1^n; S_n) \leq \log(n+1) - \frac{1}{2}\log(\pi e p(1-p)n) + \frac{2c(p)}{n}\,. \tag{4.140}$$

Combining (4.138) and (4.140), and dividing by $\log(n)$, yields:

$$R \leq P_{\mathsf{error}}^n R + \frac{\log(2) + \log(n+1)}{\log(n)} - \frac{\log(\pi e p(1-p)n)}{2\log(n)} + \frac{2c(p)}{n\log(n)}$$

where letting $n \to \infty$ produces $R \leq \frac{1}{2}$. Therefore, we have $C_{\mathsf{perm}}(\mathsf{BSC}(p)) \leq \frac{1}{2}$.

Converse for $p = \frac{1}{2}$: The output of the BSC is independent of the input, and $I(X_1^n; S_n) = 0$. So, dividing both sides of (4.138) by $\log(n)$ yields:

$$R \leq \frac{\log(2)}{\log(n)} + P_{\mathsf{error}}^n R$$

where letting $n \to \infty$ produces $R \leq 0$. Therefore, we have $C_{\mathsf{perm}}(\mathsf{BSC}(\frac{1}{2})) = 0$.

Converse for $p \in \{0,1\}$: Starting from (4.139), we get the inequality $I(X_1^n; S_n) \leq \log(n+1)$, since $H(S_n | X_1^n) = 0$. As before, combining (4.138) and this inequality, and dividing by $\log(n)$, yields:

$$R \leq \frac{\log(2)}{\log(n)} + P_{\mathsf{error}}^n R + \frac{\log(n+1)}{\log(n)}$$

where letting $n \to \infty$ produces $R \leq 1$. Therefore, we have $C_{\mathsf{perm}}(\mathsf{BSC}(p)) \leq 1$.

Achievability for $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$: For any $\epsilon \in (0, \frac{1}{2})$, suppose we have:

1. $|\mathcal{M}| = n^{\frac{1}{2}-\epsilon}$ messages,
2. a randomized encoder $f_n : \mathcal{M} \to \{0,1\}^n$ such that:
$$\forall m \in \mathcal{M}, \;\; f_n(m) = X_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}\left(\frac{m}{n^{\frac{1}{2}-\epsilon}}\right),$$
3. an ML decoder $g_n : \{0,1\}^n \to \mathcal{M}$ such that:
$$\forall y_1^n \in \{0,1\}^n, \;\; g_n(y_1^n) = \arg\max_{m \in \mathcal{M}} P_{Y_1^n|M}(y_1^n|m)$$
where the tie-breaking rule (when there are many maximizers) does not affect P_{error}^n.

This completely specifies the communication system model in subsection 4.8.2. We now analyze the average probability of error for this simple encoding and decoding scheme.

Let us condition on the event $M = m \in \mathcal{M}$. Then, $X_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(m/n^{\frac{1}{2}-\epsilon})$, and $Z_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p * (m/n^{\frac{1}{2}-\epsilon}))$ since the BSC is memoryless, where $r * s \triangleq r(1-s) + s(1-r)$ denotes the convolution of $r, s \in [0,1]$. Moreover, $Y_1^n \overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}(p*(m/n^{\frac{1}{2}-\epsilon}))$ because it is the output of passing Z_1^n through a random permutation. The conditional probability that our ML decoder makes an error is upper bounded by:

$$\begin{aligned}
\mathbb{P}(\hat{M} \neq M | M = m) &= \mathbb{P}(g_n(Y_1^n) \neq m | M = m) \\
&\leq \mathbb{P}\Big(\exists\, i \in \mathcal{M}\backslash\{m\},\, P_{Y_1^n|M}(Y_1^n|i) \geq P_{Y_1^n|M}(Y_1^n|m) \Big| M = m\Big) \\
&\leq \sum_{i \in \mathcal{M}\backslash\{m\}} \mathbb{P}\Big(P_{Y_1^n|M}(Y_1^n|i) \geq P_{Y_1^n|M}(Y_1^n|m) \Big| M = m\Big) \qquad (4.141)
\end{aligned}$$

where the second inequality is an upper bound because we regard the equality case, $P_{Y_1^n|M}(Y_1^n|i) = P_{Y_1^n|M}(Y_1^n|m)$ for $i \neq m$, as an error (even though the ML decoder may return the correct message in this scenario), and the third inequality follows from the union bound. To show that this upper bound vanishes, for any message $i \neq m$, consider a binary hypothesis test with likelihoods:

$$\begin{aligned}
\text{Given } H = 0: \;\; Y_1^n &\overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}\left(p * \frac{m}{n^{\frac{1}{2}-\epsilon}}\right) \\
\text{Given } H = 1: \;\; Y_1^n &\overset{\text{i.i.d.}}{\sim} \mathsf{Bernoulli}\left(p * \frac{i}{n^{\frac{1}{2}-\epsilon}}\right)
\end{aligned}$$

where the hypotheses $H = 0$ and $H = 1$ correspond to the messages $M = m$ and $M = i$, respectively. The magnitude of the difference between the two Bernoulli parameters is:

$$\left| p * \frac{m}{n^{\frac{1}{2}-\epsilon}} - p * \frac{i}{n^{\frac{1}{2}-\epsilon}} \right| = \frac{|1-2p||m-i|}{n^{\frac{1}{2}-\epsilon}} = \frac{1}{n^{\frac{1}{2}-\epsilon_n}}$$

where (for sufficiently large n depending on p):

$$\epsilon_n = \epsilon + \frac{\log(|1-2p||m-i|)}{\log(n)} \in \left(0, \frac{1}{2}\right).$$

Using Lemma 4.2, if $H \sim \mathsf{Bernoulli}(\frac{1}{2})$, then $P_{\mathsf{ML}}^n = \mathbb{P}(\hat{H}_{\mathsf{ML}}^n(Y_1^n) \neq H)$ satisfies:

$$P_{\mathsf{ML}}^n = \frac{1}{2}\mathbb{P}\Big(\hat{H}_{\mathsf{ML}}^n(Y_1^n) = 1 \Big| M = m\Big) + \frac{1}{2}\mathbb{P}\Big(\hat{H}_{\mathsf{ML}}^n(Y_1^n) = 0 \Big| M = i\Big) \leq \frac{3}{2n^{2\epsilon_n}}$$

which implies that the false-alarm probability satisfies:

$$\mathbb{P}\Big(\hat{H}_{\mathsf{ML}}^n(Y_1^n) = 1 \Big| M = m\Big) = \mathbb{P}\Big(P_{Y_1^n|M}(Y_1^n|i) \geq P_{Y_1^n|M}(Y_1^n|m) \Big| M = m\Big) \leq \frac{3}{n^{2\epsilon_n}} \tag{4.142}$$

where the equality follows from breaking ties, i.e. cases where we get $P_{Y_1^n|M}(Y_1^n|i) = P_{Y_1^n|M}(Y_1^n|m)$, by assigning $\hat{H}_{\mathsf{ML}}^n(Y_1^n) = 1$ (which does not affect the analysis of P_{ML}^n in Lemma 4.2).

Combining (4.141) and (4.142) yields:

$$\begin{aligned}
\mathbb{P}(\hat{M} \neq M | M = m) &\leq \sum_{i \in \mathcal{M}\setminus\{m\}} \frac{3}{n^{2\epsilon_n}} \\
&= 3 \sum_{i \in \mathcal{M}\setminus\{m\}} \left(\frac{1}{n^{\epsilon + \frac{\log(|1-2p||m-i|)}{\log(n)}}}\right)^2 \\
&= \frac{3}{(1-2p)^2 n^{2\epsilon}} \sum_{i \in \mathcal{M}\setminus\{m\}} \frac{1}{(m-i)^2} \\
&\leq \frac{3}{(1-2p)^2 n^{2\epsilon}} \sum_{k=1}^{\infty} \frac{2}{k^2} \\
&= \frac{\pi^2}{(1-2p)^2 n^{2\epsilon}}
\end{aligned} \tag{4.143}$$

where the fourth inequality holds because $k = m - i$ ranges over a subset of all non-zero integers. Finally, taking expectations with respect to M in (4.143) produces:

$$P_{\mathsf{error}}^n \leq \frac{\pi^2}{(1-2p)^2 n^{2\epsilon}} \tag{4.144}$$

which implies that $\lim_{n\to\infty} P_{\mathsf{error}}^n = 0$. Therefore, the rate $R = \frac{1}{2} - \epsilon$ is achievable for every $\epsilon \in (0, \frac{1}{2})$, and $C_{\mathsf{perm}}(\mathsf{BSC}(p)) \geq \frac{1}{2}$.

Achievability for $p \in \{0, 1\}$: Assume without loss of generality that $p = 0$ since a similar argument holds for $p = 1$. In this case, the BSC is just the deterministic identity channel, and we can use the obvious encoder-decoder pair:

1. $|\mathcal{M}| = n + 1$ messages,

2. encoder $f_n : \mathcal{M} \to \{0,1\}^n$ such that:

$$\forall m \in \mathcal{M}, \;\; f_n(m) = (\underbrace{1,\ldots,1}_{m-1\,1\text{'s}}, \underbrace{0,\ldots,0}_{n-m+1\,0\text{'s}}),$$

3. decoder $g_n : \{0,1\}^n \to \mathcal{M}$ such that:

$$\forall y_1^n \in \{0,1\}^n, \;\; g_n(y_1^n) = 1 + \sum_{i=1}^{n} y_i$$

which achieves $P_{\mathsf{error}}^n = 0$. Hence, the rate:

$$R = \lim_{n\to\infty} \frac{\log(n+1)}{\log(n)} = 1$$

is achievable, and $C_{\mathsf{perm}}(\mathsf{BSC}(p)) \geq 1$. ■

We make a few pertinent remarks regarding Theorem 4.7. Firstly, in the $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ regime, the randomized encoder and ML decoder presented in the achievability proof constitute a computationally tractable coding scheme. Indeed, unlike traditional channel coding, the ML decoder requires at most $O(n)$ likelihood ratio tests in our setup, which means that the decoder operates in polynomial time in n. More precisely, the interval $[0, n]$ can be partitioned into sub-intervals so that each sub-interval is the decoding region for a message in $\mathcal{M}$. The ML decoder can be shown to generate the message $\hat{M}$ that corresponds to the decoding region that contains the sufficient statistic S_n. Therefore, communication via permutation channels does not require the development of sophisticated coding schemes to achieve capacity. Furthermore, our achievability proof also implies the existence of a good deterministic code using the *probabilistic method*.

Secondly, although we have presented Theorem 4.7 under an average probability of error criterion, our proof establishes the permutation channel capacity of a BSC under a maximal probability of error criterion as well; see e.g. (4.143). More generally, the permutation channel capacity of a DMC remains the same under a maximal probability of error criterion. This follows from a straightforward *expurgation* argument similar to [230, Theorem 18.3, Corollary 18.1] or [53, Section 7.7, p.204].

Thirdly, in the $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ regime, we intuitively expect the rate of decay of P_{error}^n to be dominated by the rate of decay of the probability of error in distinguishing between two consecutive messages. Although we do not derive precise rates in this section, Lemma 4.2 and (4.144) indicate that this intuition is accurate.

Fourthly, the proof of Lemma 4.2 (and the discussion following it) portrays that the distinguishability between two consecutive messages is determined by a careful comparison of the difference between means and the variance. This suggests that the *central limit theorem* (CLT) can be used to obtain (at least informally) the $|\mathcal{M}| \approx \sqrt{n}$ scaling in the $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ regime. The CLT is in fact implicitly used in our converse

proof when we apply Lemma 4.3, because estimates for the entropy of a binomial distribution can be obtained using the CLT.

Lastly, Theorem 4.7 illustrates a few somewhat surprising facts about permutation channel capacity. While traditional channel capacity is convex as a function of the channel (with fixed dimensions), permutation channel capacity is clearly non-convex and discontinuous as a function the channel. Moreover, for the most part, the permutation channel capacity of a BSC does not depend on p. This is because the scaling (with n) of the difference between the Bernoulli parameters of two encoded messages does not change after passing through the memoryless BSC. However, (4.144) suggests that p does affect the rate of decay of P_{error}^n.

■ 4.8.4 Permutation Channel Capacity of BEC

In this subsection, we let $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,\mathsf{e}\}$, where e denotes the erasure symbol, within the formalism of subsection 4.8.2. Moreover, we let the DMC be a BEC, which is defined by the conditional distributions:

$$\forall z \in \mathcal{Y}, \forall x \in \mathcal{X}, \;\; P_{Z|X}(z|x) = \begin{cases} 1-\delta & , \quad z = x \\ 0 & , \quad z = 1-x \\ \delta & , \quad z = \mathsf{e} \end{cases} \tag{4.145}$$

where $\delta \in [0,1]$ is the erasure probability. We denote such a BEC as $\mathsf{BEC}(\delta)$. The ensuing theorem establishes bounds on the permutation channel capacity of BECs.

Theorem 4.8 (Bounds on Permutation Channel Capacity of BEC). *For $\delta \in (0,1)$, we have:*

$$\frac{1}{2} \leq C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) \leq 1 \, .$$

Furthermore, the extremal permutation channel capacities are $C_{\mathsf{perm}}(\mathsf{BEC}(0)) = 1$ and $C_{\mathsf{perm}}(\mathsf{BEC}(1)) = 0$.

Proof.

Converse for $\delta \in (0,1)$: As in the converse proof for BSCs, suppose we are given a sequence of encoder-decoder pairs $\{(f_n, g_n) : n \in \mathbb{N}\}$ on message sets of cardinality $|\mathcal{M}| = n^R$ such that $\lim_{n\to\infty} P_{\mathsf{error}}^n = 0$. For every $y \in \{0,1,\mathsf{e}\}$, define the function $S_n^y : \{0,1,\mathsf{e}\}^n \to [n+1]$:

$$S_n^y(y_1^n) \triangleq \sum_{i=1}^{n} \mathbb{1}\{y_i = y\} \, .$$

Consider the Markov chain $M \to X_1^n \to Z_1^n \to Y_1^n \to (S_n^1(Y_1^n), S_n^{\mathsf{e}}(Y_1^n))$. Observe that for every $y_1^n \in \{0,1,\mathsf{e}\}^n$ and $m \in \mathcal{M}$:

$$P_{Y_1^n|M}(y_1^n|m) = \frac{\mathbb{P}(S_n^1(Z_1^n) = S_n^1(y_1^n), S_n^{\mathsf{e}}(Z_1^n) = S_n^{\mathsf{e}}(y_1^n)|M=m)}{\binom{n}{S_n^1(y_1^n), S_n^{\mathsf{e}}(y_1^n), S_n^0(y_1^n)}}$$

where the term in the denominator is a multinomial coefficient, and $S_n^0(y_1^n) = n - S_n^1(y_1^n) - S_n^{\mathsf{e}}(y_1^n)$. As before, since $P_{Y_1^n|M}(y_1^n|m)$ depends on y_1^n through $S_n^1(y_1^n)$ and $S_n^{\mathsf{e}}(y_1^n)$, the Fisher-Neyman factorization theorem implies that $(S_n^1(Y_1^n), S_n^{\mathsf{e}}(Y_1^n))$ is a sufficient statistic of Y_1^n for performing inference about M. Then, following the standard Fano's inequality argument (see the derivation of (4.138)), we get:

$$R\log(n) \leq \log(2) + P_{\mathsf{error}}^n R\log(n) + I(X_1^n; S_n^1, S_n^{\mathsf{e}}) \tag{4.146}$$

where we let $S_n^1 = S_n^1(Y_1^n)$ and $S_n^{\mathsf{e}} = S_n^{\mathsf{e}}(Y_1^n)$ (with abuse of notation). To upper bound $I(X_1^n; S_n^1, S_n^{\mathsf{e}})$, notice that:

$$\begin{aligned} I(X_1^n; S_n^1, S_n^{\mathsf{e}}) &= I(X_1^n; S_n^{\mathsf{e}}) + I(X_1^n; S_n^1|S_n^{\mathsf{e}}) \\ &= H(S_n^1|S_n^{\mathsf{e}}) - H(S_n^1|X_1^n, S_n^{\mathsf{e}}) && (4.147) \\ &\leq \log(n+1) && (4.148) \end{aligned}$$

where the first line follows from the chain rule, the second line holds because the number of erasures, $S_n^{\mathsf{e}} = \sum_{i=1}^n \mathbb{1}\{Z_i = \mathsf{e}\}$ *a.s.*, is independent of X_1^n, and the third line uses the facts that $S_n^1 \in [n+1]$ and $H(S_n^1|X_1^n, S_n^{\mathsf{e}}) \geq 0$. Combining (4.146) and (4.148), and dividing by $\log(n)$, yields:

$$R \leq \frac{\log(2)}{\log(n)} + P_{\mathsf{error}}^n R + \frac{\log(n+1)}{\log(n)}$$

where letting $n \to \infty$ produces $R \leq 1$. Therefore, we have $C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) \leq 1$.

Case $\delta = 1$: In this case, the BEC erases all its input symbols so that $S_n^1 = 0$ and $S_n^{\mathsf{e}} = n$ *a.s.* This implies that $I(X_1^n; S_n^1, S_n^{\mathsf{e}}) = 0$. Hence, dividing both sides of (4.146) by $\log(n)$ yields:

$$R \leq \frac{\log(2)}{\log(n)} + P_{\mathsf{error}}^n R$$

where letting $n \to \infty$ produces $R \leq 0$. Therefore, we have $C_{\mathsf{perm}}(\mathsf{BEC}(1)) = 0$.

Case $\delta = 0$: In this case, the BEC is just the deterministic identity channel $\mathsf{BSC}(0)$. Hence, $C_{\mathsf{perm}}(\mathsf{BEC}(0)) = C_{\mathsf{perm}}(\mathsf{BSC}(0)) = 1$ using Theorem 4.7.

Achievability for $\delta \in (0,1)$: For the achievability proof, we employ a useful representation of BSCs using BECs. Observe that a $\mathsf{BSC}(\frac{\delta}{2})$ can be equivalently construed as a channel that copies its input bit with probability $1-\delta$, and generates a completely independent $\mathsf{Bernoulli}(\frac{1}{2})$ output bit with probability δ. Indeed, the decomposition (D.15) of the BSC's stochastic transition probability matrix in appendix D.3 demonstrates this equivalence. A consequence of this idea is that a $\mathsf{BSC}(\frac{\delta}{2})$ is statistically equivalent to a $\mathsf{BEC}(\delta)$ followed by a channel that outputs an independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit for every input erasure symbol, and copies all other input symbols.

Thus, for our $\mathsf{BEC}(\delta)$ permutation channel, let us use the randomized encoder from the achievability proof for a $\mathsf{BSC}(\frac{\delta}{2})$ with $\frac{\delta}{2} \in (0, \frac{1}{2})$ (in subsection 4.8.3). Furthermore, let us use a randomized decoder which first generates independent $\mathsf{Bernoulli}(\frac{1}{2})$ bits

to replace every erasure symbol in Y_1^n, and then applies the ML decoder from the achievability proof for a $\mathsf{BSC}(\frac{\delta}{2})$ to the resulting codeword (which belongs to $\{0,1\}^n$). By our previous discussion, it is straightforward to verify that the P_{error}^n for this encoder-decoder pair under the $\mathsf{BEC}(\delta)$ model is equal to the P_{error}^n analyzed in the achievability proof for a $\mathsf{BSC}(\frac{\delta}{2})$. (We omit the details of this equivalence for brevity.) This portrays that $C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) \geq \frac{1}{2}$ using the achievability result of Theorem 4.7. ■

We conjecture that $C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) = \frac{1}{2}$ in the $\delta \in (0,1)$ regime, i.e. the achievability result is tight, and the permutation channel capacities of the BSC and BEC are equal (in the non-trivial regimes of their parameters). Our converse bound, $C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) \leq 1$, is intuitively trivial, since there are only $n+1$ distinct empirical distributions of codewords in $\{0,1\}^n$ (which non-rigorously shows the upper bound on capacity). So, we believe that this bound can be tightened.

To elucidate the difficulty in improving this bound, consider the expression in (4.147), which along with the fact that $S_n^1 \in [n+1]$, produces:

$$I(X_1^n; S_n^1, S_n^{\mathsf{e}}) \leq \log(n+1) - H(S_n^1 | X_1^n, S_n^{\mathsf{e}}) \,. \tag{4.149}$$

As in the proof of the converse for the BSC, if we can lower bound $H(S_n^1|X_1^n, S_n^{\mathsf{e}})$ by:

$$H(S_n^1 | X_1^n, S_n^{\mathsf{e}}) \geq \frac{1}{2}\log(n) + o(\log(n)) \tag{4.150}$$

then combining (4.146), (4.149), and (4.150) will yield the desired bound $C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) \leq \frac{1}{2}$. It is straightforward to verify that given $X_1^n = x_1^n \in \{0,1\}^n$ such that $S_n^1(x_1^n) = m \in [n+1]$ and $S_n^{\mathsf{e}} = k \in [n+1]$, S_n^1 has a hypergeometric distribution:

$$P_{S_n^1 | X_1^n, S_n^{\mathsf{e}}}(r | x_1^n, k) = \frac{\binom{m}{r}\binom{n-m}{n-k-r}}{\binom{n}{k}} \tag{4.151}$$

for every $\max\{0, m-k\} \leq r \leq \min\{m, n-k\}$. Furthermore, $S_n^{\mathsf{e}} \sim \mathsf{binomial}(n, \delta)$ because $\{\mathbb{1}\{Z_i = \mathsf{e}\} : i \in \{1, \ldots, n\}\}$ are i.i.d. $\mathsf{Bernoulli}(\delta)$ for a memoryless $\mathsf{BEC}(\delta)$, and S_n^{e} is independent of X_1^n (as mentioned earlier). However, it is unclear how to use these facts to find an estimate of the form (4.150) since we cannot immediately apply a CLT based argument (as we might have done to obtain Lemma 4.3).

■ 4.8.5 Conclusion and Future Directions

In closing, we first briefly reiterate our main contributions in this section. Propelled by existing literature on coding for permutation channels, we formulated the information theoretic notion of permutation channel capacity for the problem of reliably communicating through a DMC followed by a random permutation transformation. We then proved that the permutation channel capacity of a BSC is $C_{\mathsf{perm}}(\mathsf{BSC}(p)) = \frac{1}{2}$ for

$p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ in Theorem 4.7. Furthermore, we derived bounds on the permutation channel capacity of a BEC, $\frac{1}{2} \leq C_{\mathsf{perm}}(\mathsf{BEC}(\delta)) \leq 1$, for $\delta \in (0, 1)$ in Theorem 4.8.

We next propose some directions for future research. Firstly, our proof technique for Theorem 4.7 can be extended to establish the permutation channel capacity of DMCs with entry-wise strictly positive stochastic transition probability matrices. In particular, this will entail employing multivariate versions of the results used (either implicitly or explicitly) in our argument such as the second moment method bound in (4.134) and the CLT. Secondly, the exact permutation channel capacity of the BEC should be determined. As conveyed in the previous subsection, this will presumably involve a more careful analysis of the converse proof. Thirdly, our ultimate objective is to establish the permutation channel capacity of general DMCs (whose row stochastic matrices can have zero entries).[71] Evidently, achieving this goal will first require us to completely resolve the permutation channel capacity of BECs. Finally, there are several other open problems that parallel aspects of classical information theoretic development such as:

1. Finding tight bounds on the probability of error (akin to *error exponent analysis*), cf. [95, Chapter 5].

2. Developing *strong converse* results, cf. [230, Section 22.1], [95, Theorem 5.8.5].

3. Establishing *exact asymptotics* for the maximum achievable value of $|\mathcal{M}|$ (akin to "finite blocklength analysis"), cf. [228], [269, Chapter II.4], and the references therein.

4. Extending the permutation channel model by replacing DMCs with other kinds of memoryless channels or networks, e.g. AWGN channels or multiple-access channels (MACs), and by using more general and realistic algebraic operations that are applied to the output codewords, e.g. random permutations that belong to subgroups of the symmetric group.

We remark that the extension of permutation channel models to network settings can lead to interesting algebraic considerations. To briefly elaborate on this, consider a simple single-hop network model, the memoryless *noisy k-user binary adder MAC* with $k \in \mathbb{N} \backslash \{1\}$, cf. [43], followed by a random permutation block; we do not formally define this model here for brevity.[72] Our intuition from subsection 4.8.3 suggests that the (appropriate notion of) *permutation channel capacity region* of a noisy k-user binary adder MAC is achieved when each user encodes their independent message with an

[71] We remark that establishing the permutation channel capacities of DMCs with stochastic matrices that have zero entries appears to be far more intractable, because zero entries introduce a combinatorial flavor to the problem similar to (but not exactly the same as) the zero error capacity. It is well-known that calculating the zero error capacity of channels is very difficult, cf. [250], and the best known approaches use semidefinite programming relaxations such as the Lovász ϑ function. So, completely resolving the permutation channel capacity question for DMCs may require new insights.

[72] We refer readers to [81, Chapter 4] for a classical treatment of MACs.

i.i.d. Bernoulli string, where different users use disjoint sets of Bernoulli parameters to encode their messages. Since the output codeword of the memoryless noisy k-user binary adder MAC is randomly permuted, the "basic decoding problem" is to reconstruct the different users' Bernoulli parameters based on the empirical distribution of the output codeword of the memoryless noisy k-user binary adder MAC. As a "toy version" of this basic decoding problem, consider the memoryless *noiseless* k-user binary adder MAC:

$$W = X_1 + \cdots + X_k \tag{4.152}$$

where $\{X_i \sim \mathsf{Bernoulli}(p_i) : i \in \{1, \ldots, k\}, p_i \in (0, 1)\}$ are mutually independent user input random variables, and $W \in [k+1]$ is the output random variable. Furthermore, suppose we are in the "infinite blocklength" regime where the empirical distributions of the users' input codewords and the output codeword are equal to the true distributions. Then, the basic decoding problem for the memoryless noiseless k-user binary adder MAC corresponds to reconstructing the parameters $\{p_i : i \in \{1, \ldots, k\}\}$ based on the pmf P_W of W. The next proposition provides a complete characterization of the valid output pmfs P_W of noiseless k-user binary adder MACs, as well as a simple formula to reconstruct $\{p_i : i \in \{1, \ldots, k\}\}$ from such valid output pmfs.

Proposition 4.9 (Noiseless Binary Adder MAC Output Distribution). *Suppose $W \in [k+1]$ is a discrete random variable with pmf $P_W = (P_W(0), \ldots, P_W(k)) > 0$ (element-wise). Then, $W = X_1 + \cdots + X_k$ for some independent random variables $\{X_i \sim \mathsf{Bernoulli}(p_i) : p_i \in (0, 1), i \in \{1, \ldots, k\}\}$ if and only if the probability generating function of W, $G_W : \mathbb{C} \to \mathbb{C}$:*

$$\forall z \in \mathbb{C}, \;\; G_W(z) \triangleq \mathbb{E}\Big[z^W\Big]$$

has all real roots. Furthermore, these real roots $z_1, \ldots, z_k \in \mathbb{R}$ of G_W (counted with multiplicity) determine the parameters $p_1, \ldots, p_k$ via the relations:

$$\forall i \in \{1, \ldots, k\}, \;\; z_i = \frac{p_i - 1}{p_i}$$

up to permutations of the indices.

Proof. Suppose $W = X_1 + \cdots + X_k$, where $X_i \sim \mathsf{Bernoulli}(p_i)$ with $p_i \in (0, 1)$ are independent. Then, we have:

$$\forall z \in \mathbb{C}, \;\; G_W(z) = \prod_{i=1}^{k} G_{X_i}(z) = \left(\prod_{j=1}^{k} p_j\right) \prod_{i=1}^{k} \left(z + \frac{1 - p_i}{p_i}\right)$$

which implies that G_W has all real roots: $z_i = (p_i - 1)/p_i$ for $i \in \{1, \ldots, k\}$.

Suppose G_W has all real roots: $z_1, \ldots, z_k \in \mathbb{R}$. Using the fundamental theorem of algebra, we have:

$$\forall z \in \mathbb{C}, \;\; G_W(z) = \sum_{i=0}^{k} P_W(i) z^i = \alpha \prod_{i=1}^{k} (z - z_i) \,.$$

Furthermore, $z_1, \ldots, z_k < 0$ by *Descartes' rule of signs*. (Note that none of the roots are zero because $P_W(0) > 0$.) So, we may define $p_1, \ldots, p_k \in (0,1)$ via the relations $z_i = (p_i - 1)/p_i$ for $i \in \{1, \ldots, k\}$. This yields:

$$G_W(z) = \alpha \left(\prod_{j=1}^{k} p_j \right)^{-1} \prod_{i=1}^{k} (1 - p_i + p_i z) = \prod_{i=1}^{k} G_{X_i}(z)$$

where $\alpha = \prod_{j=1}^{k} p_j$ because $G_W(1) = 1$, and we define independent random variables $X_i \sim \mathsf{Bernoulli}(p_i)$ for $i \in \{1, \ldots, k\}$ in the second equality. Hence, $W = X_1 + \cdots + X_k$ as required. This completes the proof. ■

Proposition 4.9 can be easily extended to include the edge cases where some $p_i \in \{0, 1\}$. Moreover, it generalizes the result in [6, Lemma 1], which proves the $k = 2$ case using a somewhat different approach. Lastly, we note that Proposition 4.9 illustrates how the basic decoding problem for a memoryless noiseless k-user binary adder MAC followed by a random permutation block can be solved in the "infinite blocklength" regime. Therefore, algebraic ideas akin to Proposition 4.9 are also (intuitively) useful to establish (the achievability part of) permutation channel capacity regions of memoryless noisy k-user binary adder MACs.

■ 4.9 Bibliographical Notes

Chapter 4 and appendix C are based primarily on portions of the manuscript [133]. These portions of the manuscript were published in part at the Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing 2015 [182]. On the other hand, section 4.8 of chapter 4 is based primarily on the conference paper [181]. There have been many extensions of the work in [182] by the author and his collaborators, e.g. [134, 135, 233], that are not included in this thesis for reasons of brevity and relevance. We next provide an integrated overview of the body of work on modal decompositions and their applications that we and our coauthors have produced.

The local information geometric analysis of section 4.3 originated in the context of network information theory in [33], which mainly analyzed discrete *degraded broadcast channels*, and in the context of *compound channels* in [2]. The authors of [139, 140] then used this local geometric analysis to study discrete multi-terminal networks such as general broadcast channels. Their key insight was that "single letterization" is easy to establish for so called *linear information coupling problems* (which appropriately model channels and networks under local approximations). These results were further extended from a single-hop to a discrete multi-hop network setting in [136] to provide insights on how to transmit private and common messages in such networks. In a similar vein, Gaussian broadcast channels were analyzed in [3] using a classical modal decomposition of bivariate jointly Gaussian distributions known as *Mehler's decomposition*, cf. [197].[73]

[73] Although the authors of [3] suspected that the modal decomposition of bivariate jointly Gaussian

In particular, the authors of [3] used a coordinate system of *Hermite polynomials* to demonstrate that non-Gaussian codes can achieve higher rates than Gaussian codes for various Gaussian networks, thereby disproving the strong Shamai-Laroia conjecture for the *Gaussian inter-symbol interference channel*.

The focus of this line of work then shifted to addressing problems in statistical inference and learning. For example, one of our transitioning papers was [132], which used local information geometric analysis to study inference in contexts where data was observed through a permutation channel. This paper showed the utility of maximal correlation functions as features in certain image processing and graphical model contexts. It also illustrated that the problem of reliable communication through a permutation channel resembled the problem of reliable communication through a multiple-input and multiple-output (MIMO) additive Gaussian noise channel. However, it turned out that modal decompositions did not yield an effective coding scheme to reliably communicate across permutation channels, and our more recent paper [181] elucidated the "right" way to code in such channels.

The first paper that developed local information geometry ideas explicitly in the context of statistical inference and learning was [182]. In this paper, we proposed that feature extraction can be performed using modal decompositions of bivariate distributions, elaborated on how an extension of the classical ACE algorithm, cf. [35], can be used to efficiently compute these decompositions from training data, and did some basic sample complexity analysis. Around the same time, the problem of feature extraction in hidden Markov model settings was addressed using local geometric analysis in [141]. Then, we analytically established modal decompositions of bivariate distributions where $P_{Y|X}$ was a *natural exponential family with quadratic variance function*, cf. [198, 208, 209], P_X was its corresponding *conjugate prior*, and all moments of X and Y were finite in [190, 191]. These results illustrated that various well-known orthogonal polynomial families were the singular vectors of the conditional expectation operators corresponding to jointly Gaussian, gamma-Poisson, and beta-binomial source-channel pairs.[74] Furthermore, they generalized Mehler's decomposition for jointly Gaussian source-channel pairs (which was used in [3]). At this point, it is worth mentioning that part of the work in [132] and the work in [189–191] formed the author's master's thesis [180].

The next evolution in this story was our formulation of the "universal" feature extraction problem in [135]. This formulation propelled a slew of other equivalent formulations such as local versions of Tishby, Pereira, and Bialek's *information bottleneck* [273] (which captures the notion of approximate minimal sufficiency) and Wyner's *common information* (which meaningfully measures common randomness between two random variables in a certain sense) [291]. These formulations were all solved using modal de-

distributions they used was known in the literature, e.g. in the classical theory of the *Ornstein-Uhlenbeck process*, they did not recognize it as Mehler's decomposition.

[74] Admittedly, these results were more aligned with the spirit of Lancaster's work [165, 166] rather than the more pertinent setting of correspondence analysis [24, 125].

compositions in [133, 134]. Furthermore, Gaussian versions of the aforementioned finite alphabet formulations were also studied in [133, 137].

In closing, it is worth mentioning two other applications of the local geometric analysis that pervades the aforementioned work. In [138], the authors propose a variant of the ACE algorithm for feature extraction in the setting where there are multiple random variables (rather than two). In particular, they set up the feature extraction problem as a maximization of the loss of Watanabe's *total correlation*, cf. [288], under local approximations. On a separate front, in [233], we utilize local information theoretic analysis to develop algorithms for probabilistic clustering. One of our main contributions is an *alternating maximization* algorithm for clustering (inspired by [219]) that maximizes local common information, where one projection step exploits the extended ACE algorithm (see Algorithm 2) and the other projection step relies on a linear program.

Chapter 5

Information Contraction in Networks: Broadcasting on DAGs

THUS far, we have studied SDPIs and contraction coefficients from various perspectives. Indeed, chapter 2 has analyzed contraction coefficients of source-channel pairs, chapter 3 has generalized SDPIs and contraction coefficients of channels for KL divergence, and chapter 4 has examined the elegant geometry of maximal correlation which pertains to the contraction of χ^2-divergence. However, all of these perspectives have focused on the contraction of information along a point-to-point channel or Markov chain. In contrast, in this chapter, we study the contraction of information within the broader class of Bayesian networks. Since tight bounds on the contraction coefficients for KL divergence and TV distance in Bayesian network settings have already been developed by Evans and Schulman [85] and Polyanskiy and Wu [231], respectively, we do not establish any general SDPIs for Bayesian networks here. Instead, we analyze the contraction of TV distance (or equivalently, the decay of Dobrushin contraction coefficients, cf. (2.46) and (2.49) in chapter 2) along certain structured bounded indegree Bayesian networks.

Specifically, we study a generalization of the well-known problem of *broadcasting on trees* [83] to the setting of bounded indegree directed acyclic graphs (DAGs). In the broadcasting on trees problem, we are given a noisy tree T whose vertices are Bernoulli random variables and edges are independent BSCs with common crossover probability $\delta \in (0, \frac{1}{2})$. Given that the root is an unbiased random bit, the objective is to decode the bit at the root from the bits at the kth layer of the tree as $k \to \infty$. The authors of [83] characterize the sharp threshold for when such reconstruction is possible:

- If $(1-2\delta)^2 \operatorname{br}(T) > 1$, then the minimum probability of error in decoding is bounded away from $\frac{1}{2}$ for all k,
- If $(1-2\delta)^2 \operatorname{br}(T) < 1$, then the minimum probability of error in decoding tends to $\frac{1}{2}$ as $k \to \infty$,

where $\operatorname{br}(T)$ denotes the *branching number* of the tree (see [179, Chapter 1.2]), and the condition $(1-2\delta)^2 \operatorname{br}(T) \gtrless 1$ is known as the *Kesten-Stigum threshold* in the regular tree setting. This result on reconstruction on trees generalizes results from random

processes and statistical physics that hold for regular trees, cf. [151] (which proves achievability) and [31] (which proves the converse), and has had numerous extensions and further generalizations including [28, 143, 144, 146, 210, 211, 225, 257, 258]. (We refer readers to [214] for a survey of the reconstruction problem on trees.) A consequence of this result is that reconstruction is impossible for trees with sub-exponentially many vertices at each layer. Indeed, if L_k denotes the number of vertices at layer k and $\lim_{k\to\infty} L_k^{1/k} = 1$, then it is straightforward to show that $\mathrm{br}(T) = 1$, which in turn implies that $(1-2\delta)^2 \mathrm{br}(T) < 1$.

Instead of analyzing trees, we consider the problem of broadcasting on *bounded indegree DAGs*. As in the setting of trees, all vertices in our graphs are Bernoulli random variables and all edges are independent BSCs. Furthermore, the values of variables located at vertices with indegree 2 or more are obtained by applying *Boolean processing functions* to their noisy inputs. Hence, compared to the setting of trees, broadcasting on DAGs has two principal differences:

1. In trees, layer sizes scale exponentially in the depth, while in DAGs, they are usually polynomial (or at least sub-exponential) in the depth.

2. In trees, the indegree of each vertex is 1, while in DAGs, each vertex has several incoming signals.

The latter enables the possibility of information fusion at the vertices of DAGs, and our main goal is to understand whether the benefits of 2 overpower the shortcoming of 1 and permit reconstruction of the root bit with sub-exponential layer size.

This chapter has three main contributions. Firstly, via a probabilistic argument using random DAGs, we demonstrate the existence of bounded indegree DAGs with $L_k = \Omega(\log(k))$ which permit recovery of the root bit for sufficiently low δ's. Secondly, we provide explicit deterministic constructions of such DAGs using regular bipartite lossless expander graphs. In particular, the constituent expander graphs for the first r layers of such DAGs can be constructed in either deterministic quasi-polynomial time or randomized polylogarithmic time in r. Together, these results imply that in terms of economy of storing information, *DAGs are doubly-exponentially more efficient than trees*. Thirdly, we show the impossibility result that no such recovery is possible on a two-dimensional (2D) regular grid if all intermediate vertices with indegree 2 use logical AND as the processing function, or all use XOR as the processing function. (This leaves only NAND as the remaining symmetric processing function.)

■ 5.1 Motivation

Broadcasting on DAGs has several natural interpretations. Perhaps most pertinently, it captures the feasibility of reliably communicating through Bayesian networks in the field of *communication networks*. Indeed, suppose a sender communicates a sequence of bits to a receiver through a large network. If broadcasting is impossible on this network, then the "wavefront of information" for each bit decays irrecoverably through

the network, and the receiver cannot reconstruct the sender's message regardless of the coding scheme employed.

The problem of broadcasting on DAGs is also closely related to the problem of *reliable computation using noisy circuits*, whose study was initiated in the seminal work [286] (also see [85]). The relation between the two models can be understood in the following way. Suppose we want to remember a bit using a noisy circuit of depth k. The "von Neumann approach" is to take multiple perfect clones of the bit and recursively apply noisy gates in order to reduce the overall noise [86, 115]. In contrast, the broadcasting perspective is to start from a single bit and repeatedly create noisy clones and apply perfect gates to these clones so that one can recover the bit reasonably well from the vertices at depth k. Thus, the broadcasting model can be construed as a noisy circuit that remembers a bit using perfect logic gates at the vertices and edges or wires that independently make errors. (It is worth mentioning that broadcasting DAG circuits are much smaller, and hence more desirable, than broadcasting tree circuits because bounded degree logic gates can be used to reduce noise.)

Furthermore, special cases of the broadcasting model have found applications in various discrete probability questions. For example, broadcasting on trees corresponds to *ferromagnetic Ising models in statistical physics*. Specifically, if we associate bits $\{0, 1\}$ with spins $\{-1, +1\}$, then the joint distribution of the vertices of any finite broadcasting subtree (e.g. up to depth $k \in \mathbb{N}$) corresponds to the *Boltzmann-Gibbs distribution* of the configuration of spins in the subtree (where we assume that the strictly positive common interaction strength is fixed and that there is no external magnetic field) [83, Section 2.2].[75] In the theory of Ising models, weak limits of Boltzmann-Gibbs distributions on finite subgraphs of an infinite graph with different boundary conditions yield different *Gibbs measures or states* on the infinite graph, cf. [91, Chapters 3 and 6]. For instance, our broadcasting distribution on the infinite tree corresponds to the Gibbs measure with *free boundary conditions*, which is obtained by taking a weak limit of the broadcasting distributions over finite subtrees. Moreover, it is well-known that under general conditions, the Dobrushin-Lanford-Ruelle (DLR) formalism for defining Gibbs measures (or *DLR states*) using *Gibbsian specifications* produces a convex *Choquet simplex* of Gibbs measures corresponding to any particular specification [98, Chapters 1, 2, and 7] (also see [91, Chapter 6]).[76] Hence, it is of both mathematical and physical interest to find the extremal Gibbs measures of this simplex.[77] It turns out that reconstruction is

[75] In particular, each value of $\delta \in \left(0, \frac{1}{2}\right)$ corresponds to a unique value of temperature such that the broadcasting distribution defined by δ is equivalent to the Boltzmann-Gibbs distribution with the associated temperature parameter [83, Equation (11)].

[76] We refer readers to [255] for a classical reference on the rigorous theory of phase transitions.

[77] Indeed, extremal Gibbs measures are precisely those Gibbs measures that have trivial tail σ-algebra, i.e. tail events exhibit a zero-one law for extremal Gibbs measures, cf. [98, Section 7.1], [91, Section 6.8]. As explained in [98, Comment (7.8)] and [91, p.302], since tail events correspond to macroscopic events that are not affected by the behavior of any finite subset of spins, extremal Gibbs measures have deterministic macroscopic events. Thus, from a physical perspective, only extremal Gibbs measures are suitable to characterize the equilibrium states of a statistical mechanical system.

impossible on a broadcasting tree if and only if the Gibbs measure with free boundary conditions of the corresponding ferromagnetic Ising model is extremal [31], [83, Section 2.2]. This portrays a strong connection between broadcasting on trees and the theory of Ising models. We refer readers to [143,144,225,258] for related work, and to [83, Section 2.2] for further references on the Ising model literature.

A second example of a related discrete probability question stems from the theory of *probabilistic cellular automata* (PCA). Indeed, another motivation for our problem is to understand whether it is possible to propagate information in regular grids starting from the root—see Figure 5.1 for a 2D example. Our *conjecture* is that such propagation is possible for sufficiently low noise δ in 3 or more dimensions, and impossible for a 2D regular grid regardless of the noise level and of the choice of processing function (which is the same for every vertex). This conjecture is inspired by work on the *positive rates conjecture* for one-dimensional (1D) PCA, cf. [109, Section 1], and the existence of non-ergodic 2D PCA such as that defined by *Toom's North-East-Center (NEC) rule*, cf. [274]. Notice that broadcasting on 2D regular grids can be perceived as 1D PCA with boundary conditions that limit the layer sizes to be $L_k = k + 1$, and the impossibility of broadcasting on 2D regular grids intuitively corresponds to the ergodicity of 1D PCA (along with sufficiently fast convergence rate). Therefore, the existence of a 2D regular grid (with a choice of processing function) which remembers its initial state (bit) for infinite time would suggest the existence of non-ergodic infinite 1D PCA consisting of 2-input binary-state cells. However, the positive rates conjecture suggests that "relatively simple" 1D PCA with local interactions and strictly positive noise probabilities are ergodic, and known counter-example constructions to this conjecture require a lot more states [93], or are non-uniform in time and space [48]. This gives credence to our conjecture that broadcasting is impossible for 2D regular grids. Furthermore, much like 2D regular grids, broadcasting on three-dimensional (3D) regular grids can be perceived as 2D PCA with boundary conditions. Hence, the existence of non-ergodic 2D PCA [274] suggests the existence of 3D regular grids where broadcasting is possible, thereby lending further credence to our conjecture. In this chapter, we take some first steps towards establishing the 2D part of our conjecture.

Finally, reconstruction on trees also plays a fundamental role in understanding various questions in theoretical computer science and learning theory. For example, results on trees have been exploited in problems of *ancestral data and phylogenetic tree reconstruction*—see e.g. [63, 212, 213, 238]. In fact, the existence results obtained in this chapter suggest that it might be possible to reconstruct other biological networks, such as phylogenetic networks (see e.g. [142]) or pedigrees (see e.g. [259, 271]), even if the growth of the network is very mild. Moreover, broadcasting on trees can be used to understand phase transitions for *random constraint satisfaction* problems—see e.g. [99,162,202,205] and follow-up work. It is an interesting future endeavor to explore if there are connections between broadcasting on general DAGs and random constraint satisfaction problems. Currently, we are not aware that such connections have been established. Lastly, we note that broadcasting on trees has also been used to prove

impossibility of weak recovery (or detection) in the problem of *community detection in stochastic block models*, cf. [1, Section 5.1].

■ 5.2 Chapter Outline

We briefly outline the rest of this chapter. In the next section 5.3, we formally define the random DAG and deterministic 2D regular grid models. In section 5.4, we present our five main results (as well as some auxiliary results) pertaining to these models, and discuss several related results in the literature. Then, we prove these main results in sections 5.5, 5.6, 5.7, 5.8, and 5.9, respectively. In particular, section 5.5 analyzes broadcasting with majority processing functions when the indegree of each vertex is 3 or more, section 5.6 analyzes broadcasting with AND and OR processing functions when the indegree of each vertex is 2, section 5.7 illustrates our explicit constructions of DAGs where reconstruction of the root bit is possible using expander graphs, section 5.8 proves the impossibility of broadcasting over a deterministic 2D regular grid with all AND processing functions, and section 5.9 proves the impossibility of broadcasting over a deterministic 2D regular grid with all XOR processing functions. Finally, we conclude our discussion and list some open problems in section 5.10.

■ 5.3 Formal Definitions

Since we will use probabilistic arguments to establish the existence of bounded indegree DAGs where reconstruction of root bit is possible, we will prove many of our results for random DAGs. So, the next subsection 5.3.1 formally defines the random DAG model. On the other hand, in order to present our impossibility results on 2D regular grids, the subsequent subsection 5.3.2 formally defines the deterministic 2D regular grid model.

■ 5.3.1 Random DAG Model

A *random DAG model* consists of an infinite DAG with fixed vertices that are Bernoulli ($\{0,1\}$-valued) random variables and randomly generated edges which are independent BSCs. We first define the vertex structure of this model, where each vertex is identified with the corresponding random variable. Let the root or "source" random variable be $X_{0,0} \sim \mathsf{Bernoulli}(\frac{1}{2})$. Furthermore, we define $X_k = (X_{k,0}, \ldots, X_{k,L_k-1})$ as the vector of vertex random variables at distance (i.e. length of shortest path) $k \in \mathbb{N} \cup \{0\}$ from the root, where $L_k \in \mathbb{N}$ denotes the number of vertices at distance k. In particular, we have $X_0 = (X_{0,0})$ so that $L_0 = 1$, and we are typically interested in the regime where $L_k \to \infty$ as $k \to \infty$.

We next define the edge structure of the random DAG model. For any $k \in \mathbb{N}$ and any $j \in [L_k]$, we independently and uniformly select $d \in \mathbb{N}$ vertices $X_{k-1,i_1}, \ldots, X_{k-1,i_d}$ with replacement from X_{k-1} (i.e. $i_1, \ldots, i_d$ are i.i.d. uniform on $[L_{k-1}]$), and then construct d directed edges: $(X_{k-1,i_1}, X_{k,j}), \ldots, (X_{k-1,i_d}, X_{k,j})$. (Here, $i_1, \ldots, i_d$ are independently chosen for each $X_{k,j}$.) This random process generates the underlying DAG structure.

In the sequel, we will let G be a random variable representing this underlying (infinite) random DAG, i.e. G encodes the random configuration of the edges between the vertices.

To define a *Bayesian network* (or directed graphical model) on this random DAG, we fix some sequence of Boolean functions $f_k : \{0,1\}^d \to \{0,1\}$ for $k \in \mathbb{N}$ (that depend on the level index k, but typically not on the realization of G), and some crossover probability $\delta \in (0, \frac{1}{2})$ (since this is the interesting regime of δ).[78] Then, for any $k \in \mathbb{N}$ and $j \in [L_k]$, given $i_1, \ldots, i_d$ and $X_{k-1,i_1}, \ldots, X_{k-1,i_d}$, we define:

$$X_{k,j} = f_k(X_{k-1,i_1} \oplus Z_{k,j,1}, \ldots, X_{k-1,i_d} \oplus Z_{k,j,d}) \tag{5.1}$$

where $\oplus$ denotes addition modulo 2,[79] and $\{Z_{k,j,i} : k \in \mathbb{N},\, j \in [L_k],\, i \in \{1, \ldots, d\}\}$ are i.i.d. $\mathsf{Bernoulli}(\delta)$ random variables that are independent of everything else. This means that each edge is a $\mathsf{BSC}(\delta)$. Moreover, (5.1) characterizes the conditional distribution of $X_{k,j}$ given its parents. In this model, the Boolean processing function used at a vertex $X_{k,j}$ depends only on the level index k. A more general model can be defined where each vertex $X_{k,j}$ has its own Boolean processing function $f_{k,j} : \{0,1\}^d \to \{0,1\}$ for $k \in \mathbb{N}$ and $j \in [L_k]$. However, with the exception of a few converse results, we will mainly analyze instances of the simpler model in this chapter.

Note that although we will analyze this model for convenience, as stated, our underlying graph is really a directed multigraph rather than a DAG, because we select the parents of a vertex with replacement. It is straightforward to construct an equivalent model where the underlying graph is truly a DAG. For each vertex $X_{k,j}$ with $k \in \mathbb{N}$ and $j \in [L_k]$, we first construct d intermediate parent vertices $\{X^i_{k,j} : i \in \{1, \ldots, d\}\}$ that live between layers k and $k-1$, where each $X^i_{k,j}$ has a single edge pointing to $X_{k,j}$. Then, for each $X^i_{k,j}$, we independently and uniformly select a vertex from layer $k-1$, and construct a directed edge from that vertex to $X^i_{k,j}$. This defines a valid (random) DAG. As a result, every realization of G can be perceived as either a directed multigraph or its equivalent DAG. Furthermore, the Bayesian network on this true DAG is defined as follows: each $X_{k,j}$ is the output of a Boolean processing function f_k with inputs $\{X^i_{k,j} : i \in \{1, \ldots, d\}\}$, and each $X^i_{k,j}$ is the output of a BSC whose input is the unique parent of $X^i_{k,j}$ in layer $k-1$.

Finally, we define the "empirical probability of unity" at level $k \in \mathbb{N} \cup \{0\}$ as:

$$\sigma_k \triangleq \frac{1}{L_k} \sum_{m=0}^{L_k - 1} X_{k,m} \tag{5.2}$$

where $\sigma_0 = X_{0,0}$ is just the root vertex. Observe that given $\sigma_{k-1} = \sigma$, the variables $X_{k-1,i_1}, \ldots, X_{k-1,i_d}$ are i.i.d. $\mathsf{Bernoulli}(\sigma)$, and as a result, $X_{k-1,i_1} \oplus Z_{k,j,1}, \ldots, X_{k-1,i_d} \oplus Z_{k,j,d}$ are i.i.d. $\mathsf{Bernoulli}(\sigma * \delta)$, where $\sigma * \delta \triangleq \sigma(1-\delta) + \delta(1-\sigma)$ is the convolution of σ

[78] The cases $\delta = 0$ and $\delta = \frac{1}{2}$ are uninteresting because the former corresponds to a deterministic DAG and the latter corresponds to an independent DAG.

[79] This notation should not be confused with the use of $\oplus$ in chapter 3 to represent an arbitrary finite Abelian group operation.

and δ. Therefore, $X_{k,j}$ is the output of f_k upon inputting a d-length i.i.d. Bernoulli$(\sigma * \delta)$ string.

Under this setup, our objective is to determine whether or not the value at the root $\sigma_0 = X_{0,0}$ can be decoded from the observations X_k as $k \to \infty$. Since X_k is an exchangeable sequence of random variables given σ_0, for any $x_{0,0}, x_{k,0}, \ldots, x_{k,L_k-1} \in \{0,1\}$ and any permutation π of $[L_k]$, we have:

$$P_{X_k|\sigma_0}(x_{k,0}, \ldots, x_{k,L_k-1}|x_{0,0}) = P_{X_k|\sigma_0}(x_{k,\pi(0)}, \ldots, x_{k,\pi(L_k-1)}|x_{0,0})\,. \tag{5.3}$$

Letting $\sigma = \frac{1}{L_k}\sum_{j=0}^{L_k-1} x_{k,j}$, we can factorize $P_{X_k|\sigma_0}$ as:

$$P_{X_k|\sigma_0}(x_{k,0}, \ldots, x_{k,L_k-1}|x_{0,0}) = \binom{L_k}{L_k\sigma}^{-1} P_{\sigma_k|\sigma_0}(\sigma|x_{0,0})\,. \tag{5.4}$$

Using the Fisher-Neyman factorization theorem [150, Theorem 3.6], this implies that σ_k is a *sufficient statistic* of X_k for performing inference about σ_0. Therefore, we restrict our attention to the Markov chain $\{\sigma_k : k \in \mathbb{N} \cup \{0\}\}$ in our achievability proofs, since if decoding is possible from σ_k, then it is also possible from X_k. Given σ_k, inferring the value of σ_0 is a binary hypothesis testing problem with minimum achievable probability of error (given by Le Cam's relation, cf. [278, proof of Theorem 2.2(i)]):

$$\mathbb{P}\Big(f^k_{\mathsf{ML}}(\sigma_k) \neq \sigma_0\Big) = \frac{1}{2}\left(1 - \left\|P^+_{\sigma_k} - P^-_{\sigma_k}\right\|_{\mathsf{TV}}\right) \tag{5.5}$$

where $f^k_{\mathsf{ML}} : \{m/L_k : m \in \{0, \ldots, L_k\}\} \to \{0,1\}$ is the ML decision rule based on the empirical probability of unity at level k in the absence of knowledge of the random DAG realization G, and $P^+_{\sigma_k}$ and $P^-_{\sigma_k}$ are the conditional distributions of σ_k given $\sigma_0 = 1$ and $\sigma_0 = 0$, respectively. We say that reconstruction of the root bit σ_0 is possible when:[80]

$$\lim_{k\to\infty} \mathbb{P}\Big(f^k_{\mathsf{ML}}(\sigma_k) \neq \sigma_0\Big) < \frac{1}{2} \quad \Leftrightarrow \quad \lim_{k\to\infty} \left\|P^+_{\sigma_k} - P^-_{\sigma_k}\right\|_{\mathsf{TV}} > 0\,. \tag{5.6}$$

In the sequel, to simplify our analysis when proving that reconstruction is possible, we will sometimes use other (sub-optimal) decision rules rather than the ML decision rule.

On the other hand, we will consider the Markov chain $\{X_k : k \in \mathbb{N} \cup \{0\}\}$ conditioned on G in our converse proofs. We say that reconstruction of the root bit X_0 is impossible when:

$$\lim_{k\to\infty} \mathbb{P}\Big(h^k_{\mathsf{ML}}(X_k, G) \neq X_0 \Big| G\Big) = \frac{1}{2}\ G\text{-a.s.} \quad \Leftrightarrow \quad \lim_{k\to\infty} \left\|P^+_{X_k|G} - P^-_{X_k|G}\right\|_{\mathsf{TV}} = 0\ G\text{-a.s.} \tag{5.7}$$

[80] Note that the limits in (5.6), (5.7), and (5.8) exist because $\mathbb{P}(f^k_{\mathsf{ML}}(\sigma_k) \neq \sigma_0)$, $\mathbb{P}(h^k_{\mathsf{ML}}(X_k) \neq X_0)$, and $\mathbb{P}(h^k_{\mathsf{ML}}(X_k, G) \neq X_0|G)$ (for any fixed realization G) are monotone non-decreasing sequences in k that are bounded above by $\frac{1}{2}$. This can be deduced either from the data processing inequality for TV distance, or from the fact that a randomized decoder at level k can simulate the stochastic transition to level $k+1$.

where $h^k_{\mathsf{ML}}(\cdot, G) : \{0,1\}^{L_k} \to \{0,1\}$ is the ML decision rule based on the full state at level k given knowledge of the random DAG realization G, $P^+_{X_k|G}$ and $P^-_{X_k|G}$ denote the conditional distributions of X_k given $\{X_0 = 1, G\}$ and $\{X_0 = 0, G\}$, respectively, and the notation G-$a.s.$ implies that the conditions in (5.7) hold with probability 1 with respect to the distribution of the random DAG G. Note that applying the bounded convergence theorem to the TV distance condition in (5.7) yields $\lim_{k\to\infty} \mathbb{E}[\|P^+_{X_k|G} - P^-_{X_k|G}\|_{\mathsf{TV}}] = 0$, and employing Jensen's inequality here establishes the weaker impossibility result:

$$\lim_{k\to\infty} \mathbb{P}\Big(h^k_{\mathsf{ML}}(X_k) \neq X_0\Big) = \frac{1}{2} \quad \Leftrightarrow \quad \lim_{k\to\infty} \Big\|P^+_{X_k} - P^-_{X_k}\Big\|_{\mathsf{TV}} = 0 \tag{5.8}$$

where $h^k_{\mathsf{ML}} : \{0,1\}^{L_k} \to \{0,1\}$ is the ML decision rule based on the full state at level k in the absence of knowledge of the random DAG realization G, and $P^+_{X_k}$ and $P^-_{X_k}$ are the conditional distributions of X_k given $X_0 = 1$ and $X_0 = 0$, respectively. Since σ_k is a sufficient statistic of X_k for performing inference about σ_0 when we average over G, we have:

$$\mathbb{P}\Big(h^k_{\mathsf{ML}}(X_k) \neq X_0\Big) = \mathbb{P}\Big(f^k_{\mathsf{ML}}(\sigma_k) \neq \sigma_0\Big) \tag{5.9}$$

or equivalently:

$$\Big\|P^+_{X_k} - P^-_{X_k}\Big\|_{\mathsf{TV}} = \Big\|P^+_{\sigma_k} - P^-_{\sigma_k}\Big\|_{\mathsf{TV}} \tag{5.10}$$

and the condition in (5.8) is a counterpart of (5.6).

■ 5.3.2 Two-Dimensional Regular Grid Model

We now turn to defining deterministic DAG models. As mentioned earlier, all deterministic DAGs we analyze in this chapter will have the structure of a 2D regular grid. A *2D regular grid model* consists of an infinite DAG whose vertices are also Bernoulli random variables and whose edges are independent $\mathsf{BSC}(\delta)$'s. As with random DAG models, the root or source random variable of the grid is denoted $X_{0,0} \sim \mathsf{Bernoulli}(\frac{1}{2})$, and we let $X_k = (X_{k,0}, \ldots, X_{k,k})$ be the vector of vertex random variables at distance $k \in \mathbb{N} \cup \{0\}$ from the root. So, there are $k+1$ vertices at distance k. Furthermore, the 2D regular grid contains the (deterministic) directed edges $(X_{k,j}, X_{k+1,j})$ and $(X_{k,j}, X_{k+1,j+1})$ for every $k \in \mathbb{N} \cup \{0\}$ and every $j \in [k+1]$. The underlying DAG of such a 2D regular grid is shown in Figure 5.1.

To construct a Bayesian network on this 2D regular grid, we again fix some crossover probability parameter $\delta \in (0, \frac{1}{2})$, and two Boolean processing functions $f_1 : \{0,1\} \to \{0,1\}$ and $f_2 : \{0,1\}^2 \to \{0,1\}$. Then, for any $k \in \mathbb{N}\backslash\{1\}$ and $j \in \{1, \ldots, k-1\}$, we define:[81]

$$X_{k,j} = f_2(X_{k-1,j-1} \oplus Z_{k,j,1}, X_{k-1,j} \oplus Z_{k,j,2}) \tag{5.11}$$

and for any $k \in \mathbb{N}$, we define:

$$X_{k,0} = f_1(X_{k-1,0} \oplus Z_{k,0,2}) \text{ and } X_{k,k} = f_1(X_{k-1,k-1} \oplus Z_{k,k,1}) \tag{5.12}$$

[81] We can similarly define a more general model where every vertex $X_{k,j}$ has its own Boolean processing function $f_{k,j}$, but we will only analyze instances of the simpler model presented here.

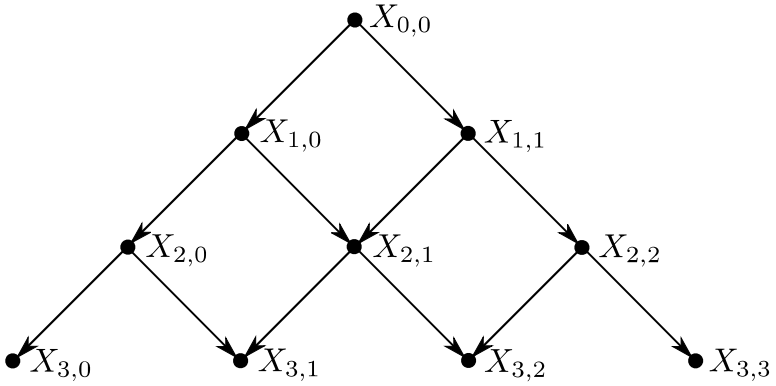


Figure 5.1. Illustration of a 2D regular grid where each vertex is a Bernoulli random variable and each edge is a BSC with parameter $\delta \in \left(0, \frac{1}{2}\right)$. Moreover, each vertex with indegree 2 uses a common Boolean processing function to combine its noisy input bits.

where $\{Z_{k,j,i} : k \in \mathbb{N}, j \in [k+1], i \in \{1,2\}\}$ are i.i.d $\mathsf{Bernoulli}(\delta)$ random variables that are independent of everything else. Together, (5.11) and (5.12) characterize the conditional distribution of any $X_{k,j}$ given its parents.

As before, the sequence $\{X_k : k \in \mathbb{N} \cup \{0\}\}$ forms a Markov chain, and our goal is to determine whether or not the value at the root X_0 can be decoded from the observations X_k as $k \to \infty$. Given X_k for any fixed $k \in \mathbb{N}$, inferring the value of X_0 is a binary hypothesis testing problem with minimum achievable probability of error:

$$\mathbb{P}\Big(h^k_{\mathsf{ML}}(X_k) \neq X_0\Big) = \frac{1}{2}\left(1 - \left\|P^+_{X_k} - P^-_{X_k}\right\|_{\mathsf{TV}}\right) \tag{5.13}$$

where $h^k_{\mathsf{ML}} : \{0,1\}^{k+1} \to \{0,1\}$ is the ML decision rule based on X_k at level k (with knowledge the 2D regular grid), and $P^+_{X_k}$ and $P^-_{X_k}$ are the conditional distributions of X_k given $X_0 = 1$ and $X_0 = 0$, respectively. Therefore, we say that reconstruction of the root bit X_0 is impossible (or "broadcasting is impossible") when:[82]

$$\lim_{k\to\infty} \mathbb{P}\Big(h^k_{\mathsf{ML}}(X_k) \neq X_0\Big) = \frac{1}{2} \quad \Leftrightarrow \quad \lim_{k\to\infty} \left\|P^+_{X_k} - P^-_{X_k}\right\|_{\mathsf{TV}} = 0 \tag{5.14}$$

where the equivalence follows from (5.13).[83] In every impossibility result in this chapter, we will prove that reconstruction of X_0 is impossible in the sense of (5.14).

In closing this section, we briefly elucidate the relationship between the feasibility of broadcasting on DAGs and the Dobrushin contraction coefficient for TV distance.

[82] As before, the limits in (5.14) exist because $\mathbb{P}(h^k_{\mathsf{ML}}(X_k) \neq X_0)$ is a monotone non-decreasing sequence in k that is bounded above by $\frac{1}{2}$. The upper bound of $\frac{1}{2}$ is a trivial consequence of the fact that a randomly generated bit cannot beat the ML decoder.

[83] In contrast to (5.14), we say that reconstruction is possible (or "broadcasting is possible") when $\lim_{k\to\infty} \mathbb{P}(h^k_{\mathsf{ML}}(X_k) \neq X_0) < \frac{1}{2}$, or equivalently, $\lim_{k\to\infty} \|P^+_{X_k} - P^-_{X_k}\|_{\mathsf{TV}} > 0$.

For any deterministic DAG (such as a 2D regular grid, or more generally, a realization of the random DAG G), the Dobrushin contraction coefficient of the transition kernel from X_0 to X_k, $P_{X_k|X_0}$, is given by:

$$\eta_{\mathsf{TV}}(P_{X_k|X_0}) = \left\| P_{X_k}^{+} - P_{X_k}^{-} \right\|_{\mathsf{TV}} \tag{5.15}$$

using the two-point characterization of η_{TV} in (2.49) in chapter 2. Therefore, as suggested by (5.14), reconstruction of the root bit X_0 is impossible for DAGs if and only if the Dobrushin contraction coefficient vanishes:

$$\lim_{k\to\infty} \eta_{\mathsf{TV}}(P_{X_k|X_0}) = 0\,. \tag{5.16}$$

From this perspective, the broadcasting on DAGs problem is entirely a question about the asymptotic vanishing of Dobrushin contraction coefficients. This observation appropriately places the results in this chapter within the context of the concepts studied in previous chapters.

■ 5.4 Main Results and Discussion

In this section, we state our main results, briefly delineate the main techniques or intuition used in the proofs, and discuss related literature.

■ 5.4.1 Results on Random DAG Models

We prove two main results on the random DAG model. The first considers the setting where the indegree of each vertex (except the root) is $d \geq 3$. In this scenario, taking a majority vote of the inputs at each vertex intuitively appears to have good "local error correction" properties. So, we fix all Boolean functions in the random DAG model to be the (d-input) *majority* rule, and prove that this model exhibits a phase transition phenomenon around a critical threshold of:

$$\delta_{\mathsf{maj}} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\left\lceil \frac{d}{2} \right\rceil \binom{d}{\lceil \frac{d}{2} \rceil}}\,. \tag{5.17}$$

Indeed, the theorem below illustrates that for $\delta < \delta_{\mathsf{maj}}$, the majority decision rule:

$$\hat{S}_k \triangleq \mathbb{1}\left\{ \sigma_k \geq \frac{1}{2} \right\} \tag{5.18}$$

can asymptotically decode σ_0, but for $\delta > \delta_{\mathsf{maj}}$, the ML decision rule with knowledge of G cannot asymptotically decode σ_0.

Theorem 5.1 (Phase Transition in Random DAG Model with Majority Rule Processing). *Let $C(\delta, d)$ and $D(\delta, d)$ be the constants defined in (5.49) and (5.45)*

in section 5.5. For a random DAG model with $d \geq 3$ and majority processing functions (where ties are broken by outputting random bits), the following phase transition phenomenon occurs around δ_{maj}:

1. *If $\delta \in (0, \delta_{\mathsf{maj}})$, and the number of vertices per level satisfies $L_k \geq C(\delta, d)\log(k)$ for all sufficiently large k (depending on δ and d), then reconstruction is possible in the sense that:*
$$\limsup_{k\to\infty} \mathbb{P}(\hat{S}_k \neq \sigma_0) < \frac{1}{2}$$
where we use the majority decoder $\hat{S}_k = \mathbb{1}\{\sigma_k \geq \frac{1}{2}\}$ at level k.

2. *If $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$, and the number of vertices per level satisfies $L_k = o(D(\delta, d)^{-k})$, then reconstruction is impossible in the sense of (5.7):*
$$\lim_{k\to\infty} \left\| P^+_{X_k|G} - P^-_{X_k|G} \right\|_{\mathsf{TV}} = 0 \quad G\text{-a.s.}$$

Theorem 5.1 is proved in section 5.5. Intuitively, the proof considers the conditional expectation function, $g : [0,1] \to [0,1]$, $g(\sigma) = \mathbb{E}[\sigma_k | \sigma_{k-1} = \sigma]$ (see (5.40) and (5.41) in section 5.5), which provides the approximate value of σ_k given the value of σ_{k-1} for large k. This function turns out to have three fixed points when $\delta \in (0, \delta_{\mathsf{maj}})$, and only one fixed point when $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$. In the former case, σ_k "moves" to the largest fixed point when $\sigma_0 = 1$, and to the smallest fixed point when $\sigma_0 = 0$. In the latter case, σ_k "moves" to the unique fixed point of $\frac{1}{2}$ regardless of the value of σ_0 (see Proposition 5.5 in section 5.5).[84] This provides the guiding intuition for why we can asymptotically decode σ_0 when $\delta \in (0, \delta_{\mathsf{maj}})$, but not when $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$.

The recursive (or fixed point) structure of g in the special case where $d = 3$ and $\delta_{\mathsf{maj}} = \frac{1}{6}$ can be traced back to the work of von Neumann in [286]. So, it is worth comparing Theorem 5.1 with von Neumann's results in [286, Section 8], where the threshold of $\frac{1}{6}$ is also significant. In [286, Section 8], von Neumann demonstrates the possibility of reliable computation by constructing a circuit with successive layers of computation and local error correction using 3-input δ-noisy majority gates (i.e. the gates independently make errors with probability δ). In his analysis, he first derives a simple recursion that captures the effect on the probability of error after applying a single noisy majority gate. Then, he uses a "heuristic" fixed point argument to show that as the depth of the circuit grows, the probability of error asymptotically stabilizes at a fixed point value less than $\frac{1}{2}$ if $\delta < \frac{1}{6}$, and the probability of error tends to $\frac{1}{2}$ if $\delta \geq \frac{1}{6}$. Moreover, he rigorously proves that reliable computation is possible for $\delta < 0.0073$.

As we mentioned in section 5.1, von Neumann's approach to remembering a random initial bit entails using multiple clones of the initial bit as inputs to a noisy circuit with one output, where the output equals the initial bit with probability greater than $\frac{1}{2}$ for

[84] Note, however, that $\sigma_k \to \frac{1}{2}$ almost surely as $k \to \infty$ does not imply the impossibility of reconstruction in the sense of (5.8), let alone (5.7). So, a different argument is required to establish such impossibility results.

"good" choices of noisy gates. It is observed in [115, Section 2] that a balanced ternary tree circuit, with k layers of 3-input noisy majority gates and 3^k inputs that are all equal to the initial bit, can be used to remember the initial bit. In fact, von Neumann's heuristic fixed point argument that yields a critical threshold of $\frac{1}{6}$ for reconstruction is rigorous in this scenario. From this starting point, Hajek and Weller prove the stronger impossibility result that reliable computation is impossible for formulae (i.e. circuits where the output of each intermediate gate is the input of only one other gate) with general 3-input δ-noisy gates when $\delta \geq \frac{1}{6}$ [115, Proposition 2]. This development can be generalized for any odd $d \geq 3$, and [86, Theorem 1] conveys that reliable computation is impossible for formulae with general d-input δ-noisy gates when $\delta \geq \delta_{\mathsf{maj}}$.

The discussion heretofore reveals that the critical thresholds in von Neumann's circuit for remembering a bit and in our model in Theorem 5.1 are both δ_{maj}. It turns out that this is a consequence of the common fixed point iteration structure of the two problems (as we will explain below). Indeed, the general recursive structure of g for any odd value of d was analyzed in [86, Section 2]. On a related front, the general recursive structure of g was also analyzed in [210] in the context of performing recursive reconstruction on periodic trees, where the critical threshold of δ_{maj} again plays a crucial role. In fact, we will follow the analysis in [210] to develop these recursions in section 5.5.

We now elucidate the common fixed point iteration structure between von Neumann's model and our model in Theorem 5.1. Suppose $d \geq 3$ is odd, and define the function $h : [0,1] \to [0,1]$, $h(p) \triangleq \mathbb{P}(\mathsf{majority}(Y_1, \ldots, Y_d) = 1)$ for $Y_1, \ldots, Y_d$ i.i.d. $\mathsf{Bernoulli}(p)$. Consider von Neumann's balanced d-ary tree circuit with k layers of d-input δ-noisy majority gates and d^k inputs that are all equal to the initial bit. In this model, it is straightforward to verify that the probability of error (i.e. output vertex $\neq$ initial bit) is $f^{(k)}(0)$, where $f : [0,1] \to [0,1]$ is given by [86, Equation (3)]:

$$f(\sigma) \triangleq \delta * h(\sigma)\,, \tag{5.19}$$

and $f^{(k)}$ denotes the k-fold composition of f with itself. On the other hand, as explained in the brief intuition for our proof of Theorem 5.1 earlier, assuming that $\sigma_0 = 0$, the relevant recursion for our model is given by the repeated composition $g^{(k)}(0)$ (which captures the average value of σ_k after k layers). According to (5.40) in section 5.5, $g(\sigma) = h(\delta * \sigma)$, which yields the relation:

$$\forall k \in \mathbb{N}, \quad f^{(k+1)}(0) = \delta * g^{(k)}(0) \tag{5.20}$$

by induction. Therefore, the fixed point iteration structures of f and g are identical, and δ_{maj} is the common critical threshold that determines when there is a unique fixed point. In particular, the fact that gates (or vertices) are noisy in von Neumann's model, while edges (or wires) are noisy in our model, has no bearing on this fixed point structure.[85]

Although both the aforementioned models use majority gates and share a common fixed point structure, it is important to recognize that our overall analysis differs from

[85] We refer readers to [73] for general results on the relation between vertex noise and edge noise.

von Neumann's analysis in a crucial way. Since our recursion pertains to conditional expectations of the proportion of 1's in different layers (rather than the probabilities of error in von Neumann's setting), our proof requires exponential concentration inequalities to formalize the intuition provided by the fixed point analysis.

We now make several other pertinent remarks about Theorem 5.1. Firstly, reconstruction is possible in the sense of (5.6) when $\delta \in (0, \delta_{\mathsf{maj}})$ since the ML decision rule achieves lower probability of error than the majority decision rule,[86] and reconstruction is impossible in the sense of (5.8) when $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$ (as explained at the end of subsection 5.3.1). Furthermore, while part 1 of Theorem 5.1 only shows that the ML decoder $f^k_{\mathsf{ML}}(\sigma_k)$ based on σ_k is optimal in the absence of knowledge of the particular graph realization G, part 2 establishes that even if the ML decoder knows the graph G and has access to the full k-layer state X_k, it cannot beat the δ_{maj} threshold in all but a zero measure set of DAGs.

Secondly, the following conjecture is still open: In the random DAG model with $L_k = O(\log(k))$ and fixed $d \geq 3$, reconstruction is impossible for all choices of Boolean processing functions when $\delta \geq \delta_{\mathsf{maj}}$. A consequence of this conjecture is that majority processing functions are optimal, i.e. they achieve the δ_{maj} reconstruction threshold. The results in [210] provide strong evidence that this conjecture is true when all vertices in the random DAG use the same odd Boolean processing function. Indeed, for fixed $\delta \in (0, \frac{1}{2})$ and any odd Boolean function $\mathsf{gate} : \{0,1\}^d \to \{0,1\}$, let $\tilde{g} : [0,1] \to [0,1]$ be defined as $\tilde{g}(\sigma) \triangleq \mathbb{P}(\mathsf{gate}(Y_1, \ldots, Y_d) = 1)$ for $Y_1, \ldots, Y_d$ i.i.d. $\mathsf{Bernoulli}(\delta * \sigma)$.[87] Then, [210, Lemma 2.4] establishes that $\tilde{g}(\sigma) \leq g(\sigma)$ for all $\sigma \geq \frac{1}{2}$ and $\tilde{g}(\sigma) \geq g(\sigma)$ for all $\sigma \leq \frac{1}{2}$, where the function g is given in (5.40) (and corresponds to the majority rule). Hence, if g has a single fixed point at $\sigma = \frac{1}{2}$, $\tilde{g}$ also has a single fixed point at $\sigma = \frac{1}{2}$. This intuitively suggests that if reconstruction of the root bit is impossible using majority processing functions, it is also impossible using any odd processing function. Furthermore, our proof of part 2 of Theorem 5.1 in section 5.5 yields that reconstruction is impossible for all choices of odd and monotone non-decreasing Boolean processing functions when $\delta > \delta_{\mathsf{maj}}$, modulo the following conjecture (which we did not verify): among all odd and monotone non-decreasing Boolean functions, the maximum Lipschitz constant of $\tilde{g}$ is attained by the majority rule at $\sigma = \frac{1}{2}$.

Thirdly, the sub-exponential layer size condition $L_k = o(D(\delta, d)^{-k})$ in part 2 of Theorem 5.1 is intuitively necessary. Suppose every Boolean processing function in our

[86] It can be seen from monotonicity and symmetry considerations that without knowledge of the random DAG realization G, the ML decision rule $f^k_{\mathsf{ML}}(\sigma_k)$ is equal to the majority decision rule $\hat{S}_k$. (So, the superior limit in part 1 of Theorem 5.1 can be replaced by a true limit.) In fact, simulations illustrate that the conditional distributions $P^+_{\sigma_k}$ and $P^-_{\sigma_k}$ have the *monotone likelihood ratio property*, i.e. the likelihood ratio $P^+_{\sigma_k}(\sigma)/P^-_{\sigma_k}(\sigma)$ is non-decreasing in σ, which also implies that $f^k_{\mathsf{ML}}(\sigma_k)$ is equal to $\hat{S}_k$. On the other hand, with knowledge of the random DAG realization G, the ML decision rule $f^k_{\mathsf{ML}}(\sigma_k, G)$ based on σ_k is not the majority decision rule.

[87] A Boolean function is said to be *odd* if flipping all its input bits also flips the output bit. The assumption that gate is odd ensures that the function $R^{\delta}_{\mathsf{gate}}(\sigma)$ in [210, Definition 2.1] is precisely equal to the function $\tilde{g}(\sigma)$.

random DAG model simply outputs the value of its first input bit. This effectively sets $d = 1$, and reduces our problem to one of broadcasting on a random tree model. If $L_k = \Omega(E(\delta)^k)$ for some large enough constant $E(\delta)$, then most realizations of the random tree will have branching numbers greater than $(1-2\delta)^{-2}$. As a result, reconstruction will be possible for most realizations of the random tree (cf. the Kesten-Stigum threshold delineated at the outset of this chapter). Thus, when we are proving impossibility results, L_k (at least intuitively) cannot be exponential in k with a very large base.

Fourthly, it is worth mentioning that for any fixed DAG with indegree $d \geq 3$ and sub-exponential L_k, for any choices of Boolean processing functions, and any choice of decoder, it is impossible to reconstruct the root bit when:

$$\delta > \frac{1}{2} - \frac{1}{2\sqrt{d}}\,. \tag{5.21}$$

This follows from Evans and Schulman's result in [85], which we will discuss further in subsection 5.4.4.

Lastly, in the context of the random DAG model studied in Theorem 5.1, the ensuing proposition illustrates that the problem of reconstruction using the information contained in just a single vertex, e.g. $X_{k,0}$, exhibits a similar phase transition phenomenon to that in Theorem 5.1.

Proposition 5.1 (Single Vertex Reconstruction). *Let $C(\delta, d)$ be the constant defined in (5.49) in section 5.5. For a random DAG model with $d \geq 3$, the following phase transition phenomenon occurs around δ_{maj}:*

1. *If $\delta \in (0, \delta_{\mathsf{maj}})$, the number of vertices per level satisfies $L_k \geq C(\delta, d)\log(k)$ for all sufficiently large k (depending on δ and d), and all Boolean processing functions are the majority rule (where ties are broken by outputting random bits), then reconstruction is possible in the sense that:*

$$\limsup_{k\to\infty} \mathbb{P}(X_{k,0} \neq X_{0,0}) < \frac{1}{2}$$

where we use a single vertex $X_{k,0}$ as the decoder at level k.

2. *If $\delta \in [\delta_{\mathsf{maj}}, \frac{1}{2})$, d is odd, and the number of vertices per level satisfies $\lim_{k\to\infty} L_k = \infty$ and $R_k \triangleq \inf_{n\geq k} L_n = O(d^{2k})$, then for all choices of Boolean processing functions (which may vary between vertices and be graph dependent), reconstruction is impossible in the sense that:*

$$\lim_{k\to\infty} \mathbb{E}\Big[\Big\|P^{+}_{X_{k,0}|G} - P^{-}_{X_{k,0}|G}\Big\|_{\mathsf{TV}}\Big] = 0$$

where $P^{+}_{X_{k,0}|G}$ and $P^{-}_{X_{k,0}|G}$ are the conditional distributions of $X_{k,0}$ given $\{X_{0,0} = 1, G\}$ and $\{X_{0,0} = 0, G\}$, respectively.

Proposition 5.1 is proved in appendix D.1. In particular, part 2 of Proposition 5.1 demonstrates that when $\delta \geq \delta_{\mathsf{maj}}$, the ML decoder based on a single vertex $X_{k,0}$ (with knowledge of the random DAG realization G) cannot reconstruct $X_{0,0}$ in all but a vanishing fraction of DAGs. It is worth mentioning that much like part 2 of Theorem 5.1, part 2 of Proposition 5.1 also implies that:

$$\liminf_{k\to\infty} \left\| P^{+}_{X_{k,0}|G} - P^{-}_{X_{k,0}|G} \right\|_{\mathsf{TV}} = 0 \quad G\text{-}a.s. \tag{5.22}$$

since applying Fatou's lemma to it yields $\mathbb{E}[\liminf_{k\to\infty} \|P^{+}_{X_{k,0}|G} - P^{-}_{X_{k,0}|G}\|_{\mathsf{TV}}] = 0$. Thus, if reconstruction is possible in the range $\delta \geq \delta_{\mathsf{maj}}$, the decoder should definitely use more than one vertex. This converse result relies on the aforementioned impossibility results on reliable computation. Specifically, the exact threshold δ_{maj} that determines whether or not reliable computation is possible using formulae is known for odd $d \geq 3$, cf. [86, 115]. Therefore, we can exploit such results to obtain a converse for odd $d \geq 3$ which holds for all choices of Boolean processing functions and at the critical value $\delta = \delta_{\mathsf{maj}}$ (although only for single vertex decoding). In contrast, when $d \geq 4$ is even, it is not even known whether such a critical threshold exists (as noted in [86, Section 7]), and hence, we cannot easily prove such converse results for even $d \geq 4$.[88]

We next present an immediate corollary of Theorem 5.1 which states that there exist constant indegree deterministic DAGs with $L_k = \Omega(\log(k))$ (i.e. $L_k \geq C(\delta, d)\log(k)$ for some large constant $C(\delta, d)$ and all sufficiently large k) such that reconstruction of the root bit is possible. Note that *deterministic DAGs* refer to Bayesian networks on specific realizations of G in the sequel. We will use the same notation as subsection 5.3.1 to analyze deterministic DAGs with the understanding that the randomness is engendered by $X_{0,0}$ and the edge BSCs, but not G. Formally, we have the following result which is proved in appendix D.2.

Corollary 5.1 (Existence of DAGs where Reconstruction is Possible). *For every indegree $d \geq 3$, every noise level $\delta \in (0, \delta_{\mathsf{maj}})$, and every sequence of level sizes satisfying $L_k \geq C(\delta, d)\log(k)$ for all sufficiently large k, there exists a deterministic DAG $\mathcal{G}$ with these parameters such that if we use majority rules as our Boolean processing functions, then there exists $\epsilon = \epsilon(\delta, d) > 0$ (that depends on δ and d) such that the probability of error in ML decoding is bounded away from $\frac{1}{2} - \epsilon$:*

$$\forall k \in \mathbb{N} \cup \{0\}, \;\; \mathbb{P}\Big(h^{k}_{\mathsf{ML}}(X_k, \mathcal{G}) \neq X_0\Big) \leq \frac{1}{2} - \epsilon$$

where $h^{k}_{\mathsf{ML}}(\cdot, \mathcal{G}) : \{0,1\}^{L_k} \to \{0,1\}$ denotes the ML decision rule at level k based on the full k-layer state X_k (given knowledge of the DAG $\mathcal{G}$).

[88] Note, however, that if all Boolean processing functions are the majority rule and the conditions of part 2 of Theorem 5.1 are satisfied, then part 2 of Theorem 5.1 implies (using the data processing inequality for TV distance and the bounded convergence theorem) that single vertex reconstruction is also impossible in the sense presented in part 2 of Proposition 5.1.

Since the critical threshold $\delta_{\mathsf{maj}} \to \frac{1}{2}$ as $d \to \infty$, a consequence of Corollary 5.1 is that for any $\delta \in (0, \frac{1}{2})$, any sufficiently large indegree d (that depends on δ), and any sequence of level sizes satisfying $L_k \geq C(\delta, d)\log(k)$ for all sufficiently large k, there exists a deterministic DAG $\mathcal{G}$ with these parameters and all majority processing functions such that reconstruction of the root bit is possible in the sense shown above.

Until now, we have restricted ourselves to the $d \geq 3$ case of the random DAG model. Our second main result considers the setting where the indegree of each vertex (except the root) is $d = 2$, because it is not immediately obvious that deterministic DAGs (for which reconstruction is possible) exist for $d = 2$. Indeed, it is not entirely clear which Boolean processing functions are good for "local error correction" in this scenario. We choose to fix all Boolean functions at even levels of the random DAG model to be the AND rule, and all Boolean functions at odd levels of the model to be the OR rule. We then prove that this random DAG model also exhibits a phase transition phenomenon around a critical threshold of:

$$\delta_{\mathsf{andor}} \triangleq \frac{3 - \sqrt{7}}{4}. \tag{5.23}$$

As before, the next theorem illustrates that for $\delta < \delta_{\mathsf{andor}}$, the "biased" majority decision rule $\hat{T}_k \triangleq \mathbb{1}\{\sigma_k \geq t\}$, where $t \in (0, 1)$ is defined in (5.68) in section 5.6, can asymptotically decode σ_0, but for $\delta > \delta_{\mathsf{andor}}$, the ML decision rule with knowledge of G cannot asymptotically decode σ_0. For simplicity, we only analyze this model at even levels in the achievability case.

Theorem 5.2 (Phase Transition in Random DAG Model with AND-OR Rule Processing). *Let $C(\delta)$ and $D(\delta)$ be the constants defined in (5.74) and (5.64) in section 5.6. For a random DAG model with $d = 2$, AND processing functions at even levels, and OR processing functions at odd levels, the following phase transition phenomenon occurs around δ_{andor}:*

1. *If $\delta \in (0, \delta_{\mathsf{andor}})$, and the number of vertices per level satisfies $L_k \geq C(\delta)\log(k)$ for all sufficiently large k (depending on δ), then reconstruction is possible in the sense that:*

$$\limsup_{k\to\infty} \mathbb{P}(\hat{T}_{2k} \neq \sigma_0) < \frac{1}{2}$$

where we use the decoder $\hat{T}_{2k} = \mathbb{1}\{\sigma_{2k} \geq t\}$ at level $2k$, which recovers the root bit by thresholding at the value $t \in (0, 1)$ in (5.68).

2. *If $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$, and the number of vertices per level satisfies $L_k = o\big(E(\delta)^{-\frac{k}{2}}\big)$ and $\liminf_{k\to\infty} L_k > \frac{2}{E(\delta)-D(\delta)}$ for any $E(\delta) \in (D(\delta), 1)$ (that depends on δ), then reconstruction is impossible in the sense of (5.7):*

$$\lim_{k\to\infty} \left\| P^{+}_{X_k|G} - P^{-}_{X_k|G} \right\|_{\mathsf{TV}} = 0 \quad G\text{-a.s.}$$

Theorem 5.2 is proved in section 5.6, and many of the remarks pertaining to Theorem 5.1 as well as the general intuition for Theorem 5.1 also hold for Theorem 5.2.

Furthermore, a proposition analogous to part 1 of Proposition 5.1 and a corollary analogous to Corollary 5.1 also hold here (but we omit explicit statements of these results for brevity).

It is straightforward to verify that the random DAG in Theorem 5.2 with alternating layers of AND and OR processing functions is equivalent to a random DAG with all NAND processing functions for the purposes of broadcasting.[89] Recall that in the discussion following Theorem 5.1, we noted how the critical threshold δ_{maj} was already known in the reliable computation literature (because it characterized when reliable computation is possible), cf. [86]. It turns out that δ_{andor} has also appeared in the reliable computation literature in a similar vein. In particular, although the existence of critical thresholds on δ for reliable computation using formulae of δ-noisy gates is not known for any even $d \geq 4$, the special case of $d = 2$ has been resolved. Indeed, Evans and Pippenger showed in [84] that reliable computation using formulae consisting of δ-noisy NAND gates is possible when $\delta < \delta_{\mathsf{andor}}$ and impossible when $\delta > \delta_{\mathsf{andor}}$. Moreover, Unger established in [279, 280] that reliable computation using formulae with general 2-input δ-noisy gates is impossible when $\delta \geq \delta_{\mathsf{andor}}$.

■ 5.4.2 Explicit Construction of Deterministic DAGs where Broadcasting is Possible

Although Corollary 5.1 illustrates the existence of DAGs where broadcasting (i.e. reconstruction of the root bit) is possible, it does not elucidate the structure of such DAGs. Moreover, Theorem 5.1 suggests that reconstruction on such deterministic DAGs should be possible using the algorithmically simple majority decision rule, but Corollary 5.1 is proved for the typically more complex ML decision rule. In this subsection, we address these deficiencies of Corollary 5.1 by presenting an explicit construction of deterministic bounded degree DAGs such that $L_k = \Theta(\log(k))$ and reconstruction of the root bit is possible using the majority decision rule.

Our construction is based on *regular bipartite lossless expander graphs*. Historically, the notion of an expander graph goes back to the work of Kolmogorov and Barzdin in [154]. Soon afterwards, Pinsker independently discovered such graphs and coined the term "expander graph" in [226].[90] Both [154] and [226, Lemma 1] prove the existence of expander graphs using probabilistic techniques. On the other hand, the first explicit construction of expander graphs appeared in [193], and more recently, lossless expander graphs were constructed using simpler ideas in [39]. We next define a pertinent variant of lossless expander graphs.

Consider a *d-regular bipartite graph* $B = (U, V, E)$, where U and V are two disjoint

[89] Indeed, we can introduce pairs of NOT gates into every edge of our DAG that goes from an AND gate to an OR gate without affecting the statistics of the model. Since an AND gate followed by a NOT gate and an OR gate whose inputs pass through NOT gates are both NAND gates, we obtain an equivalent model where all processing functions are NAND gates. We remark that analyzing this random DAG model with NAND processing functions yields a version of Theorem 5.2 with the same essential characteristics (albeit with possibly weaker conditions on L_k).

[90] In fact, expander graphs are called "expanding" graphs in [226].

sets of vertices such that $|U| = |V| = n \in \mathbb{N}$, every vertex in $U \cup V$ has degree $d \in \mathbb{N}$, and E is the set of undirected edges between U and V. Note that we allow multiple edges to exist between two vertices in B. For any subset of vertices $S \subseteq U$, we define the *neighborhood* of S as:

$$\Gamma(S) \triangleq \{v \in V : \exists u \in S,\, (u, v) \in E\} \tag{5.24}$$

which is the set of all vertices in V that are adjacent to some vertex in S. For any fraction $\alpha \in (0, 1)$ and any expansion factor $\beta > 0$, B is called an (α, β)-*expander graph* if for every subset of vertices $S \subseteq U$, we have:

$$|S| \leq \alpha n \quad \Rightarrow \quad |\Gamma(S)| \geq \beta |S| \,. \tag{5.25}$$

Note that we only require subsets of vertices in U to expand (not V). Intuitively, such expander graphs are sparse due to the d-regularity constraint, but have high connectivity due to the expansion property (5.25). Furthermore, when $\alpha \leq \frac{1}{d}$, the best expansion factor one can hope for is β as close as possible to d. Hence, $(\alpha, (1-\epsilon)d)$-expander graphs with $\alpha \leq \frac{1}{d}$ and very small $\epsilon > 0$ are known as *lossless* expander graphs [39, Section 1.1].

We utilize a slightly relaxed version of lossless expander graphs in our construction. In particular, using existing results from the literature, we establish in Corollary 5.2 of section 5.7 that for large values of the degree d and any sufficiently large n (depending on d), there exists a d-regular bipartite graph $B = (U, V, E)$ with $|U| = |V| = n$ such that for every subset of vertices $S \subseteq U$, we have:[91]

$$|S| = \frac{n}{d^{6/5}} \quad \Rightarrow \quad |\Gamma(S)| \geq (1 - \epsilon)\frac{n}{d^{1/5}} \;\text{ with }\; \epsilon = \frac{2}{d^{1/5}} \,. \tag{5.26}$$

Unlike (5.25), the expansion in (5.26) only holds for subsets $S \subseteq U$ with cardinality exactly $|S| = nd^{-6/5}$. However, we can still (loosely) perceive the graph B as a d-regular bipartite lossless (α, β)-expander graph with $\alpha = d^{-6/5}$ and $\beta = (1 - \epsilon)d$. (Strictly speaking, $nd^{-6/5}$ must be an integer, but we neglect this detail throughout our exposition for simplicity.) In the remainder of our discussion, we refer to graphs like B that satisfy (5.26) as *d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graphs* with abuse of standard nomenclature.

A d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graph B can be construed as representing two consecutive levels of a deterministic DAG upon which we are broadcasting. Indeed, we can make every edge in E directed by making them point from U to V, where U represents a particular level in the DAG and V the next level. In fact, we can construct deterministic DAGs where broadcasting is possible by concatenating several such d-regular bipartite lossless expander graphs together. The ensuing theorem details our DAG construction, and illustrates that reconstruction of the root bit is possible when we use majority Boolean processing functions and the majority decision rule $\hat{S}_k = \mathbb{1}\{\sigma_k \geq \frac{1}{2}\}$, where σ_k is defined in (5.2).

[91] We do not explicitly impose the constraint that $\epsilon = 2/d^{1/5} < 1$ because the constraint (5.27) in Theorem 5.3 implicitly ensures this.

Theorem 5.3 (DAG Construction using Expander Graphs). *Fix any noise level $\delta \in (0, \frac{1}{2})$, any sufficiently large odd degree $d = d(\delta) \geq 5$ (that depends on δ) satisfying:*

$$\frac{8}{d^{1/5}} + d^{6/5} \exp\left(-\frac{(1-2\delta)^2(d-4)^2}{8d}\right) \leq \frac{1}{2}, \tag{5.27}$$

and any sufficiently large constant $N = N(\delta) \in \mathbb{N}$ (that depends on δ) such that the constant $M \triangleq \exp(N/(4d^{12/5})) \geq 2$ and for every $n \geq N$, there exists a d-regular bipartite lossless $(d^{-6/5}, d-2d^{4/5})$-expander graph $B_n = (U_n, V_n, E_n)$ with $|U_n| = |V_n| = n$ that satisfies (5.26) for every subset $S \subseteq U_n$. Let the sequence of level sizes $\{L_k : k \in \mathbb{N} \cup \{0\}\}$ be given by $L_0 = 1$, $L_1 = N$, and:

$$\forall m \in \mathbb{N} \cup \{0\}, \ \ \forall k \in \mathbb{N} \cup \{0\} \ \text{ such that } \ M^{\lfloor 2^{m-1} \rfloor} < k \leq M^{2^m}, \ \ L_k = 2^m N \tag{5.28}$$

so that we have $L_k = \Theta(\log(k))$. Then, either in deterministic quasi-polynomial time $O(\exp(\Theta(\log(r)\log\log(r))))$, or if N additionally satisfies (5.96), in randomized poly-logarithmic time $O(\log(r)\log\log(r))$ with strictly positive success probability (5.98), we can construct the constituent expander graphs for levels $0, \ldots, r$ of an infinite deterministic DAG with level sizes $\{L_k : k \in \mathbb{N} \cup \{0\}\}$ defined above, indegrees bounded by d, outdegrees bounded by $2d$, and the following edge configuration:

1. *Every vertex in X_1 has one directed edge coming from $X_{0,0}$.*
2. *For every pair of consecutive levels k and $k+1$ such that $L_{k+1} = L_k$, the directed edges from X_k to X_{k+1} are given by the edges of B_{L_k}, where we identify the vertices in U_{L_k} with X_k and the vertices in V_{L_k} with X_{k+1}, respectively.*
3. *For every pair of consecutive levels k and $k+1$ such that $L_{k+1} = 2L_k$, we partition the vertices in X_{k+1} into two sets, $X^1_{k+1} = (X_{k+1,0}, \ldots, X_{k+1,L_k-1})$ and $X^2_{k+1} = (X_{k+1,L_k}, \ldots, X_{k+1,L_{k+1}-1})$, so that the directed edges from X_k to X^i_{k+1} are given by the edges of B_{L_k} for $i = 1, 2$, where we identify the vertices in U_{L_k} with X_k and the vertices in V_{L_k} with X^i_{k+1}, respectively, as before.*

Furthermore, for the Bayesian network defined on this infinite deterministic DAG with $X_{0,0} \sim \mathsf{Bernoulli}(\frac{1}{2})$, independent $\mathsf{BSC}(\delta)$ edges, all identity Boolean processing functions in level $k = 1$, and all majority rule Boolean processing functions in levels $k \geq 2$ (as defined in subsection 5.3.1), reconstruction is possible in the sense that:

$$\limsup_{k \to \infty} \mathbb{P}\left(\hat{S}_k \neq X_0\right) < \frac{1}{2}$$

where we use the majority decoder $\hat{S}_k = \mathbb{1}\{\sigma_k \geq \frac{1}{2}\}$ at level k.

Theorem 5.3 is proved in section 5.7. The proof of feasibility of reconstruction follows the same overarching strategy as the proof of Theorem 5.1, but obviously makes essential

use of the expansion property (5.26). We emphasize that Theorem 5.3 portrays that the constituent expander graphs of a deterministic DAG where broadcasting is possible can be constructed either in quasi-polynomial time or in randomized polylogarithmic time in the number of levels. Once the DAG is constructed however, reconstruction of the root bit is guaranteed to succeed using the majority decoder in the sense presented above. Finally, we note that the question of finding a deterministic polynomial time algorithm to construct DAGs where reconstruction is possible remains open.

■ 5.4.3 Results on 2D Regular Grids

Deterministic 2D regular grids are much harder to analyze than random DAG models due to the dependence between adjacent vertices in a given layer. As mentioned earlier, we analyze the setting where all Boolean processing functions in the 2D regular grid with two inputs are the same, and all Boolean processing functions in the 2D grid with one input are the *identity* rule. Our first result shows that reconstruction is impossible for all $\delta \in (0, \frac{1}{2})$ when AND processing functions are used.

Theorem 5.4 (Deterministic AND 2D Grid). *If $\delta \in (0, \frac{1}{2})$, and all Boolean processing functions with two inputs in the 2D regular grid are the AND rule, then reconstruction is impossible in the sense of (5.14):*

$$\lim_{k\to\infty} \left\| P^{+}_{X_k} - P^{-}_{X_k} \right\|_{\mathsf{TV}} = 0\,.$$

Theorem 5.4 is proved in section 5.8. The proof couples the 2D grid starting at $X_{0,0} = 0$ with the 2D grid starting at $X_{0,0} = 1$, and "runs" them together. Using a phase transition result concerning *bond percolation* on 2D grids, we show that we eventually reach a layer where the values of all vertices in the first grid equal the values of the corresponding vertices in the second grid. So, the two 2D grids "couple" almost surely regardless of their starting state. This implies that we cannot decode the starting state by looking at vertices in layer k as $k \to \infty$. We note that in order to prove that the two 2D grids "couple," we have to consider two different regimes of δ and provide separate arguments for each. The details of these arguments are presented in section 5.8.

Our second result shows that reconstruction is impossible for all $\delta \in (0, \frac{1}{2})$ when XOR processing functions are used.

Theorem 5.5 (Deterministic XOR 2D Grid). *If $\delta \in (0, \frac{1}{2})$, and all Boolean processing functions with two inputs in the 2D regular grid are the XOR rule, then reconstruction is impossible in the sense of (5.14):*

$$\lim_{k\to\infty} \left\| P^{+}_{X_k} - P^{-}_{X_k} \right\|_{\mathsf{TV}} = 0\,.$$

Theorem 5.5 is proved in section 5.9. In the XOR 2D grid, every vertex at level k can be written as a (binary) linear combination of the root bit and all the BSC noise

random variables in the grid up to level k. This linear relationship can be captured by a binary matrix. The main idea of the proof is to perceive this matrix as a parity check matrix of a linear code. The problem of inferring $X_{0,0}$ from X_k turns out to be equivalent to decoding the first bit of a codeword drawn uniformly from this code after observing a noisy version of the codeword. Basic facts from coding theory can then be used to complete the proof.

We remark that at first glance, Theorems 5.4 and 5.5 seem intuitively obvious from the random DAG model perspective. For example, consider the random DAG model with $d = 2$, $L_k = k + 1$, and all AND processing functions. Then, the conditional expectation function $g(\sigma) = \mathbb{E}[\sigma_k | \sigma_{k-1} = \sigma]$ has only one fixed point regardless of the value of $\delta \in (0, \frac{1}{2})$, and we intuitively expect σ_k to tend to this fixed point (which roughly captures the equilibrium between AND gates killing 1's and $\mathsf{BSC}(\delta)$'s producing new 1's) as $k \to \infty$. So, reconstruction is impossible in this random DAG model, which suggests that reconstruction is also impossible in the AND 2D grid. However, although Theorems 5.4 and 5.5 seem intuitively easy to understand in this way, we emphasize that this random DAG intuition does not capture the subtleties engendered by the regularity of the 2D grid. In fact, the random DAG intuition can even be somewhat misleading. Consider the random DAG model with $d = 2$, $L_k = k + 1$, and all NAND processing functions. This model was analyzed in Theorem 5.2, because using successive layers of AND and OR processing functions is equivalent to using all NAND processing functions. Theorem 5.2 portrays that reconstruction of the root bit is possible in a certain range of δ values. Yet, evidence from [127], which proves the ergodicity of 1D PCA with NAND gates, and numerical simulations strongly suggest that reconstruction is actually impossible for the 2D regular grid with NAND processing functions. Therefore, the 2D regular grid setting of Theorems 5.4 and 5.5 should be intuitively understood with caution. Indeed, as sections 5.8 and 5.9 illustrate, the proofs of Theorems 5.4 and 5.5 are nontrivial.

The impossibility of reconstruction in Theorems 5.4 and 5.5 also seems intuitively plausible due to the ergodicity results for numerous 1D PCA—see e.g. [107] and the references therein. However, there are two key differences between deterministic 2D regular grids and 1D PCA. Firstly, the main question in the study of 1D PCA is whether a given automaton is ergodic, i.e. whether the Markov process defined by it converges to a unique invariant probability measure on the configuration space for all initial configurations. This question of ergodicity is typically addressed by considering the convergence of finite-dimensional distributions over the sites (i.e. weak convergence). Hence, for many 1D PCA that have special characteristics (such as translation invariance, finite range, positivity, and attractiveness or monotonicity, cf. [107]), it suffices to consider the convergence of distributions on finite intervals (e.g. marginal distributions at given sites). In contrast to this setting, we are concerned with the stronger notion of convergence in TV distance. Indeed, Theorems 5.4 and 5.5 show that the TV distance between $P^+_{X_k}$ and $P^-_{X_k}$ vanishes as $k \to \infty$.

Secondly, since a 1D PCA has infinitely many sites, the problem of remembering a

bit in a 1D PCA (with binary state space) corresponds to distinguishing between the "all zeros" and "all ones" initial configurations. On the other hand, as mentioned in section 5.1, a 2D regular grid can be construed as a 1D PCA with boundary conditions; each level $k \in \mathbb{N} \cup \{0\}$ corresponds to an instance in discrete-time, and there are $L_k = k + 1$ sites at time k. Moreover, its initial configuration has only one copy of the initial bit as opposed to infinitely many copies. As a result, compared a 2D regular grid, a 1D PCA (without boundary conditions) intuitively appears to have a stronger separation between the two initial states as time progresses. The aforementioned boundary conditions form another barrier to translating results from the 1D PCA literature to 2D regular grids.

It is also worth mentioning that most results on 1D PCA pertain to the continuous-time setting—see e.g. [107, 175] and the references therein. This is because sites are updated one by one in a continuous-time automaton, but they are updated in parallel in a discrete-time automaton. So, the discrete-time setting is often harder to analyze. (Indeed, some of the only known discrete-time 1D PCA ergodicity results are in [108, Section 3], which outlines the proof of ergodicity of the 3-input majority vote model i.e. 1D PCA with 3-input majority gates,[92] and [127], which proves ergodicity of 1D PCA with NAND gates.) This is another reason why results from the 1D PCA literature cannot be easily transferred to our model.

■ 5.4.4 Further Discussion and Impossibility Results

In this subsection, we present and discuss some impossibility results pertaining to both deterministic and random DAGs. The first result illustrates that if the level sizes satisfy $L_k \leq \log(k)/(d \log(1/(2\delta)))$ for every sufficiently large k (i.e. L_k grows very "slowly"), then reconstruction is impossible regardless of the choices of Boolean processing functions and the choice of decision rule.

Proposition 5.2 (Slow Growth of Layers). *For any noise level $\delta \in (0, \frac{1}{2})$ and indegree $d \in \mathbb{N}$, if the number of vertices per level satisfies $L_k \leq \log(k)/(d \log(1/(2\delta)))$ for all sufficiently large k, then for all choices of Boolean processing functions (which may vary between vertices and be graph dependent), reconstruction is impossible in the sense that:*

1. *for a deterministic DAG:*

$$\lim_{k\to\infty} \left\| P^+_{X_k} - P^-_{X_k} \right\|_{\mathsf{TV}} = 0 \,.$$

2. *for a random DAG:*

$$\lim_{k\to\infty} \left\| P^+_{X_k|G} - P^-_{X_k|G} \right\|_{\mathsf{TV}} = 0 \quad \textit{pointwise}$$

[92] As Gray explains in [108, Section 3], his proof of ergodicity is not complete; he is "very detailed for certain parts of the argument and very sketchy in others" [108]. Although the references in [108] indicate that Gray was preparing a paper with the complete proof, this paper was never published to our knowledge. So, the ergodicity of 1D PCA with 3-input majority gates has not been rigorously established.

which means that the condition holds for every realization of the random DAG G.

This proposition is proved in appendix D.3. Part 1 of Proposition 5.2 illustrates that when L_k is sub-logarithmic, the ML decoder based on the entire k-layer state X_k with knowledge of the deterministic DAG fails to reconstruct the root bit. Similarly, part 2 of Proposition 5.2 shows that reconstruction is impossible for random DAGs even if the particular DAG realization G is known and the ML decoder can access X_k. Therefore, Proposition 5.2 illustrates that our assumption that $L_k \geq C \log(k)$, for some constant C (that depends on δ and d) and all sufficiently large k, for reconstruction to be possible in Theorems 5.1 and 5.2 is in fact necessary.

In contrast, consider a deterministic DAG with no restrictions (i.e. no bounded indegree assumption) except for the size of L_k. Then, each vertex at level k of this DAG is connected to all L_{k-1} vertices at level $k-1$. The next proposition illustrates that $L_k = \Theta(\sqrt{\log(k)})$ is the critical scaling of L_k in this scenario. In particular, reconstruction is possible when $L_k = \Omega(\sqrt{\log(k)})$ (i.e. $L_k \geq A(\delta)\sqrt{\log(k)}$ for some large constant $A(\delta)$ and all sufficiently large k), and reconstruction is impossible when $L_k = O(\sqrt{\log(k)})$ (i.e. $L_k \leq B(\delta)\sqrt{\log(k)}$ for some small constant $B(\delta)$ and all sufficiently large k). The proof of this result is deferred to appendix D.4.

Proposition 5.3 (Broadcasting in Unbounded Degree DAG Model). *Let $A(\delta)$ and $B(\delta)$ be the constants defined in (D.21) and (D.22) in appendix D.4. Consider a deterministic DAG $\mathcal{G}$ such that for every $k \in \mathbb{N}$, each vertex at level k has one incoming edge from all L_{k-1} vertices at level $k-1$. Then, for any noise level $\delta \in (0, \frac{1}{2})$, we have:*

1. *If the number of vertices per level satisfies $L_k \geq A(\delta)\sqrt{\log(k)}$ for all sufficiently large k, and all Boolean processing functions in $\mathcal{G}$ are the majority rule (where ties are broken by outputting 1), then reconstruction is possible in the sense that:*

$$\limsup_{k\to\infty} \mathbb{P}(\hat{S}_k \neq X_0) < \frac{1}{2}$$

where we use the majority decoder $\hat{S}_k = \mathbb{1}\{\sigma_k \geq \frac{1}{2}\}$ at level k.

2. *If the number of vertices per level satisfies $L_k \leq B(\delta)\sqrt{\log(k)}$ for all sufficiently large k, then for all choices of Boolean processing functions (which may vary between vertices), reconstruction is impossible in the sense of (5.14):*

$$\lim_{k\to\infty} \left\| P^+_{X_k} - P^-_{X_k} \right\|_{\mathsf{TV}} = 0\,.$$

The last impossibility result we present here is an important result from the reliable computation literature due to Evans and Schulman [85]. Evans and Schulman studied von Neumann's noisy computation model (which we briefly discussed in subsection 5.4.1), and established general conditions under which reconstruction is impossible in deterministic DAGs due to the decay of mutual information between X_0 and X_k. We

present a specialization of [85, Lemma 2] for our setting as Proposition 5.4 below. This proposition portrays that if L_k is sub-exponential and the parameters δ and d satisfy $(1-2\delta)^2 d < 1$, then reconstruction is impossible in deterministic DAGs regardless of the choices of Boolean processing functions and the choice of decision rule.

Proposition 5.4 (Decay of Mutual Information [85, Lemma 2]). *For any deterministic DAG model, we have:*

$$I(X_0; X_k) \leq \log(2)\, L_k \left((1-2\delta)^2 d\right)^k$$

where $L_k d^k$ is the total number of paths from X_0 to layer X_k, and $(1-2\delta)^{2k}$ can be construed as the overall contraction of mutual information along each path (cf. (2.54) and (2.59) in chapter 2). Therefore, if $(1-2\delta)^2 d < 1$ and $L_k = o(1/((1-2\delta)^2 d)^k)$, then for all choices of Boolean processing functions (which may vary between vertices), we have $\lim_{k\to\infty} I(X_0; X_k) = 0$, which implies via Pinsker's inequality that $\lim_{k\to\infty} \|P^+_{X_k} - P^-_{X_k}\|_{\mathsf{TV}} = 0$.

We make some pertinent remarks about this result. Firstly, Evans and Schulman's original analysis assumes that gates are noisy as opposed to edges (in accordance with von Neumann's setup), but the re-derivation of [85, Lemma 2] in [231, Corollary 7] illustrates that the result also holds for our model. In fact, the *site percolation* analysis in [231, Section 3] (which we will briefly delineate later) improves upon Evans and Schulman's estimate. Furthermore, this analysis illustrates that the bound in Proposition 5.4 also holds for all choices of random Boolean processing functions.

Secondly, while Proposition 5.4 holds for deterministic DAGs, we can easily extend it for random DAG models. Indeed, the random DAG model inherits the inequality in Proposition 5.4 pointwise:

$$I(X_0; X_k | G = \mathcal{G}) \leq \log(2)\, L_k \left((1-2\delta)^2 d\right)^k \tag{5.29}$$

for every realization of the random DAG $G = \mathcal{G}$, where $I(X_0; X_k | G = \mathcal{G})$ is the mutual information between X_0 and X_k computed using the joint distribution of X_0 and X_k given $G = \mathcal{G}$. This implies that if L_k is sub-exponential and $(1-2\delta)^2 d < 1$, then reconstruction based on X_k is impossible regardless of the choices of Boolean processing functions (which may vary between vertices and be graph dependent) and the choice of decision rule even if the decoder knows the particular random DAG realization, i.e. $\lim_{k\to\infty} \|P^+_{X_k|G} - P^-_{X_k|G}\|_{\mathsf{TV}} = 0$ pointwise (which trivially implies (5.7)). Taking expectations with respect to G in (5.29), we get:

$$I(X_0; X_k) \leq I(X_0; X_k | G) \leq \log(2)\, L_k \left((1-2\delta)^2 d\right)^k \tag{5.30}$$

where $I(X_0; X_k | G)$ is the conditional mutual information (i.e. the expected value of $I(X_0; X_k | G = \mathcal{G})$ with respect to G), and the first inequality follows from the chain rule

for mutual information (cf. Kolmogorov identity [230, Theorem 2.5]) and the fact that X_0 is independent of G. Since the second inequality in (5.30) implies (5.53), invoking the argument at the end of the proof of part 2 of Theorem 5.1 in section 5.5 also yields that reconstruction is impossible in the sense of (5.7) when L_k is sub-exponential and $(1-2\delta)^2 d < 1$. Thus, $\lim_{k\to\infty} I(X_0; X_k|G) = 0$ is a sufficient condition for (5.7). In contrast, the first inequality in (5.30) only yields the impossibility of reconstruction in the sense of (5.8) when L_k is sub-exponential and $(1-2\delta)^2 d < 1$.

Thirdly, Evans and Schulman's result in Proposition 5.4 provides an upper bound on the critical threshold of δ above which reconstruction of the root bit is impossible. Indeed, the condition, $(1-2\delta)^2 d < 1$, under which mutual information decays can be rewritten as (cf. the discussion in [85, p.2373]):

$$\delta_{\mathsf{ES}}(d) \triangleq \frac{1}{2} - \frac{1}{2\sqrt{d}} < \delta < \frac{1}{2} \tag{5.31}$$

and reconstruction is impossible for deterministic or random DAGs in this regime of δ provided L_k is sub-exponential. As a sanity check, we can verify that $\delta_{\mathsf{ES}}(2) = 0.14644... > 0.08856... = \delta_{\mathsf{andor}}$ in the context of Theorem 5.2, and $\delta_{\mathsf{ES}}(3) = 0.21132... > 0.16666... = \delta_{\mathsf{maj}}$ in the context of Theorem 5.1 with $d = 3$. Although $\delta_{\mathsf{ES}}(d)$ is a general upper bound on the critical threshold for reconstruction, in this chapter, it is not particularly useful because we analyze explicit processing functions and decision rules, and derive specific bounds that characterize the corresponding thresholds.

Fourthly, it is worth comparing $\delta_{\mathsf{ES}}(d)$ (which comes from a site percolation argument, cf. [231, Section 3]) to an upper bound on the critical threshold for reconstruction derived from bond percolation. To this end, consider the random DAG model, and recall that the $\mathsf{BSC}(\delta)$'s along each edge generate independent bits with probability 2δ (as shown in the proof of Proposition 5.2 in appendix D.3). So, we can perform bond percolation so that each edge is independently "removed" with probability 2δ. It can be shown by analyzing this bond percolation process that reconstruction is impossible (in a certain sense) when $\frac{1}{2} - \frac{1}{2d} < \delta < \frac{1}{2}$. Therefore, the Evans-Schulman upper bound of $\delta_{\mathsf{ES}}(d)$ is tighter than the bond percolation upper bound: $\delta_{\mathsf{ES}}(d) < \frac{1}{2} - \frac{1}{2d}$.

Finally, we briefly delineate how the site percolation approach in [231, Section 3] allows us to prove that reconstruction is impossible in the random DAG model for the $(1-2\delta)^2 d = 1$ case as well. Consider a site percolation process where each vertex $X_{k,j}$ (for $k \in \mathbb{N}$ and $j \in [L_k]$) is independently "open" with probability $(1-2\delta)^2$, and "closed" with probability $1-(1-2\delta)^2$. (Note that $X_{0,0}$ is open almost surely.) For every $k \in \mathbb{N}$, let p_k denote the probability that there is an "open connected path" from X_0 to X_k (i.e. there exist $j_1 \in [L_1], \ldots, j_k \in [L_k]$ such that $(X_{0,0}, X_{1,j_1}), (X_{1,j_1}, X_{2,j_2}), \ldots, (X_{k-1,j_{k-1}}, X_{k,j_k})$ are directed edges in the random DAG G and $X_{1,j_1}, \ldots, X_{k,j_k}$ are all open). It can be deduced from [231, Theorem 5] that for any $k \in \mathbb{N}$:

$$I(X_0; X_k|G) \le \log(2)\, p_k\,. \tag{5.32}$$

Next, for each $k \in \mathbb{N} \cup \{0\}$, define the random variable:

$$\lambda_k \triangleq \frac{1}{L_k} \sum_{j \in [L_k]} \mathbb{1}\{X_{k,j} \text{ is open and connected}\} \tag{5.33}$$

which is the proportion of open vertices at level k that are connected to the root by an open path. (Note that $\lambda_0 = 1$.) It is straightforward to verify (using Bernoulli's inequality) that for any $k \in \mathbb{N}$:

$$\mathbb{E}[\lambda_k | \lambda_{k-1}] = (1-2\delta)^2 \Big(1 - (1-\lambda_{k-1})^d\Big) \le (1-2\delta)^2 d \lambda_{k-1} \,. \tag{5.34}$$

Observe that by Markov's inequality and the recursion from (5.34):

$$\mathbb{E}[\lambda_k] \le (1-2\delta)^2 d \, \mathbb{E}[\lambda_{k-1}] \,, \tag{5.35}$$

we have:

$$p_k = \mathbb{P}\Big(\lambda_k \ge \frac{1}{L_k}\Big) \le L_k \mathbb{E}[\lambda_k] \le L_k \left((1-2\delta)^2 d\right)^k \tag{5.36}$$

which recovers Evans and Schulman's result (Proposition 5.4) in the context of the random DAG model. Indeed, if $(1-2\delta)^2 d < 1$ and $L_k = o(1/((1-2\delta)^2 d)^k)$, then $\lim_{k\to\infty} p_k = 0$, and as a result, $\lim_{k\to\infty} I(X_0; X_k|G) = 0$ by (5.32). On the other hand, when $(1-2\delta)^2 d = 1$, taking expectations and applying Jensen's inequality to the equality in (5.34) produces:

$$\mathbb{E}[\lambda_k] \le (1-2\delta)^2 \Big(1 - (1 - \mathbb{E}[\lambda_{k-1}])^d\Big) \,. \tag{5.37}$$

This implies that $\mathbb{E}[\lambda_k] \le F^{-1}(k)$ for every $k \in \mathbb{N} \cup \{0\}$ using the estimate in [229, Appendix A], where $F : [0,1] \to [0,\infty)$, $F(t) = \int_t^1 1/f(\tau)\, d\tau$ with $f : [0,1] \to [0,1]$, $f(t) = t - (1-2\delta)^2(1-(1-t)^d)$, and $F^{-1} : [0,\infty) \to [0,1]$ is well-defined. Since $f(t) \ge \frac{d-1}{2} t^2$ for all $t \in [0,1]$, it is straightforward to show that:

$$\mathbb{E}[\lambda_k] \le F^{-1}(k) \le \frac{2}{(d-1)k} \,. \tag{5.38}$$

Therefore, the Markov's inequality argument in (5.36) illustrates that if $(1-2\delta)^2 d = 1$ and $L_k = o(k)$, then $\lim_{k\to\infty} p_k = 0$ and reconstruction is impossible in the random DAG model due to (5.32). Furthermore, the condition on L_k can be improved to $L_k = O(k \log(k))$ using a more sophisticated Borel-Cantelli type of argument.

■ 5.5 Analysis of Majority Rule Processing in Random DAG Model

In this section, we prove Theorem 5.1. To this end, we first make some pertinent observations. Recall that we have a random DAG model with $d \ge 3$, and all Boolean functions are the majority rule, i.e. $f_k(x_1, \ldots, x_d) = \mathsf{majority}(x_1, \ldots, x_d)$ for every $k \in \mathbb{N}$. Note

that when the number of 1's is equal to the number of 0's, the majority rule outputs an independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit.[93] Suppose we are given that $\sigma_{k-1} = \sigma$ for any $k \in \mathbb{N}$. Then, for every $j \in [L_k]$, $X_{k,j} = \mathsf{majority}(Y_1, \ldots, Y_d)$ where $Y_1, \ldots, Y_d$ are i.i.d. $\mathsf{Bernoulli}(p)$ random variables with $p = \sigma * \delta$. Define the function $g : [0,1] \to [0,1]$ as follows:

$$g(\sigma) \triangleq \mathbb{E}[\mathsf{majority}(Y_1, \ldots, Y_d)] = \mathbb{P}\left(\sum_{i=1}^{d} Y_i > \frac{d}{2}\right) + \frac{1}{2}\mathbb{P}\left(\sum_{i=1}^{d} Y_i = \frac{d}{2}\right) \tag{5.39}$$

$$= \begin{cases} \displaystyle\sum_{i=\frac{d}{2}+1}^{d} \binom{d}{i} (\sigma * \delta)^i (1 - \sigma * \delta)^{d-i} + \frac{1}{2}\binom{d}{\frac{d}{2}} (\sigma * \delta)^{\frac{d}{2}} (1 - \sigma * \delta)^{\frac{d}{2}} & , \quad d \text{ even} \\ \displaystyle\sum_{i=\frac{d+1}{2}}^{d} \binom{d}{i} \underbrace{(\sigma * \delta)}_{p}{}^i \underbrace{(1 - \sigma * \delta)}_{1-p}{}^{d-i} & , \quad d \text{ odd} \end{cases} \tag{5.40}$$

which implies that $X_{k,j}$ are i.i.d. $\mathsf{Bernoulli}(g(\sigma))$ for $j \in [L_k]$, and $L_k \sigma_k \sim \mathsf{binomial}(L_k, g(\sigma))$, since we have:

$$\mathbb{P}(X_{k,j} = 1 | \sigma_{k-1} = \sigma) = \mathbb{E}[\sigma_k | \sigma_{k-1} = \sigma] = g(\sigma)\,. \tag{5.41}$$

To compute the first derivative of g, we follow the analysis in [210, Section 2]. Recall that a Boolean function $h : \{0,1\}^d \to \{0,1\}$ is *monotone non-decreasing* (respectively, *non-increasing*) if its value either increases (respectively, decreases) or remains the same whenever any of its input bits is flipped from 0 to 1. For any such monotone function $h : \{0,1\}^d \to \{0,1\}$, the *Margulis-Russo formula* states that [194, 240] (alternatively, see [113, Section 4.1]):

$$\begin{aligned} &\frac{d}{dp}\mathbb{E}[h(Y_1, \ldots, Y_d)] \\ &\quad = \sum_{i=1}^{d} \mathbb{E}[h(Y_1, \ldots, Y_{i-1}, 1, Y_{i+1}, \ldots, Y_d) - h(Y_1, \ldots, Y_{i-1}, 0, Y_{i+1}, \ldots, Y_d)]\,. \end{aligned} \tag{5.42}$$

Hence, since $h = \mathsf{majority}$ is a non-decreasing function, $g' : [0,1] \to \mathbb{R}$ is given by:

$$\begin{aligned} g'(\sigma) &= \frac{dp}{d\sigma}\frac{d}{dp}\mathbb{E}[h(Y_1, \ldots, Y_d)] \\ &= (1 - 2\delta) \sum_{i=1}^{d} \mathbb{E}[h(Y_1, \ldots, Y_{i-1}, 1, Y_{i+1}, \ldots, Y_d) - h(Y_1, \ldots, Y_{i-1}, 0, Y_{i+1}, \ldots, Y_d)] \\ &= (1 - 2\delta)\, d\, \mathbb{E}[h(1, Y_2, \ldots, Y_d) - h(0, Y_2, \ldots, Y_d)] \end{aligned}$$

[93] Although generating a random bit is a natural approach to breaking ties in the majority rule, this means that the rule is no longer purely deterministic when d is even.

$$= (1-2\delta)\, d\, \mathbb{P}(h(1, Y_2, \ldots, Y_d) = 1,\, h(0, Y_2, \ldots, Y_d) = 0) \tag{5.43}$$

$$\begin{aligned}
&= \begin{cases} (1-2\delta)\,\frac{d}{2}\left(\mathbb{P}\left(\sum_{i=2}^{d} Y_i = \frac{d}{2} - 1\right) + \mathbb{P}\left(\sum_{i=2}^{d} Y_i = \frac{d}{2}\right)\right) & , \quad d \text{ even} \\ (1-2\delta)\, d\, \mathbb{P}\left(\sum_{i=2}^{d} Y_i = \frac{d-1}{2}\right) & , \quad d \text{ odd} \end{cases} \\
&= \begin{cases} (1-2\delta)\,\frac{d}{2}\left(\binom{d-1}{\frac{d}{2}-1} p^{\frac{d}{2}-1}(1-p)^{\frac{d}{2}} + \binom{d-1}{\frac{d}{2}} p^{\frac{d}{2}}(1-p)^{\frac{d}{2}-1}\right) & , \quad d \text{ even} \\ (1-2\delta)\, d \binom{d-1}{\frac{d-1}{2}} p^{\frac{d-1}{2}}(1-p)^{\frac{d-1}{2}} & , \quad d \text{ odd} \end{cases} \\
&= \begin{cases} (1-2\delta)\,\frac{d}{4}\binom{d}{\frac{d}{2}} (p(1-p))^{\frac{d}{2}-1} & , \quad d \text{ even} \\ (1-2\delta)\,\frac{d+1}{2}\binom{d}{\frac{d+1}{2}} (p(1-p))^{\frac{d-1}{2}} & , \quad d \text{ odd} \end{cases} \\
&= \begin{cases} (1-2\delta)\,\frac{d}{4}\binom{d}{\frac{d}{2}} ((\sigma * \delta)(1-\sigma * \delta))^{\frac{d}{2}-1} & , \quad d \text{ even} \\ (1-2\delta)\,\frac{d+1}{2}\binom{d}{\frac{d+1}{2}} ((\sigma * \delta)(1-\sigma * \delta))^{\frac{d-1}{2}} & , \quad d \text{ odd} \end{cases}
\end{aligned} \tag{5.44}$$

where the second equality follows from $dp/d\sigma = 1 - 2\delta$ and (5.42), the third equality holds because $h = \mathsf{majority}$ is symmetric in its input bits, the fourth equality holds because $h = \mathsf{majority}$ is non-decreasing, and the fifth equality follows from the definition of the majority rule. Since $p \mapsto p(1-p)$ is increasing on $[0, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, 1]$, and $p = \sigma * \delta$ is linear in σ with derivative $1 - 2\delta > 0$ such that $p = \frac{1}{2}$ when $\sigma = \frac{1}{2}$, it is straightforward to verify from (5.44) that g' is positive on $[0,1]$, increasing on $[0, \frac{1}{2}]$, and decreasing on $[\frac{1}{2}, 1]$. As a result, g is increasing on $[0,1]$, convex on $[0, \frac{1}{2}]$, and concave on $[\frac{1}{2}, 1]$. Furthermore, the Lipschitz constant of g over $[0,1]$, or equivalently, the maximum value of g' over $[0,1]$ is:

$$D(\delta, d) \triangleq \max_{\sigma \in [0,1]} g'(\sigma) = g'\left(\frac{1}{2}\right) = (1-2\delta)\left(\frac{1}{2}\right)^{d-1} \left\lceil \frac{d}{2} \right\rceil \binom{d}{\lceil \frac{d}{2} \rceil} \tag{5.45}$$

regardless of whether d is even or odd.

There are two regimes of interest when we consider the contraction properties and fixed point structure of g. As defined in (5.17), let δ_{maj} be the critical noise level such that the Lipschitz constant $g'(\frac{1}{2})$ is equal to 1.[94] Then, in the $\delta \in (0, \delta_{\mathsf{maj}})$ regime, the Lipschitz constant $g'(\frac{1}{2})$ is greater than 1. Furthermore, since $g(\frac{1}{2}) = \frac{1}{2}$ and $g(1-\sigma) = 1 - g(\sigma)$ (which are straightforward to verify from (5.40)), the aforementioned properties

[94] We can also view δ_{maj} as the critical value such that the d-input majority gate with independent $\mathsf{BSC}(\delta)$'s at each input is an *amplifier* if and only if $\delta < \delta_{\mathsf{maj}}$. We refer readers to [253] for more information about amplifiers, and in particular, the relationship between amplifiers and reliable computation.

of g imply that g has three fixed points at $\sigma = 1-\hat{\sigma}, \frac{1}{2}, \hat{\sigma}$, where the largest fixed point of g is some $\hat{\sigma} \in (\frac{1}{2}, 1)$ that depends on δ (e.g. $\hat{\sigma} = \big(1+\sqrt{(1-6\delta)/(1-2\delta)^3}\big)/2$ when $d = 3$). In contrast, in the $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$ regime, the Lipschitz constant $g'(\frac{1}{2})$ is less than 1, and the only fixed point of g is $\sigma = \frac{1}{2}$. (We also mention that when $\delta = \delta_{\mathsf{maj}}$, g has only one fixed point at $\sigma = \frac{1}{2}$.)

Using these observations, we now prove Theorem 5.1.

Proof of Theorem 5.1. We begin by constructing a useful "monotone Markovian coupling" that will help establish both achievability and converse directions (see [170, Chapter 5] for basic definitions of Markovian couplings). Let $\{X_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{X_k^- : k \in \mathbb{N}\cup\{0\}\}$ denote versions of the Markov chain $\{X_k : k \in \mathbb{N}\cup\{0\}\}$ (i.e. with the same transition kernels) initialized at $X_0^+ = 1$ and $X_0^- = 0$, respectively. In particular, the marginal distributions of X_k^+ and X_k^- are $P_{X_k}^+$ and $P_{X_k}^-$, respectively. The monotone Markovian coupling $\{(X_k^-, X_k^+) : k \in \mathbb{N}\cup\{0\}\}$ between the Markov chains $\{X_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{X_k^- : k \in \mathbb{N}\cup\{0\}\}$ is generated as follows. First, condition on any random DAG realization $G = \mathcal{G}$. Recall that each edge $\mathsf{BSC}(\delta)$ of $\mathcal{G}$ either copies its input bit with probability $1-2\delta$, or produces an independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit with probability 2δ (as demonstrated in the proof of Proposition 5.2 in appendix D.3). Next, couple $\{X_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{X_k^- : k \in \mathbb{N}\cup\{0\}\}$ so that along any edge BSC of $\mathcal{G}$, say $(X_{k,j}, X_{k+1,i})$, $X_{k,j}^+$ and $X_{k,j}^-$ are either both copied with probability $1-2\delta$, or a shared independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit is produced with probability 2δ that becomes the value of both $X_{k+1,i}^+$ and $X_{k+1,i}^-$. In other words, $\{X_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{X_k^- : k \in \mathbb{N}\cup\{0\}\}$ "run" on the same underlying DAG $\mathcal{G}$ and have common BSCs. Hence, after averaging over all realizations of G, it is straightforward to verify that the Markovian coupling $\{(X_k^-, X_k^+) : k \in \mathbb{N}\cup\{0\}\}$ has the following properties:

1. The "marginal" Markov chains are $\{X_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{X_k^- : k \in \mathbb{N}\cup\{0\}\}$.
2. For every $k \in \mathbb{N}\cup\{0\}$, X_{k+1}^+ is conditionally independent of X_k^- given X_k^+, and X_{k+1}^- is conditionally independent of X_k^+ given X_k^-.
3. For every $k \in \mathbb{N}\cup\{0\}$ and every $j \in [L_k]$, $X_{k,j}^+ \geq X_{k,j}^-$ almost surely—this is the monotonicity property of the coupling.

In particular, the third property holds because $1 = X_{0,0}^+ \geq X_{0,0}^- = 0$ is true by assumption, each edge BSC preserves monotonicity (whether it copies its input or generates a new shared bit), and the majority processing functions are symmetric and monotone non-decreasing. In the sequel, probabilities of events that depend on the coupled vertex random variables $\{(X_{k,j}^-, X_{k,j}^+) : k \in \mathbb{N}\cup\{0\}, j \in [L_k]\}$ are defined with respect to this Markovian coupling. Note that this coupling also induces a monotone Markovian coupling $\{(\sigma_k^+, \sigma_k^-) : k \in \mathbb{N}\cup\{0\}\}$ between the Markov chains $\{\sigma_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{\sigma_k^- : k \in \mathbb{N}\cup\{0\}\}$ (where $\{\sigma_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{\sigma_k^- : k \in \mathbb{N}\cup\{0\}\}$ denote versions of the Markov chain $\{\sigma_k : k \in \mathbb{N}\cup\{0\}\}$ initialized at $\sigma_0^+ = 1$ and $\sigma_0^- = 0$, respectively) such that:

1. The "marginal" Markov chains are $\{\sigma_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{\sigma_k^- : k \in \mathbb{N}\cup\{0\}\}$.
2. For every $j > k \geq 1$, σ_j^+ is conditionally independent of $\sigma_0^-, \ldots, \sigma_k^-, \sigma_0^+, \ldots, \sigma_{k-1}^+$ given σ_k^+, and σ_j^- is conditionally independent of $\sigma_0^+, \ldots, \sigma_k^+, \sigma_0^-, \ldots, \sigma_{k-1}^-$ given σ_k^-.
3. For every $k \in \mathbb{N}\cup\{0\}$, $\sigma_k^+ \geq \sigma_k^-$ almost surely.

Part 1: We first prove that $\delta \in (0, \delta_{\mathsf{maj}})$ implies $\limsup_{k\to\infty} \mathbb{P}(\hat{S}_k \neq \sigma_0) < \frac{1}{2}$. To this end, we start by showing that there exists $\epsilon = \epsilon(\delta, d) > 0$ (that depends on δ and d) such that:

$$\forall k \in \mathbb{N},\ \ \mathbb{P}\Big(\sigma_k^+ \geq \hat{\sigma} - \epsilon \,\Big|\, \sigma_{k-1}^+ \geq \hat{\sigma} - \epsilon, A_{k,j}\Big) \geq 1 - \exp\Big(-2L_k\gamma(\epsilon)^2\Big) \tag{5.46}$$

where $\gamma(\epsilon) \triangleq g(\hat{\sigma}-\epsilon) - (\hat{\sigma}-\epsilon) > 0$, and $A_{k,j}$ with $0 \leq j < k$ is the non-zero probability event defined as:

$$A_{k,j} \triangleq \begin{cases} \{\sigma_j^- \leq 1-\hat{\sigma}+\epsilon\} & ,\quad 0 \leq j = k-1 \\ \{\sigma_{k-2}^+ \geq \hat{\sigma}-\epsilon, \ldots, \sigma_j^+ \geq \hat{\sigma}-\epsilon\} \cap \{\sigma_j^- \leq 1-\hat{\sigma}+\epsilon\} & ,\quad 0 \leq j \leq k-2 \end{cases}.$$

Since $g'(\hat{\sigma}) < 1$ and $g(\hat{\sigma}) = \hat{\sigma}$, $g(\hat{\sigma}-\epsilon) > \hat{\sigma}-\epsilon$ for sufficiently small $\epsilon > 0$. Fix any such $\epsilon > 0$ (which depends on δ and d because g depends on δ and d) such that $\gamma(\epsilon) > 0$. Recall that $L_k\sigma_k \sim \mathsf{binomial}(L_k, g(\sigma))$ given $\sigma_{k-1} = \sigma$. This implies that for every $k \in \mathbb{N}$ and every $0 \leq j < k$:

$$\begin{aligned}\mathbb{P}\Big(\sigma_k^+ < g\Big(\sigma_{k-1}^+\Big) - \gamma(\epsilon) \,\Big|\, \sigma_{k-1}^+ = \sigma, A_{k,j}\Big) &= \mathbb{P}(\sigma_k < g(\sigma_{k-1}) - \gamma(\epsilon) | \sigma_{k-1} = \sigma) \\ &\leq \exp\Big(-2L_k\gamma(\epsilon)^2\Big)\end{aligned}$$

where the equality follows from property 2 of our Markovian coupling $\{(\sigma_k^+, \sigma_k^-) : k \in \mathbb{N}\cup\{0\}\}$, and the inequality follows from (5.41) and Hoeffding's inequality (see Lemma C.4 in appendix C.2). As a result, we have:

$$\begin{aligned}\sum_{\sigma \geq \hat{\sigma}-\epsilon} \mathbb{P}\Big(\sigma_k^+ < g\Big(\sigma_{k-1}^+\Big) - \gamma(\epsilon) \,\Big|\, \sigma_{k-1}^+ = \sigma, A_{k,j}\Big)\, \mathbb{P}\Big(\sigma_{k-1}^+ = \sigma \,\Big|\, A_{k,j}\Big) \\ \leq \exp\Big(-2L_k\gamma(\epsilon)^2\Big) \sum_{\sigma \geq \hat{\sigma}-\epsilon} \mathbb{P}\Big(\sigma_{k-1}^+ = \sigma \,\Big|\, A_{k,j}\Big)\end{aligned}$$

which implies that:

$$\begin{aligned}\mathbb{P}\Big(\sigma_k^+ < g\Big(\sigma_{k-1}^+\Big) - \gamma(\epsilon), \sigma_{k-1}^+ \geq \hat{\sigma}-\epsilon \,\Big|\, A_{k,j}\Big) &\leq \exp\Big(-2L_k\gamma(\epsilon)^2\Big)\mathbb{P}\Big(\sigma_{k-1}^+ \geq \hat{\sigma}-\epsilon \,\Big|\, A_{k,j}\Big) \\ \mathbb{P}\Big(\sigma_k^+ < g\Big(\sigma_{k-1}^+\Big) - \gamma(\epsilon) \,\Big|\, \sigma_{k-1}^+ \geq \hat{\sigma}-\epsilon, A_{k,j}\Big) &\leq \exp\Big(-2L_k\gamma(\epsilon)^2\Big).\end{aligned}$$

Finally, notice that $\sigma_k^+ < \hat{\sigma} - \epsilon = g(\hat{\sigma} - \epsilon) - \gamma(\epsilon)$ implies that $\sigma_k^+ < g(\sigma_{k-1}^+) - \gamma(\epsilon)$ when $\sigma_{k-1}^+ \geq \hat{\sigma} - \epsilon$ (since g is non-decreasing and $g(\sigma_{k-1}^+) \geq g(\hat{\sigma} - \epsilon)$). This produces:

$$\mathbb{P}\Big(\sigma_k^+ < \hat{\sigma} - \epsilon \,\Big|\, \sigma_{k-1}^+ \geq \hat{\sigma} - \epsilon, A_{k,j}\Big) \leq \exp\Big(-2L_k\gamma(\epsilon)^2\Big)$$

which in turn establishes (5.46).

Now fix any $\tau > 0$, and choose a sufficiently large value $K = K(\epsilon, \tau) \in \mathbb{N}$ (that depends on ϵ and τ) such that:

$$\sum_{m=K+1}^{\infty} \exp\Big(-2L_m\gamma(\epsilon)^2\Big) \leq \tau\,. \tag{5.47}$$

Note that such K exists because $\sum_{m=1}^{\infty} 1/m^2 = \pi^2/6 < +\infty$, and for all sufficiently large m (depending on δ and d), we have:

$$\exp\Big(-2L_m\gamma(\epsilon)^2\Big) \leq \frac{1}{m^2} \quad \Leftrightarrow \quad L_m \geq \frac{\log(m)}{\gamma(\epsilon)^2}\,. \tag{5.48}$$

In (5.48), we use the assumption that $L_m \geq C(\delta, d)\log(m)$ for all sufficiently large m (depending on δ and d), where we define the constant $C(\delta, d)$ as:

$$C(\delta, d) \triangleq \frac{1}{\gamma(\epsilon(\delta, d))^2} > 0\,. \tag{5.49}$$

Using the continuity of probability measures, observe that:

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcap_{k>K}\Big\{\sigma_k^+ \geq \hat{\sigma} - \epsilon\Big\} \,\Bigg|\, \sigma_K^+ \geq \hat{\sigma} - \epsilon, \sigma_K^- \leq 1 - \hat{\sigma} + \epsilon\Bigg) & \\
&= \prod_{k>K} \mathbb{P}\Big(\sigma_k^+ \geq \hat{\sigma} - \epsilon \,\Big|\, \sigma_{k-1}^+ \geq \hat{\sigma} - \epsilon, A_{k,K}\Big) \\
&\geq \prod_{k>K} 1 - \exp\Big(-2L_k\gamma(\epsilon)^2\Big) \\
&\geq 1 - \sum_{k>K} \exp\Big(-2L_k\gamma(\epsilon)^2\Big) \\
&\geq 1 - \tau
\end{aligned}$$

where the first inequality follows from (5.46), the second inequality is straightforward to establish using induction, and the final inequality follows from (5.47). Therefore, we have for any $k > K$:

$$\mathbb{P}\Big(\sigma_k^+ \geq \hat{\sigma} - \epsilon \,\Big|\, \sigma_K^+ \geq \hat{\sigma} - \epsilon, \sigma_K^- \leq 1 - \hat{\sigma} + \epsilon\Big) \geq 1 - \tau\,. \tag{5.50}$$

Likewise, we can also prove mutatis mutandis that for any $k > K$:

$$\mathbb{P}\Big(\sigma_k^- \leq 1 - \hat{\sigma} + \epsilon \,\Big|\, \sigma_K^+ \geq \hat{\sigma} - \epsilon, \sigma_K^- \leq 1 - \hat{\sigma} + \epsilon\Big) \geq 1 - \tau \tag{5.51}$$

where the choices of ϵ, τ, and K in (5.51) are the same as those in (5.50) without loss of generality.

We need to show that $\limsup_{k\to\infty} \mathbb{P}(\hat{S}_k \neq \sigma_0) < \frac{1}{2}$, or equivalently, that there exists $\lambda > 0$ such that for all sufficiently large $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{P}\Big(\hat{S}_k \neq \sigma_0\Big) = \frac{1}{2}\mathbb{P}\Big(\hat{S}_k \neq \sigma_0 \,\Big|\, \sigma_0 = 1\Big) + \frac{1}{2}\mathbb{P}\Big(\hat{S}_k \neq \sigma_0 \,\Big|\, \sigma_0 = 0\Big) &\leq \frac{1-\lambda}{2} \\ \Leftrightarrow \quad \mathbb{P}\bigg(\sigma_k < \frac{1}{2} \,\bigg|\, \sigma_0 = 1\bigg) + \mathbb{P}\bigg(\sigma_k \geq \frac{1}{2} \,\bigg|\, \sigma_0 = 0\bigg) &\leq 1-\lambda \\ \Leftrightarrow \quad \mathbb{P}\Big(\sigma_k^+ \geq \frac{1}{2}\Big) - \mathbb{P}\Big(\sigma_k^- \geq \frac{1}{2}\Big) &\geq \lambda\,.\end{aligned}$$

To this end, let $E = \{\sigma_K^+ \geq \hat{\sigma} - \epsilon,\, \sigma_K^- \leq 1 - \hat{\sigma} + \epsilon\}$, and observe that for all $k > K$:

$$\begin{aligned}\mathbb{P}\Big(\sigma_k^+ \geq \frac{1}{2}\Big) - \mathbb{P}\Big(\sigma_k^- \geq \frac{1}{2}\Big) &= \mathbb{E}\bigg[\mathbb{1}\Big\{\sigma_k^+ \geq \frac{1}{2}\Big\} - \mathbb{1}\Big\{\sigma_k^- \geq \frac{1}{2}\Big\}\bigg] \\ &\geq \mathbb{E}\bigg[\bigg(\mathbb{1}\Big\{\sigma_k^+ \geq \frac{1}{2}\Big\} - \mathbb{1}\Big\{\sigma_k^- \geq \frac{1}{2}\Big\}\bigg)\mathbb{1}\{E\}\bigg] \\ &= \mathbb{E}\bigg[\mathbb{1}\Big\{\sigma_k^+ \geq \frac{1}{2}\Big\} - \mathbb{1}\Big\{\sigma_k^- \geq \frac{1}{2}\Big\}\,\bigg|\, E\bigg]\,\mathbb{P}(E) \\ &= \bigg(\mathbb{P}\Big(\sigma_k^+ \geq \frac{1}{2}\,\Big|\, E\Big) - \mathbb{P}\Big(\sigma_k^- \geq \frac{1}{2}\,\Big|\, E\Big)\bigg)\,\mathbb{P}(E) \\ &\geq \Big(\mathbb{P}\Big(\sigma_k^+ \geq \hat{\sigma} - \epsilon \,\Big|\, E\Big) - \mathbb{P}\Big(\sigma_k^- > 1 - \hat{\sigma} + \epsilon \,\Big|\, E\Big)\Big)\,\mathbb{P}(E) \\ &\geq (1-2\tau)\mathbb{P}(E) \triangleq \lambda > 0\end{aligned}$$

where the first inequality holds because $\mathbb{1}\{\sigma_k^+ \geq \frac{1}{2}\} - \mathbb{1}\{\sigma_k^- \geq \frac{1}{2}\} \geq 0$ almost surely due to the monotonicity (property 3) of the Markovian coupling $\{(\sigma_k^+, \sigma_k^-) : k \in \mathbb{N}\cup\{0\}\}$, the second inequality holds because $1 - \hat{\sigma} + \epsilon < \frac{1}{2} < \hat{\sigma} - \epsilon$ (since $\epsilon > 0$ is small), and the final inequality follows from (5.50) and (5.51). This completes the proof for the $\delta \in (0, \delta_{\mathsf{maj}})$ regime.

Part 2: We next prove that $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$ implies (5.7). First, notice that conditioned on any realization of the random DAG G, we have $X_{k,j}^+ \geq X_{k,j}^-$ almost surely for every $k \in \mathbb{N}\cup\{0\}$ and $j \in [L_k]$ (by construction of our coupling). Hence, conditioned on G, we obtain:

$$\begin{aligned}\Big\|P_{X_k|G}^+ - P_{X_k|G}^-\Big\|_{\mathsf{TV}} &\leq \mathbb{P}\Big(X_k^+ \neq X_k^- \Big| G\Big) \\ &= \mathbb{P}\Big(\exists j \in [L_k],\, X_{k,j}^+ \neq X_{k,j}^- \Big| G\Big) \\ &\leq \sum_{j=0}^{L_k-1} \mathbb{P}\Big(X_{k,j}^+ \neq X_{k,j}^- \Big| G\Big) \\ &= \mathbb{E}\bigg[\sum_{j=0}^{L_k-1} X_{k,j}^+ - X_{k,j}^- \,\bigg|\, G\bigg]\end{aligned}$$

$$= L_k \, \mathbb{E}\Big[\sigma_k^+ - \sigma_k^- \Big| G\Big]$$

where the first inequality follows from Dobrushin's maximal coupling representation of TV distance (see (2.6) in chapter 2), the third inequality follows from the union bound, and the fourth equality holds because $\mathbb{P}(X_{k,j}^+ \neq X_{k,j}^- | G) = \mathbb{P}(X_{k,j}^+ - X_{k,j}^- = 1 | G) = \mathbb{E}[X_{k,j}^+ - X_{k,j}^- | G]$ due to the monotonicity of our coupling. Then, taking expectations with respect to G yields:

$$\mathbb{E}\Big[\Big\| P_{X_k|G}^+ - P_{X_k|G}^- \Big\|_{\mathsf{TV}}\Big] \leq L_k \, \mathbb{E}\Big[\sigma_k^+ - \sigma_k^-\Big]. \tag{5.52}$$

We can bound $\mathbb{E}[\sigma_k^+ - \sigma_k^-]$ as follows. Firstly, we use the Lipschitz continuity of g (with Lipschitz constant $D(\delta, d)$) and the monotonicity of our coupling to get:

$$0 \leq \mathbb{E}\Big[\sigma_k^+ - \sigma_k^- \Big| \sigma_{k-1}^+, \sigma_{k-1}^-\Big] = g\Big(\sigma_{k-1}^+\Big) - g\Big(\sigma_{k-1}^-\Big) \leq D(\delta, d)\left(\sigma_{k-1}^+ - \sigma_{k-1}^-\right).$$

Then, we can take expectations with respect to $(\sigma_{k-1}^+, \sigma_{k-1}^-)$ on both sides of this inequality (and use the tower property on the left hand side) to obtain:

$$0 \leq \mathbb{E}\Big[\sigma_k^+ - \sigma_k^-\Big] \leq D(\delta, d) \, \mathbb{E}\Big[\sigma_{k-1}^+ - \sigma_{k-1}^-\Big].$$

Therefore, we recursively have:

$$0 \leq \mathbb{E}\Big[\sigma_k^+ - \sigma_k^-\Big] \leq D(\delta, d)^k$$

where we use the fact that $\mathbb{E}[\sigma_0^+ - \sigma_0^-] = 1$. Using (5.52) with this bound, we get:

$$\mathbb{E}\Big[\Big\| P_{X_k|G}^+ - P_{X_k|G}^- \Big\|_{\mathsf{TV}}\Big] \leq L_k D(\delta, d)^k$$

where letting $k \to \infty$ yields:

$$\lim_{k \to \infty} \mathbb{E}\Big[\Big\| P_{X_k|G}^+ - P_{X_k|G}^- \Big\|_{\mathsf{TV}}\Big] = 0 \tag{5.53}$$

because $L_k = o(D(\delta, d)^{-k})$ by assumption. (It is worth mentioning that although $L_k = o(D(\delta, d)^{-k})$ in this regime, it can diverge to infinity because the Lipschitz constant $D(\delta, d) < 1$.)

Finally, observe that $\|P_{X_k|G}^+ - P_{X_k|G}^-\|_{\mathsf{TV}} \in [0, 1]$ forms a non-increasing sequence in k for every realization of the random DAG G (since $\{X_k : k \in \mathbb{N} \cup \{0\}\}$ forms a Markov chain given G, and the data processing inequality for TV distance yields the desired monotonicity). Hence, the pointwise limit (over realizations of G) random variable, $\lim_{k\to\infty} \|P_{X_k|G}^+ - P_{X_k|G}^-\|_{\mathsf{TV}} \in [0, 1]$, has mean:

$$\mathbb{E}\left[\lim_{k \to \infty} \Big\| P_{X_k|G}^+ - P_{X_k|G}^- \Big\|_{\mathsf{TV}}\right] = 0$$

due to (5.53) and the bounded convergence theorem. Since a non-negative random variable that has zero mean must be equal to zero almost surely, we have (5.7):

$$\lim_{k\to\infty} \left\| P^{+}_{X_k|G} - P^{-}_{X_k|G} \right\|_{\mathsf{TV}} = 0 \quad G\text{-}a.s.$$

This completes the proof. ■

Finally, the next proposition portrays that the Markov chain $\{\sigma_k : k \in \mathbb{N} \cup \{0\}\}$ converges almost surely when $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$, $L_k = \omega(\log(k))$, and all processing functions are majority.

Proposition 5.5 (Majority Random DAG Model Almost Sure Convergence). *If $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$ and $L_k = \omega(\log(k))$, then $\lim_{k\to\infty} \sigma_k = \frac{1}{2}$ almost surely.*

Proposition 5.5 is proved in appendix D.5. It can be construed as a "weak" impossibility result since it demonstrates that the average number of 1's tends to $\frac{1}{2}$ in the $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$ regime regardless of the initial state of the Markov chain $\{\sigma_k : k \in \mathbb{N} \cup \{0\}\}$.

■ 5.6 Analysis of AND-OR Rule Processing in Random DAG Model

In this section, we prove Theorem 5.2. As before, we begin by making some pertinent observations. Recall that we have a random DAG model with $d = 2$, and all Boolean functions at even levels are the AND rule, and all Boolean functions at odd levels are the OR rule, i.e. $f_k(x_1, x_2) = x_1 \wedge x_2$ for every $k \in 2\mathbb{N}$, and $f_k(x_1, x_2) = x_1 \vee x_2$ for every $k \in \mathbb{N}\backslash 2\mathbb{N}$. Suppose we are given that $\sigma_{k-1} = \sigma$ for any $k \in \mathbb{N}$. Then, for every $j \in [L_k]$:

$$X_{k,j} = \begin{cases} \mathsf{Bernoulli}(\sigma * \delta) \wedge \mathsf{Bernoulli}(\sigma * \delta) & , \quad k \text{ even} \\ \mathsf{Bernoulli}(\sigma * \delta) \vee \mathsf{Bernoulli}(\sigma * \delta) & , \quad k \text{ odd} \end{cases} \tag{5.54}$$

for two i.i.d. Bernoulli random variables. Since we have:

$$\mathbb{P}(X_{k,j} = 1 | \sigma_{k-1} = \sigma) = \begin{cases} (\sigma * \delta)^2 & , \quad k \text{ even} \\ 1 - (1 - \sigma * \delta)^2 & , \quad k \text{ odd} \end{cases} \tag{5.55}$$

$$= \mathbb{E}[\sigma_k | \sigma_{k-1} = \sigma] \,, \tag{5.56}$$

$X_{k,j}$ are i.i.d. $\mathsf{Bernoulli}(g_{k \,(\mathsf{mod}\, 2)}(\sigma))$ for $j \in [L_k]$, and $L_k \sigma_k \sim \mathsf{binomial}(L_k, g_{k \,(\mathsf{mod}\, 2)}(\sigma))$, where we define $g_0 : [0,1] \to [0,1]$ and $g_1 : [0,1] \to [0,1]$ as:

$$g_0(\sigma) \triangleq (\sigma * \delta)^2 \,, \tag{5.57}$$

$$g_1(\sigma) \triangleq 1 - (1 - \sigma * \delta)^2 = 2(\sigma * \delta) - (\sigma * \delta)^2 \,. \tag{5.58}$$

The derivatives of g_0 and g_1 are:

$$g_0'(\sigma) = 2(1 - 2\delta)(\sigma * \delta) \geq 0 \,, \tag{5.59}$$

$$g_1'(\sigma) = 2(1 - 2\delta)(1 - \sigma * \delta) \geq 0 \,. \tag{5.60}$$

Consider the composition of g_0 and g_1, $g \triangleq g_0 \circ g_1 : [0,1] \to [0,1]$, given by:

$$g(\sigma) = \Big(\Big(2(\sigma * \delta) - (\sigma * \delta)^2\Big) * \delta\Big)^2 \tag{5.61}$$

which has derivative $g' : [0,1] \to \mathbb{R}$ given by:

$$\begin{aligned} g'(\sigma) &= g_0'(g_1(\sigma))g_1'(\sigma) \\ &= 4(1-2\delta)^2(g_1(\sigma) * \delta)(1 - \sigma * \delta) \geq 0\,. \end{aligned} \tag{5.62}$$

This derivative is a cubic function of σ with maximum value:

$$D(\delta) \triangleq \max_{\sigma \in [0,1]} g'(\sigma) = \begin{cases} g'\Big(\frac{1-\delta}{1-2\delta} - \sqrt{\frac{1-\delta}{3(1-2\delta)^3}}\Big) & , \quad \delta \in \Big(0, \frac{9-\sqrt{33}}{12}\Big] \\ g'(0) & , \quad \delta \in \Big(\frac{9-\sqrt{33}}{12}, \frac{1}{2}\Big) \end{cases} \tag{5.63}$$

$$= \begin{cases} \Big(\frac{4(1-\delta)(1-2\delta)}{3}\Big)^{\frac{3}{2}} & , \quad \delta \in \Big(0, \frac{9-\sqrt{33}}{12}\Big] \\ 4\delta(1-\delta)^2(1-2\delta)^2(3-2\delta) < 1 & , \quad \delta \in \Big(\frac{9-\sqrt{33}}{12}, \frac{1}{2}\Big) \end{cases} \tag{5.64}$$

which follows from standard calculus and algebraic manipulations, and Wolfram Mathematica computations. Hence, $D(\delta)$ in (5.64) is the Lipschitz constant of g over $[0,1]$. Since $4(1-\delta)(1-2\delta)/3 \in (0,1) \Leftrightarrow \delta \in ((3-\sqrt{7})/4, (9-\sqrt{33})/12]$, $D(\delta) < 1$ if and only if $\delta \in ((3-\sqrt{7})/4, 1/2)$. Moreover, $D(\delta) > 1$ if and only if $\delta \in (0, (3-\sqrt{7})/4)$ (and $D(\delta) = 1$ when $\delta = (3-\sqrt{7})/4$).

We next summarize the fixed point structure of g. Solving the equation $g(\sigma) = \sigma$ in Wolfram Mathematica produces:

$$\sigma = \frac{1 - 6\delta + 4\delta^2 \pm \sqrt{1 - 12\delta + 8\delta^2}}{2(1-2\delta)^2}, \frac{3 - 6\delta + 4\delta^2 \pm \sqrt{5 - 12\delta + 8\delta^2}}{2(1-2\delta)^2} \tag{5.65}$$

where the first pair is real when $\delta \in [0, (3-\sqrt{7})/4]$, and the second pair is always real. From these solutions, it is straightforward to verify that the only fixed points of g in the interval $[0,1]$ are:

$$t_0 \triangleq \frac{2(1-\delta)(1-2\delta) - 1 - \sqrt{4(1-\delta)(1-2\delta) - 3}}{2(1-2\delta)^2} \tag{5.66}$$

$$t_1 \triangleq \frac{2(1-\delta)(1-2\delta) - 1 + \sqrt{4(1-\delta)(1-2\delta) - 3}}{2(1-2\delta)^2} \tag{5.67}$$

$$t \triangleq \frac{2(1-\delta)(1-2\delta) + 1 - \sqrt{4(1-\delta)(1-2\delta) + 1}}{2(1-2\delta)^2} \tag{5.68}$$

where t_0 and t_1 are valid when $\delta \in (0, (3-\sqrt{7})/4]$. These fixed points satisfy $t_0 = t_1 = t$ when $\delta = (3-\sqrt{7})/4$, and $t_0 = 0$, $t_1 = 1$ when $\delta = 0$. Furthermore, observe that:

$$t_1 - t = \frac{\sqrt{a} + \sqrt{a+4} - 2}{2(1-2\delta)^2} > 0 \quad \text{and} \quad t - t_0 = \frac{\sqrt{a} - \sqrt{a+4} + 2}{2(1-2\delta)^2} > 0 \tag{5.69}$$

where $a = 4(1-\delta)(1-2\delta) - 3 > 0$ for $\delta \in (0, (3-\sqrt{7})/4)$, $t_1 - t > 0$ because $x \mapsto \sqrt{x}$ is strictly increasing ($\Rightarrow \sqrt{a} + \sqrt{a+4} > 2$), and $t - t_0 > 0$ because $x \mapsto \sqrt{x}$ is strictly subadditive ($\Rightarrow \sqrt{a}+2 > \sqrt{a+4}$). Hence, $0 < t_0 < t < t_1 < 1$ when $\delta \in (0, (3-\sqrt{7})/4)$.

Therefore, there are again two regimes of interest. Define the critical threshold $\delta_{\mathsf{andor}} \triangleq \frac{3-\sqrt{7}}{4}$. In the regime $\delta \in (0, \delta_{\mathsf{andor}})$, g has three fixed points $0 < t_0 < t < t_1 < 1$, and $D(\delta) > 1$. In contrast, in the regime $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$, g has only one fixed point at $t \in (0,1)$, and $D(\delta) < 1$.

We now prove Theorem 5.2. (The proof closely resembles the proof of Theorem 5.1 in section 5.5.)

Proof of Theorem 5.2. As in the proof of Theorem 5.1, we begin by constructing a monotone Markovian coupling $\{(X_k^-, X_k^+) : k \in \mathbb{N} \cup \{0\}\}$ between the Markov chains $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$ (which are versions of the Markov chain $\{X_k : k \in \mathbb{N} \cup \{0\}\}$ initialized at $X_0^+ = 1$ and $X_0^- = 0$, respectively), and this coupling induces a monotone Markovian coupling $\{(\sigma_k^+, \sigma_k^-) : k \in \mathbb{N} \cup \{0\}\}$ between the Markov chains $\{\sigma_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{\sigma_k^- : k \in \mathbb{N} \cup \{0\}\}$ (which are versions of the Markov chain $\{\sigma_k : k \in \mathbb{N} \cup \{0\}\}$ initialized at $\sigma_0^+ = 1$ and $\sigma_0^- = 0$, respectively). This monotone Markovian coupling satisfies the following properties:

1. The "marginal" Markov chains are $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$.
2. For every $k \in \mathbb{N} \cup \{0\}$, X_{k+1}^+ is conditionally independent of X_k^- given X_k^+, and X_{k+1}^- is conditionally independent of X_k^+ given X_k^-.
3. For every $j > k \geq 1$, σ_j^+ is conditionally independent of $\sigma_0^-, \ldots, \sigma_k^-, \sigma_0^+, \ldots, \sigma_{k-1}^+$ given σ_k^+, and σ_j^- is conditionally independent of $\sigma_0^+, \ldots, \sigma_k^+, \sigma_0^-, \ldots, \sigma_{k-1}^-$ given σ_k^-.
4. For every $k \in \mathbb{N} \cup \{0\}$ and every $j \in [L_k]$, $X_{k,j}^+ \geq X_{k,j}^-$ almost surely.
5. Due to the previous property, $\sigma_k^+ \geq \sigma_k^-$ almost surely for every $k \in \mathbb{N} \cup \{0\}$.

As before, the fourth property above holds because $1 = X_{0,0}^+ \geq X_{0,0}^- = 0$ is true by assumption, each edge BSC preserves monotonicity, and the AND and OR processing functions are symmetric and monotone non-decreasing.

Part 1: We first prove that $\delta \in (0, \delta_{\mathsf{andor}})$ implies $\limsup_{k\to\infty} \mathbb{P}(\hat{T}_{2k} \neq \sigma_0) < \frac{1}{2}$. To this end, we start by establishing that there exists $\epsilon = \epsilon(\delta) > 0$ (that depends on δ) such that:

$$\forall k \in \mathbb{N}, \;\; \mathbb{P}\Big(\sigma_{2k}^+ \geq t_1 - \epsilon \,\Big|\, \sigma_{2k-2}^+ \geq t_1 - \epsilon, A_{k,j}\Big) \geq 1 - 4\exp\left(-\frac{\hat{L}_k \gamma(\epsilon)^2}{8}\right) \tag{5.70}$$

where $\hat{L}_k \triangleq \min\{L_{2k}, L_{2k-1}\}$ for $k \in \mathbb{N}$, $\gamma(\epsilon) \triangleq g(t_1 - \epsilon) - (t_1 - \epsilon) > 0$, and $A_{k,j}$ is the non-zero probability event defined as:

$$A_{k,j} \triangleq \begin{cases} \{\sigma_{2j}^- \leq t_0 + \epsilon\} & , \, 0 \leq j = k-1 \\ \{\sigma_{2k-4}^+ \geq t_1 - \epsilon, \sigma_{2k-6}^+ \geq t_1 - \epsilon, \ldots, \sigma_{2j}^+ \geq t_1 - \epsilon\} \cap \{\sigma_{2j}^- \leq t_0 + \epsilon\} & , \, 0 \leq j \leq k-2 \end{cases}$$

where $0 \leq j < k$. Since $g'(t_1) = 4\delta(3-2\delta) < 1$ and $g(t_1) = t_1$, $g(t_1 - \epsilon) > t_1 - \epsilon$ for sufficiently small $\epsilon > 0$. Fix any such $\epsilon > 0$ (which depends on δ because g depends on δ) such that $\gamma(\epsilon) > 0$. Observe that for every $k \in \mathbb{N}$ and $\xi > 0$, we have:

$$\begin{aligned}
&\mathbb{P}(|\sigma_{2k} - g(\sigma_{2k-2})| > \xi | \sigma_{2k-2} = \sigma) \\
&\leq \mathbb{P}(|\sigma_{2k} - g_0(\sigma_{2k-1})| + |g_0(\sigma_{2k-1}) - g_0(g_1(\sigma_{2k-2}))| > \xi | \sigma_{2k-2} = \sigma) \\
&\leq \mathbb{P}(|\sigma_{2k} - g_0(\sigma_{2k-1})| + 2(1-\delta)(1-2\delta)|\sigma_{2k-1} - g_1(\sigma_{2k-2})| > \xi | \sigma_{2k-2} = \sigma) \\
&\leq \mathbb{P}\left(\left\{|\sigma_{2k} - g_0(\sigma_{2k-1})| > \frac{\xi}{2}\right\} \cup \left\{2(1-\delta)(1-2\delta)|\sigma_{2k-1} - g_1(\sigma_{2k-2})| > \frac{\xi}{2}\right\} \middle| \sigma_{2k-2} = \sigma\right) \\
&\leq \mathbb{P}\left(|\sigma_{2k} - g_0(\sigma_{2k-1})| > \frac{\xi}{2} \middle| \sigma_{2k-2} = \sigma \right) \\
&\qquad\qquad + \mathbb{P}\left(|\sigma_{2k-1} - g_1(\sigma_{2k-2})| > \frac{\xi}{4(1-\delta)(1-2\delta)} \middle| \sigma_{2k-2} = \sigma \right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(|\sigma_{2k} - g_0(\sigma_{2k-1})| > \frac{\xi}{2} \middle| \sigma_{2k-1} \right) \middle| \sigma_{2k-2} = \sigma \right] + 2\exp\left(-\frac{L_{2k-1}\xi^2}{8(1-\delta)^2(1-2\delta)^2} \right) \\
&\leq 2\exp\left(-\frac{L_{2k}\xi^2}{2} \right) + 2\exp\left(-\frac{L_{2k-1}\xi^2}{8(1-\delta)^2(1-2\delta)^2} \right) \\
&\leq 4\exp\left(-\frac{\hat{L}_k \xi^2}{8} \right)
\end{aligned} \tag{5.71}$$

where the first inequality follows from the triangle inequality and the fact that $g = g_0 \circ g_1$, the second inequality holds because the Lipschitz constant of g_0 on $[0,1]$ is $\max_{\sigma \in [0,1]} g_0'(\sigma) = g_0'(1) = 2(1-\delta)(1-2\delta)$ using (5.59), the fourth inequality follows from the union bound, the fifth and sixth inequalities follow from the Markov property, Hoeffding's inequality (see Lemma C.4 in appendix C.2), and the fact that $L_k \sigma_k \sim \mathsf{binomial}(L_k, g_{k \,(\mathrm{mod}\, 2)}(\sigma))$ given $\sigma_{k-1} = \sigma$, and the final inequality holds because $(1-\delta)^2(1-2\delta)^2 \leq 1$. Hence, for any $k \in \mathbb{N}$ and any $0 \leq j < k$, we have:

$$\begin{aligned}
\mathbb{P}\left(\sigma_{2k}^+ < g\left(\sigma_{2k-2}^+\right) - \gamma(\epsilon) \,\middle|\, \sigma_{2k-2}^+ = \sigma, A_{k,j} \right) &= \mathbb{P}(\,\sigma_{2k} < g(\sigma_{2k-2}) - \gamma(\epsilon) \,|\, \sigma_{2k-2} = \sigma) \\
&\leq \mathbb{P}(|\sigma_{2k} - g(\sigma_{2k-2})| > \gamma(\epsilon) | \sigma_{2k-2} = \sigma) \\
&\leq 4\exp\left(-\frac{\hat{L}_k \gamma(\epsilon)^2}{8} \right)
\end{aligned}$$

where the first equality follows from property 3 of the Markovian coupling, and the final inequality follows from (5.71). As shown in the proof of Theorem 5.1, this produces:

$$\begin{aligned}
\mathbb{P}\left(\sigma_{2k}^+ < g\left(\sigma_{2k-2}^+\right) - \gamma(\epsilon) \,\middle|\, \sigma_{2k-2}^+ \geq t_1 - \epsilon, A_{k,j} \right) &\leq 4\exp\left(-\frac{\hat{L}_k \gamma(\epsilon)^2}{8} \right) \\
\mathbb{P}\left(\sigma_{2k}^+ < t_1 - \epsilon \,\middle|\, \sigma_{2k-2}^+ \geq t_1 - \epsilon, A_{k,j} \right) &\leq 4\exp\left(-\frac{\hat{L}_k \gamma(\epsilon)^2}{8} \right)
\end{aligned}$$

where the second inequality follows from the first because $\sigma^+_{2k} < t_1 - \epsilon = g(t_1-\epsilon) - \gamma(\epsilon)$ implies that $\sigma^+_{2k} < g(\sigma^+_{2k-2}) - \gamma(\epsilon)$ when $\sigma^+_{2k-2} \geq t_1 - \epsilon$ (since g is non-decreasing and $g(\sigma^+_{2k-2}) \geq g(t_1-\epsilon)$). This proves (5.70).

Now fix any $\tau > 0$, and choose a sufficiently large even integer $K = K(\epsilon, \tau) \in 2\mathbb{N}$ (that depends on ϵ and τ) such that:

$$4 \sum_{m=\frac{K}{2}+1}^{\infty} \exp\left(-\frac{\hat{L}_m \gamma(\epsilon)^2}{8}\right) \leq \tau\,. \tag{5.72}$$

Note that such K exists because $\sum_{m=1}^{\infty} 1/(2m-1)^2 \leq 1 + \sum_{m=2}^{\infty} 1/(2m-2)^2 = 1 + (\pi^2/24) < +\infty$, and for sufficiently large m (depending on δ), we have:

$$\exp\left(-\frac{\hat{L}_m \gamma(\epsilon)^2}{8}\right) \leq \frac{1}{(2m-1)^2} \quad \Leftrightarrow \quad \hat{L}_m \geq \frac{16\log(2m-1)}{\gamma(\epsilon)^2}\,. \tag{5.73}$$

As before, in (5.73), we use the assumption that $L_m \geq C(\delta)\log(m)$ for all sufficiently large m (depending on δ), where we define the constant $C(\delta)$ as:

$$C(\delta) \triangleq \frac{16}{\gamma(\epsilon(\delta))^2} > 0\,. \tag{5.74}$$

Using the continuity of probability measures, observe that:

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{k>\frac{K}{2}} \left\{\sigma^+_{2k} \geq t_1 - \epsilon\right\} \,\middle|\, \sigma^+_K \geq t_1 - \epsilon, \sigma^-_K \leq t_0 + \epsilon\right) & \\
&= \prod_{k>\frac{K}{2}} \mathbb{P}\Big(\sigma^+_{2k} \geq t_1 - \epsilon \,\Big|\, \sigma^+_{2k-2} \geq t_1 - \epsilon, A_{k,\frac{K}{2}}\Big) \\
&\geq \prod_{k>\frac{K}{2}} 1 - 4\exp\left(-\frac{\hat{L}_k \gamma(\epsilon)^2}{8}\right) \\
&\geq 1 - 4\sum_{k>\frac{K}{2}} \exp\left(-\frac{\hat{L}_k \gamma(\epsilon)^2}{8}\right) \\
&\geq 1 - \tau
\end{aligned}$$

where the first inequality follows from (5.70), and the final inequality follows from (5.72). Therefore, we have for any $k > \frac{K}{2}$:

$$\mathbb{P}\Big(\sigma^+_{2k} \geq t_1 - \epsilon \,\Big|\, \sigma^+_K \geq t_1 - \epsilon, \sigma^-_K \leq t_0 + \epsilon\Big) \geq 1 - \tau\,. \tag{5.75}$$

Likewise, we can also prove mutatis mutandis that for any $k > \frac{K}{2}$:

$$\mathbb{P}\Big(\sigma^-_{2k} \leq t_0 + \epsilon \,\Big|\, \sigma^+_K \geq t_1 - \epsilon, \sigma^-_K \leq t_0 + \epsilon\Big) \geq 1 - \tau \tag{5.76}$$

where ϵ, τ, and K in (5.76) can be chosen to be the same as those in (5.75) without loss of generality.

Finally, we let $E = \{\sigma_K^+ \geq t_1 - \epsilon, \sigma_K^- \leq t_0 + \epsilon\}$, and observe that for all $k > \frac{K}{2}$:

$$\begin{aligned}
\mathbb{P}\Big(\sigma_{2k}^+ \geq t\Big) - \mathbb{P}\Big(\sigma_{2k}^- \geq t\Big) &\geq \mathbb{E}\Big[\Big(\mathbb{1}\Big\{\sigma_{2k}^+ \geq t\Big\} - \mathbb{1}\Big\{\sigma_{2k}^- \geq t\Big\}\Big)\,\mathbb{1}\{E\}\Big] \\
&= \Big(\mathbb{P}\Big(\sigma_{2k}^+ \geq t \,\Big|\, E\Big) - \mathbb{P}\Big(\sigma_{2k}^- \geq t \,\Big|\, E\Big)\Big)\,\mathbb{P}(E) \\
&\geq \Big(\mathbb{P}\Big(\sigma_{2k}^+ \geq t_1 - \epsilon \,\Big|\, E\Big) - \mathbb{P}\Big(\sigma_{2k}^- > t_0 + \epsilon \,\Big|\, E\Big)\Big)\,\mathbb{P}(E) \\
&\geq (1 - 2\tau)\mathbb{P}(E) > 0
\end{aligned}$$

where the first inequality holds because $\mathbb{1}\{\sigma_{2k}^+ \geq t\} - \mathbb{1}\{\sigma_{2k}^- \geq t\} \geq 0$ *a.s.* due to the monotonicity (property 5) of our Markovian coupling, the second inequality holds because $t_0 + \epsilon < t < t_1 - \epsilon$ (since $\epsilon > 0$ is small), and the final inequality follows from (5.75) and (5.76). As argued in the proof of Theorem 5.1, this illustrates that $\limsup_{k\to\infty} \mathbb{P}(\hat{T}_{2k} \neq \sigma_0) < \frac{1}{2}$.

Part 2: We next prove that $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$ implies:

$$\lim_{k\to\infty} \left\| P^+_{X_{2k}|G} - P^-_{X_{2k}|G} \right\|_{\mathsf{TV}} = 0 \quad G\text{-}a.s. \tag{5.77}$$

Following the proof of Theorem 5.1, we can show that:

$$\mathbb{E}\Big[\left\| P^+_{X_{2k}|G} - P^-_{X_{2k}|G} \right\|_{\mathsf{TV}}\Big] \leq L_{2k}\,\mathbb{E}\Big[\sigma_{2k}^+ - \sigma_{2k}^-\Big]\,. \tag{5.78}$$

In order to bound $\mathbb{E}[\sigma_{2k}^+ - \sigma_{2k}^-]$, we proceed as follows. Firstly, for any $k \in \mathbb{N}$, we have:

$$\begin{aligned}
\mathbb{E}\Big[\sigma_{2k}^+ - \sigma_{2k}^- \Big|\, \sigma_{2k-2}^+, \sigma_{2k-2}^-\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\sigma_{2k}^+ - \sigma_{2k}^- \Big|\, \sigma_{2k-1}^+, \sigma_{2k-1}^-\Big] \Big|\, \sigma_{2k-2}^+, \sigma_{2k-2}^-\Big] \\
&= \mathbb{E}\Big[g_0\Big(\sigma_{2k-1}^+\Big) - g_0\Big(\sigma_{2k-1}^-\Big) \Big|\, \sigma_{2k-2}^+, \sigma_{2k-2}^-\Big]
\end{aligned} \tag{5.79}$$

where the first equality follows from the tower and Markov properties, and the second equality holds because $L_{2k}\sigma_{2k} \sim \mathsf{binomial}(L_{2k}, g_0(\sigma))$ given $\sigma_{2k-1} = \sigma$. Then, recalling that $g_0(\sigma) = (\sigma * \delta)^2 = (1-2\delta)^2\sigma^2 + 2\delta(1-2\delta)\sigma + \delta^2$, we can compute:

$$\begin{aligned}
\mathbb{E}\Big[g_0\Big(\sigma_{2k-1}^+\Big) \Big|\, \sigma_{2k-2}^+, \sigma_{2k-2}^-\Big] &= \mathbb{E}\Big[g_0\Big(\sigma_{2k-1}^+\Big) \Big|\, \sigma_{2k-2}^+\Big] \\
&= \mathbb{E}\Big[(1-2\delta)^2 {\sigma_{2k-1}^+}^2 + 2\delta(1-2\delta)\sigma_{2k-1}^+ + \delta^2 \Big|\, \sigma_{2k-2}^+\Big] \\
&= (1-2\delta)^2\Big(\mathbb{VAR}\Big(\sigma_{2k-1}^+ \Big|\, \sigma_{2k-2}^+\Big) + \mathbb{E}\Big[\sigma_{2k-1}^+ \Big|\, \sigma_{2k-2}^+\Big]^2\Big) \\
&\quad + 2\delta(1-2\delta)\mathbb{E}\Big[\sigma_{2k-1}^+ \Big|\, \sigma_{2k-2}^+\Big] + \delta^2 \\
&= (1-2\delta)^2 g_1\Big(\sigma_{2k-2}^+\Big)^2 + 2\delta(1-2\delta)g_1\Big(\sigma_{2k-2}^+\Big) + \delta^2 \\
&\quad + (1-2\delta)^2\,\frac{g_1\Big(\sigma_{2k-2}^+\Big)\Big(1 - g_1\Big(\sigma_{2k-2}^+\Big)\Big)}{L_{2k-1}}
\end{aligned}$$

$$= g\Big(\sigma^+_{2k-2}\Big) + (1-2\delta)^2\,\frac{g_1\Big(\sigma^+_{2k-2}\Big)\Big(1-g_1\Big(\sigma^+_{2k-2}\Big)\Big)}{L_{2k-1}} \tag{5.80}$$

where the first equality uses property 3 of the monotone Markovian coupling, and the fourth equality uses the fact that $L_{2k-1}\sigma_{2k-1} \sim \mathsf{binomial}(L_{2k-1}, g_1(\sigma))$ given $\sigma_{2k-2} = \sigma$. Using (5.79) and (5.80), we get:

$$\begin{aligned}
&\mathbb{E}\Big[\sigma^+_{2k} - \sigma^-_{2k}\Big|\,\sigma^+_{2k-2}, \sigma^-_{2k-2}\Big] \\
&\qquad = g\Big(\sigma^+_{2k-2}\Big) - g\Big(\sigma^-_{2k-2}\Big) \\
&\qquad\quad + (1-2\delta)^2\left(\frac{g_1\Big(\sigma^+_{2k-2}\Big)\Big(1-g_1\Big(\sigma^+_{2k-2}\Big)\Big) - g_1\Big(\sigma^-_{2k-2}\Big)\Big(1-g_1\Big(\sigma^-_{2k-2}\Big)\Big)}{L_{2k-1}}\right) \\
&\qquad = g\Big(\sigma^+_{2k-2}\Big) - g\Big(\sigma^-_{2k-2}\Big) \\
&\qquad\quad + (1-2\delta)^2\left(\frac{g_1\Big(\sigma^+_{2k-2}\Big) - g_1\Big(\sigma^-_{2k-2}\Big) - \Big(g_1\Big(\sigma^+_{2k-2}\Big)^2 - g_1\Big(\sigma^-_{2k-2}\Big)^2\Big)}{L_{2k-1}}\right) \\
&\qquad \le g\Big(\sigma^+_{2k-2}\Big) - g\Big(\sigma^-_{2k-2}\Big) + (1-2\delta)^2\left(\frac{g_1\Big(\sigma^+_{2k-2}\Big) - g_1\Big(\sigma^-_{2k-2}\Big)}{L_{2k-1}}\right) \\
&\qquad \le \left(D(\delta) + \frac{2(1-\delta)(1-2\delta)^3}{L_{2k-1}}\right)\Big(\sigma^+_{2k-2} - \sigma^-_{2k-2}\Big) \\
&\qquad \le \left(D(\delta) + \frac{2}{L_{2k-1}}\right)\Big(\sigma^+_{2k-2} - \sigma^-_{2k-2}\Big)
\end{aligned} \tag{5.81}$$

where the first inequality holds because $g_1(\sigma^+_{2k-2})^2 - g_1(\sigma^-_{2k-2})^2 \ge 0$ *a.s.* (since g_1 is non-negative and non-decreasing by (5.60), and $\sigma^+_{2k-2} \ge \sigma^-_{2k-2}$ *a.s.* by property 5 of the monotone Markovian coupling), the second inequality holds because $\sigma^+_{2k-2} \ge \sigma^-_{2k-2}$ *a.s.* and g and g_1 have Lipschitz constants $D(\delta)$ and $\max_{\sigma\in[0,1]} g_1'(\sigma) = 2(1-\delta)(1-2\delta)$ respectively, and the final inequality holds because $(1-\delta)(1-2\delta)^3 \le 1$. Then, as in the proof of Theorem 5.1, we can take expectations in (5.81) to obtain:

$$0 \le \mathbb{E}\Big[\sigma^+_{2k} - \sigma^-_{2k}\Big] \le \left(D(\delta) + \frac{2}{L_{2k-1}}\right)\mathbb{E}\Big[\sigma^+_{2k-2} - \sigma^-_{2k-2}\Big]$$

which recursively produces:

$$0 \le \mathbb{E}\Big[\sigma^+_{2k} - \sigma^-_{2k}\Big] \le \prod_{i=1}^{k}\left(D(\delta) + \frac{2}{L_{2i-1}}\right)$$

where we use the fact that $\mathbb{E}[\sigma^+_0 - \sigma^-_0] = 1$.

Next, using (5.78) with this bound, we get:

$$\mathbb{E}\Big[\Big\|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\Big\|_{\mathsf{TV}}\Big] \leq L_{2k} \prod_{i=1}^{k} \left(D(\delta) + \frac{2}{L_{2i-1}}\right). \tag{5.82}$$

Recall that $L_k = o(E(\delta)^{-\frac{k}{2}})$ and $\liminf_{k\to\infty} L_k > \frac{2}{E(\delta)-D(\delta)}$ for some $E(\delta) \in (D(\delta), 1)$ (that depends on δ). Hence, there exists $K = K(\delta) \in \mathbb{N}$ (that depends on δ) such that for all $i > K$, $L_{2i-1} \geq \frac{2}{E(\delta)-D(\delta)}$. This means that we can further upper bound (5.82) as follows:

$$\forall k > K,\ \ \mathbb{E}\Big[\Big\|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\Big\|_{\mathsf{TV}}\Big] \leq L_{2k}\, E(\delta)^{k-K} \prod_{i=1}^{K} \left(D(\delta) + \frac{2}{L_{2i-1}}\right)$$

and letting $k \to \infty$ produces:

$$\lim_{k\to\infty} \mathbb{E}\Big[\Big\|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\Big\|_{\mathsf{TV}}\Big] = 0\,. \tag{5.83}$$

Finally, as in the proof of Theorem 5.1, $\|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\|_{\mathsf{TV}} \in [0,1]$ forms a non-increasing sequence in k for every realization of the random DAG G, and the pointwise limit random variable, $\lim_{k\to\infty} \|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\|_{\mathsf{TV}} \in [0,1]$, has mean:

$$\mathbb{E}\left[\lim_{k\to\infty} \Big\|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\Big\|_{\mathsf{TV}}\right] = 0$$

due to (5.83) and the bounded convergence theorem. Therefore, we must have (5.77). Moreover, since $\|P^{+}_{X_k|G} - P^{-}_{X_k|G}\|_{\mathsf{TV}} \in [0,1]$ forms a non-increasing sequence in k, we have:

$$\lim_{k\to\infty} \Big\|P^{+}_{X_k|G} - P^{-}_{X_k|G}\Big\|_{\mathsf{TV}} = \lim_{k\to\infty} \Big\|P^{+}_{X_{2k}|G} - P^{-}_{X_{2k}|G}\Big\|_{\mathsf{TV}}$$

for every realization of the random DAG G. Hence, we obtain (5.7), which completes the proof. ■

We remark that when $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$ and the condition $\liminf_{k\to\infty} L_k > \frac{2}{E(\delta)-D(\delta)}$ cannot be satisfied by any $E(\delta)$, if L_k satisfies the condition of Proposition 5.2 (in subsection 5.4.4), then part 2 of Proposition 5.2 still yields the desired converse result. Finally, the next proposition demonstrates that the Markov chain $\{\sigma_{2k} : k \in \mathbb{N} \cup \{0\}\}$ converges almost surely when $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$, $L_k = \omega(\log(k))$, all processing functions at even levels are the AND rule, and all processing functions at odd levels are the OR rule.

Proposition 5.6 (AND-OR Random DAG Model Almost Sure Convergence). *If $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$ and $L_k = \omega(\log(k))$, then $\lim_{k\to\infty} \sigma_{2k} = t$ almost surely.*

Proposition 5.6 is proved in appendix D.6, and much like Proposition 5.5, it can also be construed as a "weak" impossibility result.

■ 5.7 Deterministic Quasi-Polynomial Time and Randomized Polylogarithmic Time Constructions of DAGs where Broadcasting is Possible

In this section, we prove Theorem 5.3 by constructing deterministic bounded degree DAGs with $L_k = \Theta(\log(k))$ where broadcasting is possible. As mentioned in subsection 5.4.2, our construction is based on d-regular bipartite lossless $(d^{-6/5}, d-2d^{4/5})$-expander graphs. So, we first verify that such graphs actually exist. Recall that we represent a d-regular bipartite graph as $B = (U, V, E)$, where U and V are disjoint sets of vertices and E is the set of undirected edges. The next proposition is a specialization of [256, Proposition 1, Appendix II] which illustrates that randomly generated regular bipartite graphs are good expanders with high probability.

Proposition 5.7 (Random Expander Graph [226, Lemma 1], [256, Proposition 1, Appendix II]). *Fix any fraction $\alpha \in (0,1)$ and any degree $d \in \mathbb{N}$. Then, for every sufficiently large n (depending on α and d), the randomly generated d-regular bipartite graph $\mathsf{B} = (U, V, \mathsf{E})$ with $|U| = |V| = n$ satisfies:*

$$\begin{aligned}\mathbb{P}\Big(\forall S \subseteq U \text{ with } |S| = \alpha n,\, |\Gamma(S)| \geq n\Big(1 - (1-\alpha)^d - \sqrt{2d\alpha H(\alpha)}\Big)\Big) &> 1 - \binom{n}{n\alpha} \exp(-nH(\alpha)) \\ &\geq 1 - \frac{e}{2\pi\sqrt{\alpha(1-\alpha)n}}\end{aligned}$$

where $H(\alpha) \triangleq -\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)$ for $\alpha \in (0,1)$ denotes the binary Shannon entropy function.

We note that the probability measure $\mathbb{P}$ in Proposition 5.7 is defined by the random d-regular bipartite graph B, whose vertices U and V are fixed and edges E are random. In particular, B is generated as follows (cf. *configuration model* in [32, Section 2.4]):

1. Fix a complete bipartite graph $\hat{B} = (\hat{U}, \hat{V}, \hat{E})$ such that $|\hat{U}| = |\hat{V}| = dn$.
2. Randomly and uniformly select a perfect matching $\mathsf{M} \subseteq \hat{E}$ in $\hat{B}$.
3. Group sets of d consecutive vertices in $\hat{U}$, respectively $\hat{V}$, to generate a set of n super-vertices U, respectively V.
4. This yields a random d-regular bipartite graph $\mathsf{B} = (U, V, \mathsf{E})$, where every edge in E is an edge between super-vertices in M.

Note that we allow for the possibility that two vertices in B have multiple edges between them. The first inequality in Proposition 5.7 is proved in [256, Appendix II]. On the other hand, the second inequality in Proposition 5.7 is a straightforward consequence

of estimating the binomial coefficient using precise Stirling's formula bounds, cf. [89, Chapter II, Section 9, Equation (9.15)]:

$$\forall n \in \mathbb{N}, \quad \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n. \tag{5.84}$$

The second inequality portrays that the probability in Proposition 5.7 tends to 1 as $n \to \infty$. Moreover, strictly speaking, αn must be an integer, but we will neglect this detail throughout our exposition for simplicity (as in subsection 5.4.2). We next use this proposition to establish the existence of pertinent regular bipartite lossless expander graphs.

Corollary 5.2 (Lossless Expander Graph). *Fix any $\epsilon \in (0,1)$ and any degree $d \geq (\frac{2}{\epsilon})^5$. Then, for every sufficient large n (depending on d), the randomly generated d-regular bipartite graph $\mathsf{B} = (U, V, \mathsf{E})$ with $|U| = |V| = n$ satisfies:*

$$\mathbb{P}\Big(\forall S \subseteq U \text{ with } |S| = \frac{n}{d^{6/5}},\ |\Gamma(S)| \geq (1-\epsilon)d|S|\Big) > 1 - \frac{e}{2\pi\sqrt{d^{-6/5}(1-d^{-6/5})\,n}}.$$

Hence, for every sufficient large n (depending on d), there exists a d-regular bipartite lossless $(d^{-6/5}, (1-\epsilon)d)$-expander graph $B = (U, V, E)$ with $|U| = |V| = n$ such that for every subset of vertices $S \subseteq U$, we have:

$$|S| = \frac{n}{d^{6/5}} \quad \Rightarrow \quad |\Gamma(S)| \geq (1-\epsilon)d|S| = (1-\epsilon)\frac{n}{d^{1/5}}.$$

Corollary 5.2 is proved in appendix D.7. We remark that explicit constructions of bipartite lossless expander graphs B where only the vertices in U are d-regular can be found in the literature, cf. [39], but we require the vertices in V to be d-regular in our construction.

As we discussed in subsection 5.4.2, d-regular bipartite lossless expander graphs can be concatenated to produce a DAG where broadcasting is possible. To formally establish this, we first argue that a single d-regular bipartite lossless expander graph, when perceived as two successive layers of a deterministic DAG, exhibits a "one-step broadcasting" property. Fix any crossover probability $\delta \in (0, \frac{1}{2})$, and choose any sufficiently large *odd degree* $d = d(\delta)$ (that depends on δ) such that (5.27) (reproduced below) holds:

$$\frac{8}{d^{1/5}} + d^{6/5}\exp\left(-\frac{(1-2\delta)^2(d-4)^2}{8d}\right) \leq \frac{1}{2} \tag{5.85}$$

where the left hand side tends to 0 as $d \to \infty$ for fixed δ, and the minimum value of d satisfying this inequality increases as $\delta \to \frac{1}{2}^-$. Then, Corollary 5.2 demonstrates that for any sufficiently large n (depending on d), there exists a d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graph $B = (U, V, E)$ with $|U| = |V| = n$ such that the expansion property in (5.26) (reproduced below) holds:

$$\forall S \subseteq U, \quad |S| = \frac{n}{d^{6/5}} \quad \Rightarrow \quad |\Gamma(S)| \geq (1-\epsilon)\frac{n}{d^{1/5}} \text{ with } \epsilon = \frac{2}{d^{1/5}}. \tag{5.86}$$

Note that in the statements of Lemma 5.1 (see below) and Theorem 5.3, we assume the existence of such d-regular bipartite lossless $(d^{-6/5}, d-2d^{4/5})$-expander graphs without proof due to Corollary 5.2. Let us assume that the undirected edges in E are actually all directed from U to V, and construe B as two consecutive levels of a deterministic DAG upon which we are broadcasting (as in subsection 5.4.2). In particular, let the Bernoulli random variable corresponding to any vertex $v \in U \cup V$ be denoted by X_v, and suppose each (directed) edge of B is an independent $\mathsf{BSC}(\delta)$ as before. Furthermore, let the Boolean processing function at each vertex in V be the majority rule, which is always well-defined as d is odd. This defines a Bayesian network on B, and the ensuing lemma establishes the feasibility of "one-step broadcasting" down this Bayesian network.

Lemma 5.1 (One-Step Broadcasting in Expander DAG). *For any noise level $\delta \in (0, \frac{1}{2})$, any sufficiently large odd degree $d = d(\delta) \geq 5$ (that depends on δ) satisfying (5.27), and any sufficiently large n (depending on d), consider the Bayesian network, with independent $\mathsf{BSC}(\delta)$ noise on the edges and majority Boolean processing functions at the vertices, defined above on a d-regular bipartite lossless $(d^{-6/5}, d-2d^{4/5})$-expander graph $B = (U, V, E)$ such that $|U| = |V| = n$. Then, for every input distribution on $\{X_u : u \in U\}$, we have:*

$$\mathbb{P}\left(\sum_{v\in V} X_v > \frac{n}{d^{6/5}} \,\middle|\, \sum_{u\in U} X_u \leq \frac{n}{d^{6/5}}\right) \leq \exp\left(-\frac{n}{2d^{12/5}}\right).$$

Proof. We begin with some useful definitions. For any vertex $v \in V$, let $\mathsf{pa}(v)$ denote the multiset of vertices in U that are parents of v. (Note that $\mathsf{pa}(v)$ is a multiset because there may be multiple edges between two vertices, and $|\mathsf{pa}(v)| = d$.) Let $S \triangleq \{u \in U : X_u = 1\} \subseteq U$ denote the subset of vertices in U that take value 1, which implies that $|S| = \sum_{u\in U} X_u$. Furthermore, for any vertex $v \in V$, let $N_v \triangleq \sum_{u\in \mathsf{pa}(v)} X_u$ denote the number of parents of v in S that have value 1 (counting with repetition). Finally, let $T \triangleq \{v \in V : N_v \geq t\} \subseteq V$ denote the subset of vertices in V with at least $t \in \mathbb{N}\backslash\{1\}$ parents in S. We will assign an appropriate value to t below.

Suppose $|S| = \sum_{u\in U} X_u \leq n/d^{6/5}$ (which is the event we condition upon in the lemma statement). Consider the case where $|S| = n/d^{6/5}$. Then, applying the expansion property in (5.26) yields (the "vertex counting" bound):

$$|\Gamma(S)| = |T| + |\Gamma(S)\backslash T| \geq (1-\epsilon)\frac{n}{d^{1/5}} \tag{5.87}$$

where $T \subseteq \Gamma(S)$ by definition of T, and $\epsilon = 2d^{-1/5}$. Moreover, we also have the "edge counting" bound:

$$t|T| + |\Gamma(S)\backslash T| \leq d|S| = \frac{n}{d^{1/5}} \tag{5.88}$$

since each vertex in T has at least t edges from S, each vertex in $\Gamma(S)\backslash T$ has at least 1 edge from S, and the total number of outgoing edges from S is $d|S|$. Combining (5.87) and (5.88) produces:

$$(1-\epsilon)\frac{n}{d^{1/5}} - |T| \leq |\Gamma(S)\backslash T| \leq \frac{n}{d^{1/5}} - t|T|$$

which implies that:

$$|T| \leq \frac{n\epsilon}{d^{1/5}(t-1)} = \frac{2n}{d^{2/5}(t-1)}\,. \tag{5.89}$$

On the other hand, in the the case where $|S| < n/d^{6/5}$, if we flip the values of vertices in $U\backslash S$ to 1 and subsequently increase the cardinality of S, then the cardinality of T also increases or remains the same. Hence, if $|S| = \sum_{u\in U} X_u \leq n/d^{6/5}$, then (5.89) also holds.

Now, for any input distribution on $\{X_u : u \in U\}$, observe that:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{v\in V} X_v > \frac{n}{d^{6/5}} \,\middle|\, \sum_{u\in U} X_u \leq \frac{n}{d^{6/5}}\right)\\
&= \mathbb{P}\left(\sum_{v\in T} X_v + \sum_{v\in V\backslash T} X_v > \frac{n}{d^{6/5}} \,\middle|\, |S| \leq \frac{n}{d^{6/5}}\right)\\
&\leq \mathbb{P}\left(\sum_{v\in V\backslash T} X_v > \frac{n}{d^{6/5}} - |T| \,\middle|\, |S| \leq \frac{n}{d^{6/5}}\right)\\
&\leq \mathbb{P}\left(\sum_{v\in V\backslash T} X_v > \frac{n}{d^{6/5}} - \frac{2n}{d^{2/5}(t-1)} \,\middle|\, |S| \leq \frac{n}{d^{6/5}}\right)\\
&= \mathbb{E}\left[\mathbb{P}\left(\sum_{v\in V\backslash T} X_v > \frac{n}{d^{6/5}} - \frac{2n}{d^{2/5}(t-1)} \,\middle|\, V\backslash T, \{N_v : v \in V\backslash T\}\right) \middle| |S| \leq \frac{n}{d^{6/5}}\right] \qquad (5.90)\\
&\leq \mathbb{E}\left[\mathbb{P}\left(\sum_{v\in V\backslash T} X_v > \frac{n}{d^{6/5}} - \frac{2n}{d^{2/5}(t-1)} \,\middle|\, V\backslash T, \{\forall v \in V\backslash T,\ N_v = t-1\}\right) \middle| |S| \leq \frac{n}{d^{6/5}}\right]\\
&= \mathbb{E}\left[\mathbb{P}\left(\mathsf{binomial}(|V\backslash T|, \mathbb{P}(X_v = 1|N_v = t-1)) > \frac{n}{d^{6/5}} - \frac{2n}{d^{2/5}(t-1)} \middle| V\backslash T\right) \middle| |S| \leq \frac{n}{d^{6/5}}\right]\\
&\leq \mathbb{P}\left(\mathsf{binomial}(n, \mathbb{P}(X_v = 1|N_v = t-1)) > \frac{n}{d^{6/5}} - \frac{2n}{d^{2/5}(t-1)}\right)\\
&\leq \mathbb{P}\left(\mathsf{binomial}\left(n, \exp\left(-2d(1-2\delta)^2\left(\frac{1}{2} - \frac{t-1}{d}\right)^2\right)\right) > \frac{n}{d^{6/5}} - \frac{2n}{d^{2/5}(t-1)}\right) \qquad (5.91)
\end{aligned}$$

where the steps hold due to the following reasons:

1. In the first equality, T and $V\backslash T$ are random sets.
2. The second inequality holds because $X_v \in \{0,1\}$ for all $v \in T$.
3. The third inequality follows from (5.89).
4. The fourth equality uses the fact that $\{X_v : v \in V\backslash T\}$ are conditionally independent of the event $\{|S| \leq n/d^{6/5}\}$ given $V\backslash T$ and $\{N_v : v \in V\backslash T\}$, and the

conditional expectation in the fourth equality is over the random set $V\backslash T$ and the random variables $\{N_v : v \in V\backslash T\}$.

5. The fifth inequality holds because $N_v \leq t-1$ for every $v \in V\backslash T$, and a straightforward monotone coupling argument shows that the distribution $P_{X_v|N_v=t-1}$ stochastically dominates the distribution $P_{X_v|N_v=k}$ for any $k < t-1$. Furthermore, the conditional expectation in the fifth inequality is only over the random set $V\backslash T$.

6. The sixth equality holds because $\{X_v : v \in V\backslash T\}$ are conditionally i.i.d. given $V\backslash T$ and the event $\{\forall v \in V\backslash T,\, N_v = t-1\}$.

7. The seventh inequality holds because $|V\backslash T| \leq n$, and a simple monotone coupling argument establishes that a $\mathsf{binomial}(n, \mathbb{P}(X_v = 1|N_v = t-1))$ random variable stochastically dominates a $\mathsf{binomial}(|V\backslash T|, \mathbb{P}(X_v = 1|N_v = t-1))$ random variable.

8. The eighth inequality holds because a $\mathsf{binomial}(n,p)$ random variable stochastically dominates a $\mathsf{binomial}(n,q)$ random variable when $p \geq q$ (again via a monotone coupling argument), and Hoeffding's inequality (see Lemma C.4 in appendix C.2) yields:

$$\begin{aligned}\mathbb{P}(X_v = 1|N_v = t-1) &= \mathbb{P}\left(\sum_{i=1}^{t-1} Z_i + \sum_{j=1}^{d-t+1} Y_j > \frac{d}{2}\right) \\ &\leq \exp\left(-2d(1-2\delta)^2\left(\frac{1}{2} - \frac{t-1}{d}\right)^2\right) \end{aligned} \tag{5.92}$$

where Z_i are i.i.d. $\mathsf{Bernoulli}(1-\delta)$, Y_j are i.i.d. $\mathsf{Bernoulli}(\delta)$, $\{Z_i : i \in \{1,\ldots,t-1\}\}$ and $\{Y_j : j \in \{1,\ldots,d-t+1\}\}$ are independent, we assume that $\frac{t-1}{d} < \frac{1}{2}$, and we use the fact that X_v is the majority of its parents' values after passing them through independent $\mathsf{BSC}(\delta)$'s.

Finally, applying Hoeffding's inequality (see Lemma C.4 in appendix C.2) once more to (5.91) yields:

$$\begin{aligned}&\mathbb{P}\left(\sum_{v\in V} X_v > \frac{n}{d^{6/5}} \,\middle|\, \sum_{u\in U} X_u \leq \frac{n}{d^{6/5}}\right) \\ &\quad\leq \exp\left(-2n\left(\frac{1}{d^{6/5}} - \frac{2}{d^{2/5}(t-1)} - \exp\left(-2d(1-2\delta)^2\left(\frac{1}{2} - \frac{t-1}{d}\right)^2\right)\right)^2\right)\end{aligned} \tag{5.93}$$

where we assume that:

$$\frac{1}{d^{6/5}} - \frac{2}{d^{2/5}(t-1)} > \exp\left(-2d(1-2\delta)^2\left(\frac{1}{2} - \frac{t-1}{d}\right)^2\right). \tag{5.94}$$

Next, let $t = 1 + \lceil \frac{d}{4} \rceil$ so that:[95]

$$\frac{1}{4} \leq \frac{t-1}{d} \leq \frac{1}{4} + \frac{1}{d}\,.$$

Since we have assumed in the lemma statement that $d \geq 5$, the upper bound on $\frac{t-1}{d}$ illustrates that $\frac{t-1}{d} < \frac{1}{2}$, which ensures that (5.92) is valid. Furthermore, using both the upper and lower bounds on $\frac{t-1}{d}$, notice that (5.94) is also valid if we have:

$$\frac{1}{d^{6/5}} - \frac{8}{d^{7/5}} > \exp\left(-\frac{(1-2\delta)^2(d-4)^2}{8d}\right) \quad \Leftrightarrow \quad 1 > \frac{8}{d^{1/5}} + d^{6/5}\exp\left(-\frac{(1-2\delta)^2(d-4)^2}{8d}\right)$$

which is true by our assumption in (5.27). In fact, a simple computation shows that:

$$\begin{aligned}\frac{1}{d^{6/5}} - \frac{2}{d^{2/5}(t-1)} - \exp\left(-2d(1-2\delta)^2\left(\frac{1}{2} - \frac{t-1}{d}\right)^2\right) & \\ \geq \frac{1}{d^{6/5}} - \frac{8}{d^{7/5}} - \exp\left(-\frac{(1-2\delta)^2(d-4)^2}{8d}\right) & \\ \geq \frac{1}{2d^{6/5}} &\end{aligned}$$

where the second inequality is equivalent to (5.27). Therefore, we have from (5.93):

$$\mathbb{P}\left(\sum_{v\in V} X_v > \frac{n}{d^{6/5}} \,\middle|\, \sum_{u\in U} X_u \leq \frac{n}{d^{6/5}}\right) \leq \exp\left(-\frac{n}{2d^{12/5}}\right)$$

which completes the proof. ■

Intuitively, Lemma 5.1 parallels (5.46) in the proof of Theorem 5.1 in section 5.5. The lemma portrays that if the proportion of 1's is small in a given layer, then it remains small in the next layer with high probability when the edges between the layers are defined by a regular bipartite lossless expander graph. We next prove Theorem 5.3 by constructing deterministic bounded degree DAGs with $L_k = \Theta(\log(k))$ and showing using Lemma 5.1 that the root bit can be reconstructed using the majority decision rule $\hat{S}_k = \mathbb{1}\{\sigma_k \geq \frac{1}{2}\}$. In particular, we delineate two simple algorithms to construct the constituent expander graphs of such DAGs: a *deterministic quasi-polynomial time* algorithm and a *randomized polylogarithmic time* algorithm.

Proof of Theorem 5.3. Fix any $\delta \in (0, \frac{1}{2})$, any sufficiently large $d = d(\delta) \geq 5$ satisfying (5.27), and any sufficiently large constant $N = N(\delta) \in \mathbb{N}$ such that $M = \exp(N/(4d^{12/5})) \geq 2$ and for every $n \geq N$, there exists a d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graph $B_n = (U_n, V_n, E_n)$ with $|U_n| = |V_n| = n$ that satisfies (5.26) for every subset $S \subseteq U_n$. Furthermore, fix the level sizes so that $L_0 = 1$,

[95] The choice of t is arbitrary and we could have chosen any t such that $0 < \frac{t-1}{d} < \frac{1}{2}$.

$L_1 = N$, and $\{L_k : k \in \mathbb{N}\backslash\{1\}\}$ are defined by (5.28). It is straightforward to verify that $L_k = \Theta(\log(k))$ (for fixed δ). The remainder of the proof is split into three parts. We first present two simple algorithms to generate the constituent expander graphs of the deterministic DAG described in the theorem statement, and then argue that broadcasting is possible on the resulting DAG.

Deterministic Quasi-Polynomial Time Algorithm: We will require two useful facts:

1. For fixed sets of labeled vertices U_n and V_n with $|U_n| = |V_n| = n$, the total number of d-regular bipartite graphs $B_n = (U_n, V_n, E_n)$ is given by the multinomial coefficient:
$$\binom{nd}{d, d, \ldots, d} = \frac{(nd)!}{(d!)^n} \leq n^{nd} = \exp(dn \log(n))$$
where we allow multiple edges between two vertices, and the inequality follows from e.g. [59, Lemma 2.2]. To see this, first attach d edges to each vertex in U_n, and then successively count the number of ways to choose d edges for each vertex in V_n.[96]

2. Checking whether a given d-regular bipartite graph $B_n = (U_n, V_n, E_n)$ with $|U_n| = |V_n| = n$ satisfies (5.26) for all subsets $S \subseteq U_n$ using brute force takes running time $O(n^2 \exp(nH(d^{-6/5})))$. To see this, note that there are $\binom{n}{nd^{-6/5}} \leq \exp(nH(d^{-6/5}))$ (cf. [59, Lemma 2.2]) subsets $S \subseteq U_n$ with $|S| = nd^{-6/5}$, and verifying (5.26) takes $O(n^2)$ time for each such subset S.

Consider any level $M^{2^{m-1}} < r \leq M^{2^m}$ with some associated $m \in \mathbb{N}$. We show that the distinct expander graphs making up levels $0, \ldots, r$ of the deterministic DAG in the theorem statement can be constructed in quasi-polynomial time in r. In particular, we need to generate $m+1$ d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graphs $B_N, B_{2N}, \ldots, B_{2^m N}$. So, for each $i \in \{0, \ldots, m\}$, we generate $B_{2^i N}$ by exhaustively enumerating over the all possible d-regular bipartite graphs with $|U_{2^i N}| = |V_{2^i N}| = 2^i N$ until we find one that satisfies the desired expansion condition. (Note that such expander graphs are guaranteed to exist due to Corollary 5.2.) Using the aforementioned facts 1 and 2, generating all $m+1$ desired graphs takes running time:

$$O((m+1)L_r^2 \exp(L_r H(d^{-6/5})) \exp(dL_r \log(L_r))) = O(\exp(\Theta(\log(r) \log\log(r)))) \quad (5.95)$$

where we also use the facts that $L_r = 2^m N = \Theta(\log(r))$ and $m = \Theta(\log\log(r))$ since $M^{2^{m-1}} < r \leq M^{2^m}$.[97] Therefore, we can construct the constituent expander graphs

[96] Since the vertices in U_n and V_n are labeled, the total number of non-isomorphic d-regular bipartite graphs is smaller than $(nd)!/(d!)^n$. However, it is larger than $(nd)!/((2n)!(d!)^n)$, and the quasi-polynomial nature of our running time does not change with a more careful calculation of the number of non-isomorphic d-regular bipartite graphs.

[97] In our descriptions and analyses of the two algorithms, the big-O and big-Θ asymptotic notation conceal constants that depend on the fixed crossover probability parameter δ.

in levels $0, \ldots, r$ of our DAG in quasi-polynomial time with brute force. Note that we neglect details of how intermediate graphs are represented in our analysis. Moreover, we are not concerned with optimizing the quasi-polynomial running time.

Randomized Polylogarithmic Time Algorithm: We will require another useful fact:

3. A random d-regular bipartite graph $\mathsf{B} = (U_n, V_n, \mathsf{E})$ with $|U_n| = |V_n| = n$ can be generated according to the distribution described after Proposition 5.7 in $O(n)$ time. To see this, as outlined after Proposition 5.7, we must first generate a uniform random perfect matching in a complete bipartite graph $\hat{B} = (\hat{U}_{dn}, \hat{V}_{dn}, \hat{E})$ such that $|\hat{U}_{dn}| = |\hat{V}_{dn}| = dn$. Observe that the edges in a perfect matching can be written as a permutation of the sequence $(1, 2, \ldots, dn)$, because each index and its corresponding value in the (permuted) sequence encodes an edge. So, perfect matchings in $\hat{B}$ are in bijective correspondence with permutations of the sequence $(1, 2, \ldots, dn)$. Therefore, we can generate a uniform random perfect matching by generating a uniform random permutation of $(1, 2, \ldots, dn)$ in $O(dn)$, or equivalently $O(n)$, time using the *Fisher-Yates-Durstenfeld-Knuth shuffle*, cf. [153, Section 3.4.2, p.145] and the references therein. (Note that we do not take the running time of the random number generation process into account.) All that remains is to create super-vertices, which can also be done in $O(n)$ time.

Suppose that the constant $N = N(\delta)$ also satisfies the additional condition:

$$N > \frac{e^2}{\left(6 - 4\sqrt{2}\right)\pi^2 d^{-6/5}\left(1 - d^{-6/5}\right)} \tag{5.96}$$

where N still depends only on δ (through the dependence of d on δ). Consider any level $M^{2^{m-1}} < r \leq M^{2^m}$ with some associated $m \in \mathbb{N}$. We present a *Monte Carlo algorithm* that constructs the distinct expander graphs making up levels $0, \ldots, r$ of the deterministic DAG in the theorem statement with strictly positive success probability (that depends on δ but not on r) in polylogarithmic time in r. As in the previous algorithm, we ideally want to output $m+1$ d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graphs $B_N, B_{2N}, \ldots, B_{2^m N}$. So, using the aforementioned fact 3, for each $i \in \{0, \ldots, m\}$, we can generate a random d-regular bipartite graph $\mathsf{B} = (U_{2^i N}, V_{2^i N}, \mathsf{E})$ with $|U_{2^i N}| = |V_{2^i N}| = 2^i N$ according to the distribution in Corollary 5.2 in at most $O(2^m N)$ time. The total running time of the algorithm is thus:

$$O((m+1)2^m N) = O(\log(r) \log\log(r)) \tag{5.97}$$

since $2^m N = \Theta(\log(r))$ and $m = \Theta(\log\log(r))$ as before. Furthermore, by Corollary 5.2, the outputted random graphs satisfy (5.26) for all relevant subsets of vertices with probability at least:

$$\prod_{i=0}^{m}\left(1 - \frac{e}{2\pi\sqrt{d^{-6/5}\left(1 - d^{-6/5}\right)2^i N}}\right) \geq 1 - \frac{e}{2\pi\sqrt{d^{-6/5}\left(1 - d^{-6/5}\right)N}} \sum_{i=0}^{m}\left(\frac{1}{\sqrt{2}}\right)^i$$

$$\geq 1 - \frac{e}{2\pi\sqrt{d^{-6/5}(1-d^{-6/5})\,N}} \sum_{i=0}^{\infty} \left(\frac{1}{\sqrt{2}}\right)^i$$
$$= 1 - \frac{e}{\left(2-\sqrt{2}\right)\pi\sqrt{d^{-6/5}(1-d^{-6/5})\,N}} > 0 \tag{5.98}$$

where the first inequality is easily proved by induction, and the quantity in the final equality is strictly positive by assumption (5.96). Hence, our Monte Carlo algorithm constructs the constituent expander graphs in levels $0, \ldots, r$ of our DAG with strictly positive success probability in polylogarithmic time. Once again, note that we neglect details of how intermediate graphs are represented in our analysis. Moreover, we do not account for the running time of actually printing out levels $0, \ldots, r$ of the DAG.

Finally, the aforementioned fact 2 conveys that testing whether the $m+1$ d-regular random bipartite graphs our Monte Carlo algorithm generates are lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graphs takes polynomial running time:

$$O\Big((m+1)2^{2m}N^2 \exp\Big(2^m N H(d^{-6/5})\Big)\Big) = O\Big(\log\log(r)\log(r)^2 r^{8d^{12/5}H(d^{-6/5})}\Big) \tag{5.99}$$

where we use the fact that $2^m N < 2N\log(r)/\log(M) = 8d^{12/5}\log(r)$ since $r > M^{2^{m-1}}$ and $\log(M) = N/(4d^{12/5})$. Therefore, by repeatedly running our Monte Carlo algorithm until a valid set of $m+1$ d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graphs is produced, we obtain a *Las Vegas algorithm* that runs in expected polynomial time $O(\log\log(r)\log(r)^2 r^{8d^{12/5}H(d^{-6/5})})$.

Feasibility of Broadcasting: We now prove that broadcasting is possible on the Bayesian network defined on the DAG constructed in the theorem statement. As before, we follow the proof of Theorem 5.1 in section 5.5. So, we first construct a monotone Markovian coupling $\{(X_k^-, X_k^+) : k \in \mathbb{N}\cup\{0\}\}$ between the Markov chains $\{X_k^+ : k \in \mathbb{N}\cup\{0\}\}$ and $\{X_k^- : k \in \mathbb{N}\cup\{0\}\}$ (which denote versions of the Markov chain $\{X_k : k \in \mathbb{N}\cup\{0\}\}$ initialized at $X_0^+ = 1$ and $X_0^- = 0$, respectively) such that along any edge BSC of the deterministic DAG, say $(X_{k,j}, X_{k+1,i})$, $X_{k,j}^+$ and $X_{k,j}^-$ are either both copied with probability $1-2\delta$, or a shared independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit is produced with probability 2δ that becomes the value of both $X_{k+1,i}^+$ and $X_{k+1,i}^-$. This coupling satisfies the three properties delineated at the outset of the proof of Theorem 5.1 in section 5.5. Furthermore, let σ_k^+ and σ_k^- for $k \in \mathbb{N}\cup\{0\}$ be random variables with distributions $P_{\sigma_k|\sigma_0=1}$ and $P_{\sigma_k|\sigma_0=0}$, respectively (which means that $\sigma_0^+ = 1$ and $\sigma_0^- = 0$).

Notice that Lemma 5.1 implies the following result:

$$\mathbb{P}\left(\sigma_k^- \leq \frac{1}{d^{6/5}} \,\middle|\, \sigma_{k-1}^- \leq \frac{1}{d^{6/5}}\right) \geq 1 - \exp\left(-\frac{L_{k-1}}{2d^{12/5}}\right) \tag{5.100}$$

for every pair of consecutive levels $k-1$ and k such that $L_k = L_{k-1}$. Moreover, for every pair of consecutive levels $k-1$ and k such that $L_k = 2L_{k-1}$, we have:

$$\mathbb{P}\left(\sigma_k^- > \frac{1}{d^{6/5}} \,\middle|\, \sigma_{k-1}^- \leq \frac{1}{d^{6/5}}\right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\frac{1}{L_{k-1}} \sum_{i=0}^{L_{k-1}-1} X_{k,i}^{-} + \frac{1}{L_{k-1}} \sum_{j=L_{k-1}}^{L_k-1} X_{k,j}^{-} > \frac{2}{d^{6/5}} \,\middle|\, \sigma_{k-1}^{-} \leq \frac{1}{d^{6/5}} \right) \\
&= \mathbb{P}\left(\left\{ \frac{1}{L_{k-1}} \sum_{i=0}^{L_{k-1}-1} X_{k,i}^{-} > \frac{1}{d^{6/5}} \right\} \cup \left\{ \frac{1}{L_{k-1}} \sum_{j=L_{k-1}}^{L_k-1} X_{k,j}^{-} > \frac{1}{d^{6/5}} \right\} \,\middle|\, \sigma_{k-1}^{-} \leq \frac{1}{d^{6/5}} \right) \\
&\leq \mathbb{P}\left(\frac{1}{L_{k-1}} \sum_{i=0}^{L_{k-1}-1} X_{k,i}^{-} > \frac{1}{d^{6/5}} \,\middle|\, \sigma_{k-1}^{-} \leq \frac{1}{d^{6/5}} \right) \\
&\quad + \mathbb{P}\left(\frac{1}{L_{k-1}} \sum_{j=L_{k-1}}^{L_k-1} X_{k,j}^{-} > \frac{1}{d^{6/5}} \,\middle|\, \sigma_{k-1}^{-} \leq \frac{1}{d^{6/5}} \right) \\
&\leq 2 \exp\left(-\frac{L_{k-1}}{2d^{12/5}} \right)
\end{aligned}
$$

where the first inequality follows from the union bound, and the final inequality follows from Lemma 5.1 and the construction of our DAG (recall that two separate d-regular bipartite lossless $(d^{-6/5}, d - 2d^{4/5})$-expander graphs make up the edges between X_{k-1} and X_k^1, and between X_{k-1} and X_k^2, respectively). This implies that:

$$
\mathbb{P}\left(\sigma_k^{-} \leq \frac{1}{d^{6/5}} \,\middle|\, \sigma_{k-1}^{-} \leq \frac{1}{d^{6/5}} \right) \geq 1 - 2 \exp\left(-\frac{L_{k-1}}{2d^{12/5}} \right) \tag{5.101}
$$

for every pair of consecutive levels $k-1$ and k such that $L_k = 2L_{k-1}$, as well as for every pair of consecutive levels $k-1$ and k such that $L_k = L_{k-1}$ (by slackening the bound in (5.100)). Hence, the bound in (5.101) holds for all levels $k \geq 2$.

Now fix any $\tau > 0$, and choose a sufficiently large value $K = K(\delta, \tau) \in \mathbb{N}$ (that depends on δ and τ) such that:

$$
2 \sum_{k=K+1}^{\infty} \exp\left(-\frac{L_{k-1}}{2d^{12/5}} \right) \leq \tau \,. \tag{5.102}
$$

Note that such K exists because $2 \sum_{k=1}^{\infty} 1/k^2 = \pi^2/3 < +\infty$, and for every $m \in \mathbb{N} \cup \{0\}$ and every $M^{\lfloor 2^{m-1} \rfloor} < k \leq M^{2^m}$, we have:

$$
\exp\left(-\frac{L_k}{2d^{12/5}} \right) \leq \frac{1}{k^2} \quad \Leftrightarrow \quad k \leq \exp\left(\frac{2^m N}{4d^{12/5}} \right) = M^{2^m}
$$

where the right hand side holds due to the construction of our deterministic DAG. Using the continuity of probability measures, observe that:

$$
\begin{aligned}
&\mathbb{P}\left(\bigcap_{k>K} \left\{ \sigma_k^{-} \leq \frac{1}{d^{6/5}} \right\} \,\middle|\, \sigma_K^{+} \geq 1 - \frac{1}{d^{6/5}}, \, \sigma_K^{-} \leq \frac{1}{d^{6/5}} \right) \\
&\qquad = \prod_{k>K} \mathbb{P}\left(\sigma_k^{-} \leq \frac{1}{d^{6/5}} \,\middle|\, \sigma_{k-1}^{-} \leq \frac{1}{d^{6/5}}, \, A_k \right)
\end{aligned}
$$

$$\begin{aligned}
&\geq \prod_{k>K} 1 - 2\exp\left(-\frac{L_{k-1}}{2d^{12/5}}\right) \\
&\geq 1 - 2\sum_{k>K} \exp\left(-\frac{L_{k-1}}{2d^{12/5}}\right) \\
&\geq 1 - \tau
\end{aligned}$$

where A_k for $k > K$ is the non-zero probability event defined as:

$$A_k \triangleq \begin{cases} \left\{\sigma_K^+ \geq 1 - \frac{1}{d^{6/5}}\right\} & , \quad k = K+1 \\ \left\{\sigma_{k-2}^- \leq \frac{1}{d^{6/5}}, \ldots, \sigma_K^- \leq \frac{1}{d^{6/5}}\right\} \cap \left\{\sigma_K^+ \geq 1 - \frac{1}{d^{6/5}}\right\} & , \quad k \geq K+2 \end{cases},$$

the first inequality follows from (5.101), and the final inequality follows from (5.102). When using (5.101) in the calculation above, we can neglect the effect of the conditioning event A_k, because a careful perusal of the proof of Lemma 5.1 (which yields (5.101) as a consequence) shows that (5.101) continues to hold when we condition on events like A_k. Indeed, in step (5.90) of the proof, the random variables $\{X_v : v \in V\backslash T\}$ are conditionally independent of the *σ-algebra generated by random variables in previous layers of the DAG* given $V\backslash T$ and $\{N_v : v \in V\backslash T\}$. Moreover, this observation extends appropriately to the current Markovian coupling setting. We have omitted these details from Lemma 5.1 for the sake of clarity. Therefore, we have for any $k > K$:

$$\mathbb{P}\left(\sigma_k^- \leq \frac{1}{d^{6/5}} \,\middle|\, \sigma_K^+ \geq 1 - \frac{1}{d^{6/5}}, \sigma_K^- \leq \frac{1}{d^{6/5}}\right) \geq 1 - \tau\,. \tag{5.103}$$

Likewise, due to the symmetry of the role of 0's and 1's in our deterministic DAG model, we can also prove mutatis mutandis that for any $k > K$:

$$\mathbb{P}\left(\sigma_k^+ \geq 1 - \frac{1}{d^{6/5}} \,\middle|\, \sigma_K^+ \geq 1 - \frac{1}{d^{6/5}}, \sigma_K^- \leq \frac{1}{d^{6/5}}\right) \geq 1 - \tau \tag{5.104}$$

where τ and K in (5.104) can be chosen to be the same as those in (5.103) without loss of generality.

Finally, define the event $E = \{\sigma_K^+ \geq 1 - \frac{1}{d^{6/5}}, \sigma_K^- \leq \frac{1}{d^{6/5}}\}$, and observe that for all $k > K$:

$$\begin{aligned}
\mathbb{P}\left(\sigma_k^+ \geq \frac{1}{2}\right) - \mathbb{P}\left(\sigma_k^- \geq \frac{1}{2}\right) &\geq \mathbb{E}\left[\left(\mathbb{1}\left\{\sigma_k^+ \geq \frac{1}{2}\right\} - \mathbb{1}\left\{\sigma_k^- \geq \frac{1}{2}\right\}\right)\mathbb{1}\{E\}\right] \\
&= \left(\mathbb{P}\left(\sigma_k^+ \geq \frac{1}{2} \,\middle|\, E\right) - \mathbb{P}\left(\sigma_k^- \geq \frac{1}{2} \,\middle|\, E\right)\right)\mathbb{P}(E) \\
&\geq \left(\mathbb{P}\left(\sigma_k^+ \geq 1 - \frac{1}{d^{6/5}} \,\middle|\, E\right) - \mathbb{P}\left(\sigma_k^- > \frac{1}{d^{6/5}} \,\middle|\, E\right)\right)\mathbb{P}(E) \\
&\geq (1 - 2\tau)\mathbb{P}(E) > 0
\end{aligned}$$

where the first inequality holds because $\mathbb{1}\{\sigma_k^+ \geq \frac{1}{2}\} - \mathbb{1}\{\sigma_k^- \geq \frac{1}{2}\} \geq 0$ *a.s.* due to the monotonicity of our Markovian coupling, the third inequality holds because $\frac{1}{d^{6/5}} < \frac{1}{2} <$

$1-\frac{1}{d^{6/5}}$ (since $d \geq 5$), and the final inequality follows from (5.103) and (5.104). As argued in the proof of Theorem 5.1 in section 5.5, this illustrates that $\limsup_{k\to\infty} \mathbb{P}(\hat{S}_k \neq X_0) < \frac{1}{2}$, which completes the proof. ■

■ 5.8 Analysis of 2D Regular Grid with AND Processing Functions

We now turn to proving Theorem 5.4. Recall that we are given a deterministic 2D regular grid where all Boolean processing functions with two inputs are the AND rule, and all Boolean processing functions with one input are the identity rule, i.e. $f_2(x_1, x_2) = x_1 \wedge x_2$ and $f_1(x) = x$.

As in our proof of Theorem 5.1 in section 5.5, we begin by constructing a useful monotone Markovian coupling of the Markov chains $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$, which denote versions of the Markov chain $\{X_k : k \in \mathbb{N} \cup \{0\}\}$ initialized at $X_0^+ = 1$ and $X_0^- = 0$, respectively. (Note that the marginal distributions of X_k^+ and X_k^- are $P_{X_k}^+$ and $P_{X_k}^-$, respectively.) We define the coupled 2D grid variables $\{Y_{k,j} = (X_{k,j}^-, X_{k,j}^+) : k \in \mathbb{N} \cup \{0\}, j \in [k+1]\}$, and let the Markovian coupling be $\{Y_k = (Y_{k,0}, \ldots, Y_{k,k}) : k \in \mathbb{N} \cup \{0\}\}$. Since the underlying 2D regular grid is fixed, we couple $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$ so that along any edge BSC, say $(X_{k,j}, X_{k+1,j})$, $X_{k,j}^+$ and $X_{k,j}^-$ are either both copied with probability $1-2\delta$, or a shared independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit is produced with probability 2δ that becomes the value of both $X_{k+1,j}^+$ and $X_{k+1,j}^-$. As before, the Markovian coupling $\{Y_k : k \in \mathbb{N} \cup \{0\}\}$ exhibits the following properties:

1. The "marginal" Markov chains are $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$.
2. For every $k \in \mathbb{N} \cup \{0\}$, X_{k+1}^+ is conditionally independent of X_k^- given X_k^+, and X_{k+1}^- is conditionally independent of X_k^+ given X_k^-.
3. For every $k \in \mathbb{N} \cup \{0\}$ and every $j \in [k+1]$, $X_{k,j}^+ \geq X_{k,j}^-$ almost surely.

Here, the third property holds because AND processing functions are symmetric and monotone non-decreasing. In this section, probabilities of events that depend on the coupled 2D grid variables $\{Y_{k,j} : k \in \mathbb{N} \cup \{0\}, j \in [k+1]\}$ are defined with respect to this Markovian coupling.

Since the marginal Markov chains $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$ run on the same 2D regular grid with common BSCs, we keep track of the Markov chain $\{Y_k : k \in \mathbb{N} \cup \{0\}\}$ in a single coupled 2D regular grid. This 2D grid has the same underlying graph as the 2D grid described in subsection 5.3.2. Its vertices are the coupled 2D grid variables $\{Y_{k,j} = (X_{k,j}^-, X_{k,j}^+) : k \in \mathbb{N} \cup \{0\}, j \in [k+1]\}$, and we relabel the alphabet of these variables for simplicity. So, each $Y_{k,j} = (X_{k,j}^-, X_{k,j}^+) \in \mathcal{Y}$ with:

$$\mathcal{Y} \triangleq \{0_{\mathsf{c}}, 1_{\mathsf{u}}, 1_{\mathsf{c}}\} \tag{5.105}$$

where $0_{\mathsf{c}} = (0,0)$, $1_{\mathsf{u}} = (0,1)$, and $1_{\mathsf{c}} = (1,1)$. (Note that we do not require a letter $0_{\mathsf{u}} = (1,0)$ in this alphabet due to the monotonicity in the coupling.) Furthermore, each

edge of the coupled 2D grid is a channel $W \in \mathbb{R}_{\mathsf{sto}}^{3\times 3}$ between the alphabets $\mathcal{Y}$ and $\mathcal{Y}$ that captures the action of a shared $\mathsf{BSC}(\delta)$, where the stochastic matrix W has the form:

$$W = \begin{array}{c} \\ 0_{\mathsf{c}} \\ 1_{\mathsf{u}} \\ 1_{\mathsf{c}} \end{array} \begin{array}{c} \begin{array}{ccc} 0_{\mathsf{c}} & 1_{\mathsf{u}} & 1_{\mathsf{c}} \end{array} \\ \begin{bmatrix} 1-\delta & 0 & \delta \\ \delta & 1-2\delta & \delta \\ \delta & 0 & 1-\delta \end{bmatrix} \end{array} \tag{5.106}$$

where the (i,j)th entry gives the probability of output j given input i. It is straightforward to verify that W describes the aforementioned Markovian coupling. Finally, the AND rule can be equivalently described on the alphabet $\mathcal{Y}$ as:

x_1	x_2	$x_1 \wedge x_2$
0_{c}	$\star$	0_{c}
1_{u}	1_{u}	1_{u}
1_{u}	1_{c}	1_{u}
1_{c}	1_{c}	1_{c}

(5.107)

where $\star$ denotes any letter in $\mathcal{Y}$, and the symmetry of the AND rule covers all other possible input combinations. This coupled 2D grid model completely characterizes the Markov chain $\{Y_k : k \in \mathbb{N}\cup\{0\}\}$, which starts at $Y_{0,0} = 1_{\mathsf{c}}$ $a.s.$ We next prove Theorem 5.4 by further analyzing this model.

Proof of Theorem 5.4. We first bound the TV distance between $P_{X_k}^+$ and $P_{X_k}^-$ using Dobrushin's maximal coupling characterization of TV distance, cf. (2.6) in chapter 2:

$$\left\| P_{X_k}^+ - P_{X_k}^- \right\|_{\mathsf{TV}} \le \mathbb{P}\left(X_k^+ \neq X_k^-\right) = 1 - \mathbb{P}\left(X_k^+ = X_k^-\right).$$

The events $\{X_k^+ = X_k^-\}$ are non-decreasing in k, i.e. $\{X_k^+ = X_k^-\} \subseteq \{X_{k+1}^+ = X_{k+1}^-\}$ for all $k \in \mathbb{N}\cup\{0\}$. Indeed, suppose for any $k \in \mathbb{N}\cup\{0\}$, the event $\{X_k^+ = X_k^-\}$ occurs. Since we have:

$$\begin{aligned} \left\{X_k^+ = X_k^-\right\} &= \left\{Y_k \in \{0_{\mathsf{c}}, 1_{\mathsf{c}}\}^{k+1}\right\} \\ &= \{\text{there are no } 1_{\mathsf{u}}\text{'s in level } k \text{ of the coupled 2D grid}\}, \end{aligned}$$

the channel (5.106) and the rule (5.107) imply that there are no 1_{u}'s in level $k+1$. Hence, the event $\{X_{k+1}^+ = X_{k+1}^-\}$ occurs as well. Letting $k \to \infty$, we can use the continuity of $\mathbb{P}$ with the events $\{X_k^+ = X_k^-\}$ to get:

$$\lim_{k\to\infty} \left\| P_{X_k}^+ - P_{X_k}^- \right\|_{\mathsf{TV}} \le 1 - \lim_{k\to\infty} \mathbb{P}\left(X_k^+ = X_k^-\right) = 1 - \mathbb{P}(A)$$

where we define:

$$A \triangleq \{\exists k \in \mathbb{N}, \text{ there are no } 1_{\mathsf{u}}\text{'s in level } k \text{ of the coupled 2D grid}\}. \tag{5.108}$$

Therefore, it suffices to prove that $\mathbb{P}(A) = 1$.

To prove this, we recall a well-known result from [77, Section 3] on *oriented bond percolation* in 2D lattices. Given the underlying DAG of our deterministic 2D regular grid from subsection 5.3.2, suppose we independently keep each edge "open" with some probability $p \in [0, 1]$, and delete it ("closed") with probability $1 - p$. Define the event:

$$\Omega_\infty \triangleq \{\text{there is an infinite open path starting at the root}\}$$

and the quantities:

$$R_k \triangleq \sup\{j \in [k+1] : \text{there is an open path from the root to the vertex } (k, j)\}$$
$$L_k \triangleq \inf\{j \in [k+1] : \text{there is an open path from the root to the vertex } (k, j)\}$$

which are the rightmost and leftmost vertices at level $k \in \mathbb{N} \cup \{0\}$, respectively, that are connected to the root. (Here, we refer to the vertex $X_{k,j}$ using (k, j) as we do not associate a random variable to it.) It is proved in [77, Section 3] that the occurrence of Ω_∞ experiences a phase transition phenomenon as the open probability parameter p varies from 0 to 1.

Lemma 5.2 (Oriented Bond Percolation [77, Section 3]). *For the aforementioned bond percolation process on the 2D regular grid, there exists a critical threshold* $\delta_{\mathsf{perc}} \in \left(\frac{1}{2}, 1\right)$ *around which we observe the following phase transition phenomenon:*

1. *If* $p > \delta_{\mathsf{perc}}$, *then* $\mathbb{P}_p(\Omega_\infty) > 0$ *and:*

$$\mathbb{P}_p\left(\lim_{k\to\infty} \frac{R_k}{k} = \frac{1+\alpha(p)}{2} \text{ and } \lim_{k\to\infty} \frac{L_k}{k} = \frac{1-\alpha(p)}{2} \,\middle|\, \Omega_\infty\right) = 1 \tag{5.109}$$

   for some constant $\alpha(p) > 0$, *where* $\alpha(p)$ *is defined in [77, Section 3, Equation (6)], and* $\mathbb{P}_p$ *is the probability measure defined by the bond percolation process.*

2. *If* $p < \delta_{\mathsf{perc}}$, *then* $\mathbb{P}_p(\Omega_\infty) = 0$.

We will use Lemma 5.2 to prove $\mathbb{P}(A) = 1$ by considering two cases.

Case 1: Suppose $1 - 2\delta < \delta_{\mathsf{perc}}$ (i.e. $\delta > (1 - \delta_{\mathsf{perc}})/2$) in our coupled 2D grid. The root of the coupled 2D grid is $Y_{0,0} = 1_{\mathsf{u}}$ *a.s.*, and we consider an oriented bond percolation process on the grid (as described above) with $p = 1 - 2\delta$. In particular, we say that each edge of the grid is open if and only if the corresponding BSC copies its input (with probability $1 - 2\delta$). In this context, Ω_∞^c is the event that there exists $k \in \mathbb{N}$ such that none of the vertices at level k are connected to the root via a sequence of BSCs that are copies. Suppose the event Ω_∞^c occurs. Since (5.106) and (5.107) portray that a 1_{u} moves from level k to level $k+1$ only if one of its outgoing edges is open (and the corresponding BSC is a copy), there exists $k \in \mathbb{N}$ such that none of the vertices at level k are 1_{u}'s. This proves that $\Omega_\infty^c \subseteq A$. Therefore, using part 2 of Lemma 5.2, we get $\mathbb{P}(A) = 1$.

Case 2: Suppose $1-\delta > \delta_{\mathsf{perc}}$ (i.e. $\delta < 1-\delta_{\mathsf{perc}}$) in our coupled 2D grid. Consider an oriented bond percolation process on the grid (as described earlier) with $p = 1-\delta$, where an edge is open if and only if the corresponding BSC is either copying or generating a 0 as the new bit (i.e. this BSC takes a 0_{c} to a 0_{c}, which happens with probability $1-\delta$ as shown in (5.106)). Let B_k for $k \in \mathbb{N}$ be the event that the BSC from $Y_{k-1,0}$ to $Y_{k,0}$ generates a new bit which equals 0. Then, $\mathbb{P}(B_k) = \delta$ and $\{B_k : k \in \mathbb{N}\}$ are mutually independent. So, the second Borel-Cantelli lemma tells us that infinitely many of the events $\{B_k : k \in \mathbb{N}\}$ occur almost surely. Furthermore, $B_k \subseteq \{Y_{k,0} = 0_{\mathsf{c}}\}$ for every $k \in \mathbb{N}$.

We next define the following sequence of random variables for all $i \in \mathbb{N}$:

$$L_i \triangleq \min\{k \geq T_{i-1}+1 : B_k \text{ occurs}\}$$
$$T_i \triangleq 1 + \max\{k \geq L_i : \exists j \in [k+1],\, Y_{k,j} \text{ is connected to } Y_{L_i,0} \text{ by an open path}\}$$

where we set $T_0 \triangleq 0$. Note that when $T_{i-1} = \infty$, we let $L_i = \infty$ *a.s.*, and when $T_{i-1} < \infty$, $L_i < \infty$ *a.s.* because infinitely many of the events $\{B_k : k \in \mathbb{N}\}$ occur almost surely. We also note that when $L_i < \infty$, the set:

$$\{k \geq L_i : \exists j \in [k+1],\, Y_{k,j} \text{ is connected to } Y_{L_i,0} \text{ by an open path}\}$$

is non-empty since $Y_{L_i,0}$ is always connected to itself, and $T_i - L_i - 1$ denotes the length of the longest open path connected to $Y_{L_i,0}$ (which could be infinity). Lastly, when $L_i = \infty$, we let $T_i = \infty$ *a.s.*

Let $\mathcal{F}_k$ for every $k \in \mathbb{N} \cup \{0\}$ be the σ-algebra generated by the random variables $(Y_0, \ldots, Y_k)$ and all the BSCs above level k (where we include all events determining whether these BSCs are copies, and all events determining the independent bits they produce). Then, $\{\mathcal{F}_k : k \in \mathbb{N} \cup \{0\}\}$ is a *filtration*. It is straightforward to verify that L_i and T_i are *stopping times* with respect to $\{\mathcal{F}_k : k \in \mathbb{N} \cup \{0\}\}$ for all $i \in \mathbb{N}$. We can show this inductively. $T_0 = 0$ is trivially a stopping time, and if T_{i-1} is a stopping time, then L_i is clearly a stopping time. So, it suffices to prove that T_i is a stopping time given L_i is a stopping time. For any finite $m \in \mathbb{N}$, $\{T_i = m\}$ is the event that $L_i \leq m-1$ and the length of the longest open path connected to $Y_{L_i,0}$ is $m-1-L_i$. This event is contained in $\mathcal{F}_m$ because the event $\{L_i \leq m-1\}$ is contained in $\mathcal{F}_{m-1} \subseteq \mathcal{F}_m$ (since L_i is a stopping time), and the length of the longest open path can be determined from $\mathcal{F}_m$ (rather than $\mathcal{F}_{m-1}$). Hence, T_i is indeed a stopping time when L_i is a stopping time.

Now observe that:

$$\begin{aligned}
\mathbb{P}(\exists k \in \mathbb{N},\, T_k = \infty) &= \mathbb{P}(T_1 = \infty) + \sum_{m=2}^{\infty} \mathbb{P}(\exists k \in \mathbb{N}\backslash\{1\},\, T_k = \infty | T_1 = m)\mathbb{P}(T_1 = m) \\
&= \mathbb{P}(T_1 = \infty) + \sum_{m=2}^{\infty} \mathbb{P}(\exists k \in \mathbb{N},\, T_k + m = \infty)\mathbb{P}(T_1 = m) \\
&= \mathbb{P}(T_1 = \infty) + (1 - \mathbb{P}(T_1 = \infty))\mathbb{P}(\exists k \in \mathbb{N},\, T_k = \infty) \qquad (5.110)
\end{aligned}$$

where the first equality uses the law of total probability, the third equality follows from straightforward calculations, and the second equality follows from the fact that for all $m \in \mathbb{N}\backslash\{1\}$:

$$\mathbb{P}(\exists k \in \mathbb{N}\backslash\{1\},\, T_k = \infty | T_1 = m) = \mathbb{P}(\exists k \in \mathbb{N},\, T_k + m = \infty)\,.$$

This relation holds because the random variables $\{(L_i, T_i) : i \in \mathbb{N}\backslash\{1\}\}$ given $T_1 = m$ have the same distribution as the random variables $\{(L_{i-1} + m, T_{i-1} + m) : i \in \mathbb{N}\backslash\{1\}\}$. In particular, the conditional distribution of L_i given $T_1 = m$ corresponds to the distribution of $L_{i-1} + m$, and the conditional distribution of T_i given $T_1 = m$ corresponds to the distribution of $T_{i-1} + m$. These distributional equivalences implicitly use the fact that $\{T_i : i \in \mathbb{N}\}$ are stopping times. Indeed, the conditioning on $\{T_1 = m\}$ in these equivalences can be removed because the event $\{T_1 = m\}$ is in $\mathcal{F}_m$ since T_1 is a stopping time, and $\{T_1 = m\}$ is therefore independent of the events $\{B_k : k > m\}$ and the events that determine when the BSCs below level m are open.

Next, rearranging (5.110), we get:

$$\mathbb{P}(\exists k \in \mathbb{N},\, T_k = \infty)\mathbb{P}(T_1 = \infty) = \mathbb{P}(T_1 = \infty)\,.$$

Since $\mathbb{P}(T_1 = \infty) = \mathbb{P}(\Omega_\infty) > 0$ by part 1 of Lemma 5.2, we have:

$$\mathbb{P}(\exists k \in \mathbb{N},\, T_k = \infty) = 1\,. \tag{5.111}$$

For every $k \in \mathbb{N}$, define the events:

$$\Omega_k^{\mathsf{left}} \triangleq \{\text{there exists an infinite open path starting at the vertex } Y_{k,0}\}\,,$$
$$\Omega_k^{\mathsf{right}} \triangleq \{\text{there exists an infinite open path starting at the vertex } Y_{k,k}\}\,.$$

If $\{\exists k \in \mathbb{N},\, T_k = \infty\}$ occurs, we can choose the smallest $m \in \mathbb{N}$ such that $T_m = \infty$, and for this m, there is an infinite open path starting at $Y_{L_m,0} = 0_{\mathsf{c}}$ (where $Y_{L_m,0} = 0_{\mathsf{c}}$ because B_{L_m} occurs). Hence, using (5.111), we have:

$$\mathbb{P}\Big(\exists k \in \mathbb{N},\, \{Y_{k,0} = 0_{\mathsf{c}}\} \cap \Omega_k^{\mathsf{left}}\Big) = 1\,.$$

Likewise, we can also prove that:

$$\mathbb{P}\Big(\exists k \in \mathbb{N},\, \{Y_{k,k} = 0_{\mathsf{c}}\} \cap \Omega_k^{\mathsf{right}}\Big) = 1$$

which implies that:

$$\mathbb{P}\Big(\exists k \in \mathbb{N}, \exists m \in \mathbb{N},\, \{Y_{k,0} = Y_{m,m} = 0_{\mathsf{c}}\} \cap \Omega_k^{\mathsf{left}} \cap \Omega_m^{\mathsf{right}}\Big) = 1\,. \tag{5.112}$$

To finish the proof, consider $k, m \in \mathbb{N}$ such that $Y_{k,0} = Y_{m,m} = 0_{\mathsf{c}}$, and suppose Ω_k^{left} and $\Omega_m^{\mathsf{right}}$ both happen. For every $n > \max\{k, m\}$, define the quantities:

$$R_n^{\mathsf{left}} \triangleq \sup\{j \in [n+1] : \text{there is an open path from } Y_{k,0} \text{ to } Y_{n,j}\}$$

$$L_n^{\mathsf{right}} \triangleq \inf\{j \in [n+1] : \text{there is an open path from } Y_{m,m} \text{ to } Y_{n,j}\}$$

which are the rightmost and leftmost vertices at level n that are connected to $Y_{k,0}$ and $Y_{m,m}$, respectively, by open paths. Using (5.109) from part 1 of Lemma 5.2, we know that almost surely:

$$\lim_{n\to\infty} \frac{R_n^{\mathsf{left}}}{n} = \lim_{n\to\infty} \frac{R_n^{\mathsf{left}}}{n-k} = \frac{1+\alpha(1-\delta)}{2}\,,$$
$$\lim_{n\to\infty} \frac{L_n^{\mathsf{right}}}{n} = \lim_{n\to\infty} \frac{L_n^{\mathsf{right}}-m}{n-m} = \frac{1-\alpha(1-\delta)}{2}\,.$$

This implies that almost surely:

$$\lim_{n\to\infty} \frac{R_n^{\mathsf{left}} - L_n^{\mathsf{right}}}{n} = \alpha(1-\delta) > 0$$

which means that for some sufficiently large level $n > \max\{k, m\}$, the rightmost open path from $Y_{k,0}$ meets the leftmost open path from $Y_{m,m}$:

$$\left|R_n^{\mathsf{left}} - L_n^{\mathsf{right}}\right| \leq 1\,.$$

By construction, all the vertices in these two open paths are equal to 0_{c}. Furthermore, since (5.106) and (5.107) demonstrate that AND gates and BSCs output 0_{c}'s or 1_{c}'s when their inputs are 0_{c}'s or 1_{c}'s, it is straightforward to inductively establish that all vertices at level n that are either to left of R_n^{left} or to the right of L_n^{right} take values in $\{0_{\mathsf{c}}, 1_{\mathsf{c}}\}$. This shows that every vertex at level n must be equal to 0_{c} or 1_{c} because the two aforementioned open paths meet. Hence, there exists a level $n \in \mathbb{N}$ with no 1_{u}'s, i.e. the event A occurs. Therefore, we get $\mathbb{P}(A) = 1$ using (5.112).

Combining the two cases completes the proof as $\mathbb{P}(A) = 1$ for any $\delta \in (0, \frac{1}{2})$. ■

We remark that this proof can be perceived as using the technique presented in [170, Theorem 5.2]. Indeed, let $T \triangleq \inf\{k \in \mathbb{N} : X_k^+ = X_k^-\}$ be a stopping time (with respect to the filtration $\{\mathcal{F}_k : k \in \mathbb{N} \cup \{0\}\}$ defined earlier) denoting the first time that the marginal Markov chains $\{X_k^+ : k \in \mathbb{N} \cup \{0\}\}$ and $\{X_k^- : k \in \mathbb{N} \cup \{0\}\}$ meet. (Note that $\{T = \infty\}$ corresponds to the event that these chains never meet.) Since the events $\{X_k^+ = X_k^-\}$ for $k \in \mathbb{N}$ form a non-decreasing sequence of sets, $\{T > k\} = \{X_k^+ \neq X_k^-\}$. We can use this relation to obtain the following bound on the TV distance between $P_{X_k}^+$ and $P_{X_k}^-$ (cf. (2.6) in chapter 2):

$$\left\|P_{X_k}^+ - P_{X_k}^-\right\|_{\mathsf{TV}} \leq \mathbb{P}\Big(X_k^+ \neq X_k^-\Big) = \mathbb{P}(T > k) = 1 - \mathbb{P}(T \leq k) \tag{5.113}$$

where letting $k \to \infty$ and using the continuity of $\mathbb{P}$ produces:

$$\lim_{k\to\infty} \left\|P_{X_k}^+ - P_{X_k}^-\right\|_{\mathsf{TV}} \leq 1 - \mathbb{P}(\exists k \in \mathbb{N},\, T \leq k) = 1 - \mathbb{P}(T < \infty)\,. \tag{5.114}$$

These bounds correspond to the ones shown in [170, Theorem 5.2]. Since the event $A = \{\exists k \in \mathbb{N}, T \leq k\} = \{T < \infty\}$, our proof that A happens almost surely also demonstrates that the two marginal Markov chains meet after a finite amount of time almost surely.

■ 5.9 Analysis of 2D Regular Grid with XOR Processing Functions

We finally turn to proving Theorem 5.5. We will use some basic coding theory ideas in this section, and refer readers to [237] for an introduction to the subject. We let $\mathbb{F}_2 = \{0, 1\}$ denote the Galois field of order 2 (i.e. integers with addition and multiplication modulo 2), $\mathbb{F}_2^n$ with $n \in \mathbb{N}\backslash\{1\}$ denote the vector space over $\mathbb{F}_2$ of column vectors with n entries from $\mathbb{F}_2$, and $\mathbb{F}_2^{m\times n}$ with $m, n \in \mathbb{N}\backslash\{1\}$ denote the space of $m \times n$ matrices with entries in $\mathbb{F}_2$. (All matrix and vector operations will be performed modulo 2.) Now fix some matrix $H \in \mathbb{F}_2^{m\times n}$ that has the following block structure:

$$H = \begin{bmatrix} 1 & B_1 \\ \mathbf{0} & B_2 \end{bmatrix} \tag{5.115}$$

where $\mathbf{0}$ denotes the zero vector (of appropriate dimension), $B_1 \in \mathbb{F}_2^{1\times(n-1)}$, and $B_2 \in \mathbb{F}_2^{(m-1)\times(n-1)}$. Consider the following two problems:

1. **Coding Problem:** Let $\mathcal{C} \triangleq \{x \in \mathbb{F}_2^n : Hx = \mathbf{0}\}$ be the *linear code* defined by the *parity check matrix* H. Let $X = [X_1 \ X_2^T]^T$ with $X_1 \in \mathbb{F}_2$ and $X_2 \in \mathbb{F}_2^{n-1}$ be a codeword drawn uniformly from $\mathcal{C}$. Assume that there exists a codeword $x = [1 \ x_2^T]^T \in \mathcal{C}$ for some $x_2 \in \mathbb{F}_2^{n-1}$ (i.e. $B_1 x_2 = 1$ and $B_2 x_2 = \mathbf{0}$). Then, since $\mathcal{C} \ni x' \mapsto x' + x \in \mathcal{C}$ is a bijective map that flips the first bit of its input, X_1 is a $\mathsf{Bernoulli}(\frac{1}{2})$ random variable. We observe the codeword X through an additive noise channel model and see $Y_1 \in \mathbb{F}_2$ and $Y_2 \in \mathbb{F}_2^{n-1}$:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = X + \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} X_1 + Z_1 \\ X_2 + Z_2 \end{bmatrix} \tag{5.116}$$

   where $Z_1 \in \mathbb{F}_2$ is a $\mathsf{Bernoulli}(\frac{1}{2})$ random variable, $Z_2 \in \mathbb{F}_2^{n-1}$ is a vector of i.i.d. $\mathsf{Bernoulli}(\delta)$ random variables that are independent of Z_1, and both Z_1, Z_2 are independent of X. Our problem is to decode X_1 with minimum probability of error after observing Y_1, Y_2. This can be achieved by using the ML decoder for X_1 based on Y_1, Y_2.

2. **Inference Problem:** Let $X' \in \mathbb{F}_2$ be a $\mathsf{Bernoulli}(\frac{1}{2})$ random variable, and $Z \in \mathbb{F}_2^{n-1}$ be a vector of i.i.d. $\mathsf{Bernoulli}(\delta)$ random variables that are independent of X'. Suppose we see the observations $S_1' \in \mathbb{F}_2$ and $S_2' \in \mathbb{F}_2^{m-1}$ through the model:

$$\begin{bmatrix} S_1' \\ S_2' \end{bmatrix} = H \begin{bmatrix} X' \\ Z \end{bmatrix} = \begin{bmatrix} X' + B_1 Z \\ B_2 Z \end{bmatrix}. \tag{5.117}$$

   Our problem is to decode X' with minimum probability of error after observing S_1', S_2'. This can be achieved by using the ML decoder for X' based on S_1', S_2'.

As we will soon see, the inference problem above corresponds to our setting of reconstruction in the 2D regular grid with XOR processing functions. The next lemma

illustrates that this inference problem is in fact "equivalent" to the aforementioned coding problem, and this connection will turn out to be useful since the coding problem admits simpler analysis.

Lemma 5.3 (Equivalence of Problems). *For the coding problem in (5.116) and the inference problem in (5.117), the following statements hold:*

1. *The minimum probabilities of error for the coding and inference problems are equal.*
2. *Suppose the random variables in the coding and inference problems are coupled so that $X_1 = X'$ a.s. and $Z_2 = Z$ a.s. (i.e. these variables are shared by the two problems), X_2 is generated from a conditional distribution $P_{X_2|X_1}$ such that X is uniform on $\mathcal{C}$, Z_1 is generated independently, (Y_1, Y_2) is defined by (5.116), and (S_1', S_2') is defined by (5.117). Then, $S_1' = B_1 Y_2$ a.s. and $S_2' = B_2 Y_2$ a.s.*
3. *Under the aforementioned coupling, (S_1', S_2') is a sufficient statistic of (Y_1, Y_2) for performing inference about X_1 (in the coding problem).*

Proof.

Part 1: We first show that the minimum probabilities of error for the two problems are equal. The inference problem has prior is $X' \sim \mathsf{Bernoulli}(\frac{1}{2})$, and the following likelihoods for every $s_1' \in \mathbb{F}_2$ and every $s_2' \in \mathbb{F}_2^{m-1}$:

$$P_{S_1', S_2'|X'}(s_1', s_2'|0) = \sum_{z \in \mathbb{F}_2^{n-1}} P_Z(z) \mathbb{1}\{B_1 z = s_1', B_2 z = s_2'\}, \tag{5.118}$$

$$P_{S_1', S_2'|X'}(s_1', s_2'|1) = \sum_{z \in \mathbb{F}_2^{n-1}} P_Z(z) \mathbb{1}\{B_1 z = s_1' + 1, B_2 z = s_2'\}. \tag{5.119}$$

On the other hand, the coding problem has prior $X_1 \sim \mathsf{Bernoulli}(\frac{1}{2})$, and the following likelihoods for every $y_1 \in \mathbb{F}_2$ and every $y_2 \in \mathbb{F}_2^{n-1}$:

$$\begin{aligned} P_{Y_1, Y_2|X_1}(y_1, y_2|0) &= P_{Y_1|X_1}(y_1|0) P_{Y_2|X_1}(y_2|0) \\ &= \frac{1}{2} \sum_{x_2 \in \mathbb{F}_2^{n-1}} P_{Y_2|X_2}(y_2|x_2) P_{X_2|X_1}(x_2|0) \\ &= \frac{1}{2} \sum_{x_2 \in \mathbb{F}_2^{n-1}} P_{Z_2}(y_2 - x_2) \mathbb{1}\{B_1 x_2 = 0, B_2 x_2 = \mathbf{0}\} \frac{2}{|\mathcal{C}|} \\ &= \frac{1}{|\mathcal{C}|} \sum_{z_2 \in \mathbb{F}_2^{n-1}} P_{Z_2}(z_2) \mathbb{1}\{B_1 z_2 = B_1 y_2, B_2 z_2 = B_2 y_2\}, \end{aligned} \tag{5.120}$$

$$P_{Y_1, Y_2|X_1}(y_1, y_2|1) = \frac{1}{|\mathcal{C}|} \sum_{z_2 \in \mathbb{F}_2^{n-1}} P_{Z_2}(z_2) \mathbb{1}\{B_1 z_2 = B_1 y_2 + 1, B_2 z_2 = B_2 y_2\}, \tag{5.121}$$

where the third equality uses the fact that X_2 is uniform over a set of cardinality $|\mathcal{C}|/2$ given any value of X_1, because $X_1 \sim \mathsf{Bernoulli}(\frac{1}{2})$ and X is uniform on $\mathcal{C}$. For the coding

problem, define $S_1 \triangleq B_1 Y_2$ and $S_2 \triangleq B_2 Y_2$. Due to the Fisher-Neyman factorization theorem [150, Theorem 3.6], (5.120) and (5.121) demonstrate that (S_1, S_2) is a sufficient statistic of (Y_1, Y_2) for performing inference about X_1.

Continuing in the context of the coding problem, define the set $\mathcal{C}' = \{x \in \mathbb{F}_2^{n-1} : B_1 x = 0, B_2 x = \mathbf{0}\}$, which is also a linear code, and for any fixed $s_1 \in \mathbb{F}_2$ and $s_2 \in \mathbb{F}_2^{m-1}$, define the set $\mathcal{S}(s_1, s_2) = \{(y_1, y_2) \in \mathbb{F}_2 \times \mathbb{F}_2^{n-1} : B_1 y_2 = s_1, B_2 y_2 = s_2\}$. If there exists $y_2' \in \mathbb{F}_2^{n-1}$ such that $B_1 y_2' = s_1$ and $B_2 y_2' = s_2$, then $\mathcal{S}(s_1, s_2) = \{(y_1, y_2 + y_2') \in \mathbb{F}_2 \times \mathbb{F}_2^{n-1} : y_2 \in \mathcal{C}'\}$, which means that $|\mathcal{S}(s_1, s_2)| = 2|\mathcal{C}'| = |\mathcal{C}|$ (where the final equality holds because each vector in $\mathcal{C}'$ corresponds to a codeword in $\mathcal{C}$ whose first letter is 0, and we have assumed that there are an equal number of codewords in $\mathcal{C}$ with first letter 1). Hence, for every $s_1 \in \mathbb{F}_2$ and every $s_2 \in \mathbb{F}_2^{m-1}$, the likelihoods of (S_1, S_2) given X_1 can be computed from (5.120) and (5.121):

$$\begin{aligned} P_{S_1,S_2|X_1}(s_1, s_2|0) &= \sum_{y_1 \in \mathbb{F}_2,\, y_2 \in \mathbb{F}_2^{n-1}} P_{Y_1,Y_2|X_1}(y_1, y_2|0) \mathbb{1}\{B_1 y_2 = s_1, B_2 y_2 = s_2\} \\ &= \frac{|\mathcal{S}(s_1, s_2)|}{|\mathcal{C}|} \sum_{z_2 \in \mathbb{F}_2^{n-1}} P_{Z_2}(z_2) \mathbb{1}\{B_1 z_2 = s_1, B_2 z_2 = s_2\} \\ &= \sum_{z_2 \in \mathbb{F}_2^{n-1}} P_{Z_2}(z_2) \mathbb{1}\{B_1 z_2 = s_1, B_2 z_2 = s_2\}, \end{aligned} \tag{5.122}$$

$$P_{S_1,S_2|X_1}(s_1, s_2|1) = \sum_{z_2 \in \mathbb{F}_2^{n-1}} P_{Z_2}(z_2) \mathbb{1}\{B_1 z_2 = s_1 + 1, B_2 z_2 = s_2\}, \tag{5.123}$$

where the second equality follows from (5.120) and the third equality clearly holds in the $|\mathcal{S}(s_1, s_2)| = 0$ case as well. The likelihoods (5.122) and (5.123) are exactly the same as the likelihoods (5.118) and (5.119), respectively, that we computed earlier for the inference problem. Thus, the sufficient statistic (S_1, S_2) of (Y_1, Y_2) for X_1 in the coding problem is equivalent to the observation (S_1', S_2') in the inference problem in the sense that they are defined by the same probability model. As a result, the minimum probabilities of error in these formulations must be equal.

Part 2: We now assume that the random variables in the two problems are coupled as in the lemma statement. To prove that $S_1' = S_1$ *a.s.* and $S_2' = S_2$ *a.s.*, observe that:

$$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} B_1 Y_2 \\ B_2 Y_2 \end{bmatrix} = \begin{bmatrix} B_1 X_2 + B_1 Z_2 \\ B_2 X_2 + B_2 Z_2 \end{bmatrix} = \begin{bmatrix} X_1 + B_1 Z_2 \\ B_2 Z_2 \end{bmatrix} = H \begin{bmatrix} X_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} S_1' \\ S_2' \end{bmatrix}$$

where the second equality uses (5.116), the third equality holds because $B_1 X_2 = X_1$ and $B_2 X_2 = \mathbf{0}$ since $X \in \mathcal{C}$ is a codeword, and the last equality uses (5.117) and the fact that $X_1 = X'$ *a.s.* and $Z_2 = Z$ *a.s.* This proves part 2.

Part 3: Since (S_1, S_2) is a sufficient statistic of (Y_1, Y_2) for performing inference about X_1 in the coding problem, and $S_1' = S_1$ *a.s.* and $S_2' = S_2$ *a.s.* under the coupling in the lemma statement, (S_1', S_2') is also a sufficient statistic of (Y_1, Y_2) for performing inference about X_1 under this coupling. This completes the proof. ■

Recall that we are given a deterministic 2D regular grid where all Boolean processing functions with two inputs are the XOR rule, and all Boolean processing functions with one input are the identity rule, i.e. $f_2(x_1, x_2) = x_1 \oplus x_2$ and $f_1(x) = x$. We next prove Theorem 5.5 using Lemma 5.3.

Proof of Theorem 5.5. We first prove that the problem of decoding the root bit in the XOR 2D grid is captured by the inference problem defined in (5.117). Let E_k denote the set of all directed edges in the 2D regular grid above level $k \in \mathbb{N}$. Furthermore, let us associate each edge $e \in E_k$ with an independent $\mathsf{Bernoulli}(\delta)$ random variable $Z_e \in \mathbb{F}_2$. Since a $\mathsf{BSC}(\delta)$ can be modeled as addition of an independent $\mathsf{Bernoulli}(\delta)$ bit (in $\mathbb{F}_2$), the random variables $\{Z_e : e \in E_k\}$ define the BSCs of the 2D regular grid up to level k. Moreover, each vertex at level $k \in \mathbb{N}$ of the XOR 2D grid is simply a sum (in $\mathbb{F}_2$) of its parent vertices and the random variables on the edges between it and its parents:

$$\begin{aligned} \forall j \in \{1, \ldots, k-1\}, \quad X_{k,j} &= X_{k-1,j-1} \oplus X_{k-1,j} \oplus Z_{(X_{k-1,j-1}, X_{k,j})} \oplus Z_{(X_{k-1,j}, X_{k,j})} \,, \\ X_{k,0} &= X_{k-1,0} \oplus Z_{(X_{k-1,0}, X_{k,0})} \,, \\ X_{k,k} &= X_{k-1,k-1} \oplus Z_{(X_{k-1,k-1}, X_{k,k})} \,. \end{aligned}$$

These recursive formulae for each vertex in terms of its parent vertices can be unwound so that each vertex is represented as a linear combination (in $\mathbb{F}_2$) of the root bit and all the edge random variables:

$$\forall k \in \mathbb{N}, \forall j \in [k+1], \quad X_{k,j} = \left(\binom{k}{j} \pmod 2 \right) X_{0,0} + \sum_{e \in E_k} b_{k,j,e} Z_e \tag{5.124}$$

where the coefficient of $X_{0,0}$ can be computed by realizing that the coefficients of the vertices in the "2D regular grid above $X_{k,j}$" (with $X_{k,j}$ as the root) are defined by the recursion of Pascal's triangle, and $b_{k,j,e} \in \mathbb{F}_2$ are some fixed coefficients. We do not require detailed knowledge of the values of $\{b_{k,j,e} \in \mathbb{F}_2 : k \in \mathbb{N}, j \in [k+1], e \in E_k\}$, but they can also be evaluated via straightforward counting if desired.

In the remainder of this proof, we will fix k to be a power of 2: $k = 2^m$ for $m \in \mathbb{N}$. Then, we have:

$$\binom{k}{j} \equiv \binom{2^m}{j} \equiv \begin{cases} 1 & , \quad j \in \{0, k\} \\ 0 & , \quad j \in \{1, \ldots, k-1\} \end{cases} \pmod 2 \tag{5.125}$$

since by *Lucas' theorem* (see [90]), the parity of $\binom{k}{j}$ is 0 if and only if at least one of the digits of j in base 2 is strictly greater than the corresponding digit of k in base 2, and the base 2 representation of $k = 2^m$ is $10 \cdots 0$ (with m 0's). So, for each k, we can define a binary matrix $H_k \in \mathbb{F}_2^{(k+1) \times (|E_k|+1)}$ whose rows are indexed by the vertices at level k and columns are indexed by 1 (first index corresponding to $X_{0,0}$) followed by the edges in E_k, and whose rows are made up of the coefficients in (5.124) (where the first

entry of each row is given by (5.125)). Clearly, we can write (5.124) in matrix-vector form using H_k for every k:

$$\begin{bmatrix} X_{k,0} \\ X_{k,1} \\ \vdots \\ X_{k,k-1} \\ X_{k,k} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \text{—} & b_{k,0,e} & \text{—} \\ 0 & \text{—} & b_{k,1,e} & \text{—} \\ \vdots & & \vdots & \\ 0 & \text{—} & b_{k,k-1,e} & \text{—} \\ 1 & \text{—} & b_{k,k,e} & \text{—} \end{bmatrix}}_{\triangleq H_k} \begin{bmatrix} X_{0,0} \\ | \\ Z_e \\ | \end{bmatrix} \tag{5.126}$$

where the vector on the right hand side of (5.126) has first element $X_{0,0}$ followed by the random variables $\{Z_e : e \in E_k\}$ (indexed consistently with H_k). Our XOR 2D grid reconstruction problem is to decode $X_{0,0}$ from the observations $(X_{k,0}, \ldots, X_{k,k})$ with minimum probability of error. Note that we can apply a row operation to H_k that replaces the last row of H_k with the sum of the first and last rows of H_k to get the binary matrix $H'_k \in \mathbb{F}_2^{(k+1)\times(|E_k|+1)}$, and correspondingly, we can replace $X_{k,k}$ with $X_{k,0} + X_{k,k}$ in (5.126) to get the "equivalent" formulation:

$$\begin{bmatrix} X_{k,0} \\ X_{k,1} \\ \vdots \\ X_{k,k-1} \\ X_{k,0} + X_{k,k} \end{bmatrix} = H'_k \begin{bmatrix} X_{0,0} \\ | \\ Z_e \\ | \end{bmatrix} \tag{5.127}$$

for every k. Indeed, since we only perform invertible operations to obtain (5.127) from (5.126), the minimum probability of error for ML decoding $X_{0,0}$ from the observations $(X_{k,0}, \ldots, X_{k,k})$ under the model (5.126) is equal to the minimum probability of error for ML decoding $X_{0,0}$ from the observations $(X_{k,0}, \ldots, X_{k,k-1}, X_{k,0} + X_{k,k})$ under the model (5.127). Furthermore, since H'_k is of the form (5.115), the equivalent XOR 2D grid reconstruction problem in (5.127) is exactly of the form of the inference problem in (5.117).

We next transform the XOR 2D grid reconstruction problem in (5.126), or equivalently, (5.127), into a coding problem. By Lemma 5.3, the inference problem in (5.127) is "equivalent" to a coupled coding problem analogous to (5.116). To describe this coupled coding problem, consider the linear code defined by the parity check matrix H'_k:

$$\mathcal{C}_k \triangleq \left\{ w \in \mathbb{F}_2^{|E_k|+1} : H'_k w = \mathbf{0} \right\} = \left\{ w \in \mathbb{F}_2^{|E_k|+1} : H_k w = \mathbf{0} \right\}$$

where the second equality shows that the parity check matrix H_k also generates $\mathcal{C}_k$ (because row operations do not change the nullspace of a matrix). As required by the coding problem, this linear code contains a codeword of the form $[1 \ w_2^T]^T \in \mathcal{C}_k$ for some $w_2 \in \mathbb{F}_2^{|E_k|}$. To prove this, notice that such a codeword exists if and only if the

first column $[1\,0\cdots 0]^T$ of H'_k is in the span of the remaining columns of H'_k. Assume for the sake of contradiction that such a codeword does not exist. Then, we can decode $X_{0,0}$ in the setting of (5.127) with zero probability of error, because the observation vector on the left hand side of (5.127) is in the span of the second to last columns of H'_k if and only if $X_{0,0} = 0$.[98] This leads to a contradiction since it is clear that we cannot decode the root bit with zero probability of error in the XOR 2D grid. Hence, a codeword of the form $[1 \;\; w_2^T]^T \in \mathcal{C}_k$ for some $w_2 \in \mathbb{F}_2^{|E_k|}$ always exists. Next, we let $W_k = [X_{0,0} \;\; —W_{k,e}—]^T \in \mathcal{C}_k$ be a codeword that is drawn uniformly from $\mathcal{C}_k$, where the first element of W_k is $X_{0,0}$ and the remaining elements of W_k are $\{W_{k,e} : e \in E_k\}$. In the coupled coding problem, we observe W_k through the additive noise channel model:

$$Y_k \triangleq \begin{bmatrix} Y_{0,0}^k \\ | \\ Y_{k,e} \\ | \end{bmatrix} = W_k + \begin{bmatrix} Z_{0,0}^k \\ | \\ Z_e \\ | \end{bmatrix} \tag{5.128}$$

where $\{Z_e : e \in E_k\}$ are the BSC random variables that are independent of W_k, $Z_{0,0}^k$ is a completely independent $\mathsf{Bernoulli}(\frac{1}{2})$ random variable, $Y_{0,0}^k = X_{0,0} \oplus Z_{0,0}^k$, and $Y_{k,e} = W_{k,e} \oplus Z_e$ for $e \in E_k$. Our goal is to decode the first bit of the codeword, $X_{0,0}$, with minimum probability of error from the observation Y_k. Since we have coupled the coding problem (5.128) and the inference problem (5.127) according to the coupling in part 2 of Lemma 5.3, part 3 of Lemma 5.3 shows that $(X_{k,0}, \ldots, X_{k,k-1}, X_{k,0} + X_{k,k})$, or equivalently:

$$\begin{bmatrix} X_{k,0} \\ \vdots \\ X_{k,k} \end{bmatrix} = \begin{bmatrix} \sum_{e \in E_k} b_{k,0,e} Y_{k,e} \\ \vdots \\ \sum_{e \in E_k} b_{k,k,e} Y_{k,e} \end{bmatrix}, \tag{5.129}$$

is a sufficient statistic of Y_k for performing inference about $X_{0,0}$ in the coding problem (5.128). Hence, the ML decoder for $X_{0,0}$ based on the sufficient statistic $(X_{k,0}, \ldots, X_{k,k-1}, X_{k,0} + X_{k,k})$ (without loss of generality), which achieves the minimum probability of error in the coding problem (5.128), makes an error if and only if the ML decision rule for $X_{0,0}$ based on $(X_{k,0}, \ldots, X_{k,k-1}, X_{k,0} + X_{k,k})$, which achieves the minimum probability of error in the inference problem (5.127), makes an error. Therefore, as shown in part 1 of Lemma 5.3, the minimum probabilities of error in the XOR 2D grid reconstruction problem (5.126) and the coding problem (5.128) are equal, and it suffices to analyze the coding problem (5.128).

In the coding problem (5.128), we observe the codeword W_k after passing it through memoryless BSCs. We now establish a "cleaner" model where W_k is passed through

[98] It is worth mentioning that in the ensuing coding problem in (5.128), if such a codeword does not exist, we can also decode the first codeword bit with zero probability of error because all codewords must have the first bit equal to 0.

memoryless BECs. Recall that each $\mathsf{BSC}(\delta)$ copies its input with probability $1-2\delta$ and generates an independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit with probability 2δ (as shown in the proof of Proposition 5.2 in appendix D.3), i.e. for any $e \in E_k$, instead of setting $Z_e \sim \mathsf{Bernoulli}(\delta)$, we can generate Z_e as follows:

$$Z_e = \begin{cases} 0 & , \quad \text{with probability } 1-2\delta \\ \mathsf{Bernoulli}\left(\frac{1}{2}\right) & , \quad \text{with probability } 2\delta \end{cases}$$

where $\mathsf{Bernoulli}(\frac{1}{2})$ denotes an independent uniform bit. Suppose we know which BSCs among $\{Z_e : e \in E_k\}$ generate independent bits in (5.128). Then, we can perceive each BSC in $\{Z_e : e \in E_k\}$ as an independent $\mathsf{BEC}(2\delta)$, which erases its input with probability 2δ and produces the erasure symbol e if and only if the corresponding $\mathsf{BSC}(\delta)$ generates an independent bit, and copies its input with probability $1-2\delta$ otherwise. (Note that the BSC defined by $Z_{0,0}^k$ corresponds to a $\mathsf{BEC}(1)$ which always erases its input.) Consider observing the codeword W_k under this *BEC model*, where $X_{0,0}$ is erased *a.s.*, and the remaining bits of W_k are erased independently with probability 2δ, i.e. we observe $Y_k' = [\mathsf{e} \; — Y_{k,e}' —]^T \in \{0,1,\mathsf{e}\}^{|E_k|+1}$, where the first entry corresponds to the erased value of $X_{0,0}$, and for every $e \in E_k$, $Y_{k,e}' = W_{k,e}$ with probability $1-2\delta$ and $Y_{k,e}' = \mathsf{e}$ with probability 2δ. Clearly, we can obtain Y_k from Y_k' by replacing every instance of e in Y_k' with an independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit. Since the BECs reveal additional information about which BSCs generate independent bits, the minimum probability of error in ML decoding $X_{0,0}$ based on Y_k' under the BEC model lower bounds the minimum probability of error in ML decoding $X_{0,0}$ based on Y_k under the BSC model (5.128).[99] In the rest of the proof, we establish conditions under which the minimum probability of error for the BEC model is $\frac{1}{2}$, and then show as a consequence that the minimum probability of error in the XOR 2D grid reconstruction problem in (5.126) tends to $\frac{1}{2}$ as $k \to \infty$.

Let $I_k \subseteq E_k$ denote the set of indices where the corresponding elements of W_k are not erased in the BEC model:

$$I_k \triangleq \left\{ e \in E_k : Y_{k,e}' = W_{k,e} \right\}.$$

The ensuing lemma is a standard exercise in coding theory which shows that the ML decoder for $X_{0,0}$ only fails under the BEC model when a special codeword exists in $\mathcal{C}_k$; see the discussion in [237, Section 3.2].

Lemma 5.4 (Bit-wise ML Decoding [237, Section 3.2]). *Suppose we condition on some realization of Y_k' (in the BEC model), which determines a corresponding realization of the set of indices I_k. Then, the ML decoder for $X_{0,0}$ based on Y_k' (with codomain $\mathbb{F}_2$) makes an error with probability $\frac{1}{2}$ if and only if there exists a codeword $w \in \mathcal{C}_k$ with first element $w_1 = 1$ and $w_e = 0$ for all $e \in I_k$.*

[99] Indeed, the ML decoder for $X_{0,0}$ based on Y_k' has a smaller (or equal) probability of error than the decoder which first translates Y_k' into Y_k by replacing every e with an independent $\mathsf{Bernoulli}(\frac{1}{2})$ bit, and then applies the ML decoder for $X_{0,0}$ based on Y_k as in the coding problem (5.128). We also remark that the relation $\mathsf{BEC}(2\delta) \succeq_{\ln} \mathsf{BSC}(\delta)$ is well-known in information theory, cf. (2.59) and (3.22) in chapters 2 and 3.

We next illustrate that such a special codeword exists whenever two particular erasures occur. Let $e_1 \in E_k$ and $e_2 \in E_k$ denote the edges $(X_{k-1,0}, X_{k,0})$ and $(X_{k-1,k-1}, X_{k,k})$ in the 2D regular grid, respectively. Consider the vector $\omega^k \in \mathbb{F}_2^{|E_k|+1}$ such that $\omega_1^k = 1$ (i.e. the first bit is 1), $\omega_{e_1}^k = \omega_{e_2}^k = 1$, and all other elements of ω^k are 0. Then, $\omega^k \in \mathcal{C}_k$ because:

$$H_k\,\omega^k = \begin{bmatrix} 1 & — & b_{k,0,e} & — \\ 0 & — & b_{k,1,e} & — \\ \vdots & & \vdots & \\ 0 & — & b_{k,k-1,e} & — \\ 1 & — & b_{k,k,e} & — \end{bmatrix} \omega^k = \begin{bmatrix} 1 \oplus b_{k,0,e_1} \oplus b_{k,0,e_2} \\ b_{k,1,e_1} \oplus b_{k,1,e_2} \\ \vdots \\ b_{k,k-1,e_1} \oplus b_{k,k-1,e_2} \\ 1 \oplus b_{k,k,e_1} \oplus b_{k,k,e_2} \end{bmatrix} = \mathbf{0}$$

where we use the facts that $b_{k,0,e_1} = 1$, $b_{k,0,e_2} = 0$, $b_{k,k,e_1} = 0$, $b_{k,k,e_2} = 1$, and for any $j \in \{1, \ldots, k-1\}$, $b_{k,j,e_1} = 0$ and $b_{k,j,e_2} = 0$. (Note that the value of b_{k,j,e_i} for $i \in \{0,1\}$ and $j \in [k+1]$ is determined by checking the dependence of vertex $X_{k,j}$ on the variable Z_{e_i} in (5.124), which is straightforward because e_i is an edge between the last two layers at the side of the 2D regular grid up to level k). Since ω^k has two 1's at the indices e_1 and e_2 (besides the first bit), if the BECs corresponding to the indices e_1 and e_2 erase their inputs, i.e. $e_1, e_2 \notin I_k$, then $\omega^k \in \mathcal{C}_k$ satisfies the conditions of Lemma 5.4 and the ML decoder for $X_{0,0}$ based on Y_k' under the BEC model makes an error with probability $\frac{1}{2}$. Hence, we define the event:

$$\begin{aligned} B_k &\triangleq \left\{ Y'_{k,e_1} = Y'_{k,e_2} = \mathsf{e} \right\} \\ &= \{\text{BECs corresponding to edges } e_1 \in E_k \text{ and } e_2 \in E_k \text{ erase their inputs}\} \\ &= \{\text{BSCs corresponding to edges } e_1 \in E_k \text{ and } e_2 \in E_k \text{ generate independent bits}\}. \end{aligned}$$

As the ML decoder for $X_{0,0}$ based on Y_k' under the BEC model makes an error with probability $\frac{1}{2}$ conditioned on B_k, we must have:

$$P_{Y_k'|X_{0,0}}(y'|0) = P_{Y_k'|X_{0,0}}(y'|1)$$

for all realizations $y' \in \{0, 1, \mathsf{e}\}^{|E_k|+1}$ of Y_k' such that B_k occurs, i.e. $y'_1 = y'_{e_1} = y'_{e_2} = \mathsf{e}$. This implies that Y_k' is conditionally independent of $X_{0,0}$ given B_k (where we also use the fact that $X_{0,0}$ is independent of B_k). Furthermore, it is straightforward to verify that Y_k is also conditionally independent of $X_{0,0}$ given B_k, because Y_k can be obtained from Y_k' by replacing e's with completely independent $\mathsf{Bernoulli}(\frac{1}{2})$ bits. Thus, since (5.129) shows that X_k is a deterministic function of Y_k, X_k is conditionally independent of $X_{0,0}$ given B_k.

To finish the proof, notice that $\mathbb{P}(B_k) = (2\delta)^2$ for every k, and the events $\{B_k : k = 2^m, m \in \mathbb{N}\}$ are mutually independent because the BSCs in the 2D regular grid are all independent. So, infinitely many of the events $\{B_k : k = 2^m, m \in \mathbb{N}\}$ occur *a.s.* by the second Borel-Cantelli lemma. Let us define:

$$\forall n \in \mathbb{N}, \quad A_n \triangleq \bigcup_{m=1}^{n} B_{2^m}$$

where the continuity of the underlying probability measure $\mathbb{P}$ yields $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$. Then, since X_k is conditionally independent of $X_{0,0}$ given B_k, and X_r is conditionally independent of $X_{0,0}$ and B_k given X_k for any $r > k$, we have that X_{2^m} is conditionally independent of $X_{0,0}$ given A_m for every $m \in \mathbb{N}$. Hence, we obtain:

$$\forall m \in \mathbb{N}, \;\; \mathbb{P}\Big(h_{\mathsf{ML}}^{2^m}(X_{2^m}) \neq X_{0,0} \,\Big|\, A_m\Big) = \frac{1}{2}$$

where $h_{\mathsf{ML}}^{k} : \mathbb{F}_2^{k+1} \to \mathbb{F}_2$ denotes the ML decoder for $X_{0,0}$ based on X_k for the XOR 2D grid reconstruction problem in (5.126). Finally, observe that:

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}\Big(h_{\mathsf{ML}}^{2^m}(X_{2^m}) \neq X_{0,0}\Big) &= \lim_{m\to\infty} \mathbb{P}\Big(h_{\mathsf{ML}}^{2^m}(X_{2^m}) \neq X_{0,0} \,\Big|\, A_m\Big)\, \mathbb{P}(A_m) \\
&\qquad + \mathbb{P}\Big(h_{\mathsf{ML}}^{2^m}(X_{2^m}) \neq X_{0,0} \,\Big|\, A_m^c\Big)\, \mathbb{P}(A_m^c) \\
&= \lim_{m\to\infty} \mathbb{P}\Big(h_{\mathsf{ML}}^{2^m}(X_{2^m}) \neq X_{0,0} \,\Big|\, A_m\Big) \\
&= \frac{1}{2}\,.
\end{aligned}$$

This completes the proof since the above condition establishes (5.14). ■

■ 5.10 Conclusion and Future Directions

To conclude, we recapitulate the main contributions of this chapter. For random DAG models with indegree $d \geq 3$, we considered the intuitively reasonable setting where all Boolean processing functions are the majority rule. We proved in Theorem 5.1 that reconstruction of the root bit for this model is possible using the majority decision rule when $\delta < \delta_{\mathsf{maj}}$ and $L_k = \Omega(\log(k))$, and impossible using the ML decision rule in all but a zero measure set of DAGs when $\delta > \delta_{\mathsf{maj}}$ and L_k is sub-exponential. On the other hand, when the indegree $d = 2$ so that the choices of Boolean processing functions are unclear, we derived a similar phase transition in Theorem 5.2 for random DAG models with AND processing functions at all even levels and OR processing functions at all odd levels. These main results on random DAG models established the existence of deterministic DAGs where broadcasting is possible via the probabilistic method. For example, we conveyed in Corollary 5.1 that for any indegree $d \geq 3$, any noise level $\delta < \delta_{\mathsf{maj}}$, and $L_k = \Theta(\log(k))$, there exists a deterministic DAG with all majority processing functions such that reconstruction of the root bit is possible. In fact, Proposition 5.2 showed that the scaling $L_k = \Theta(\log(k))$ is optimal for such DAGs where broadcasting is possible. Furthermore, for any $\delta \in (0, \frac{1}{2})$ and any sufficiently large bounded indegrees and outdegrees, we constructed explicit deterministic DAGs with $L_k = \Theta(\log(k))$ and all majority processing functions such that broadcasting is possible in Theorem 5.3. Our construction utilized regular bipartite lossless expander graphs between successive layers of the DAGs, and we showed that the constituent expander graphs can be generated in either deterministic quasi-polynomial time or randomized polylogarithmic time in

the number of levels. Finally, we made partial progress towards our conjecture that broadcasting is impossible in 2D regular grids where all vertices with two inputs use the same Boolean processing function. In particular, we proved impossibility results for 2D regular grids with all AND and all XOR processing functions in Theorems 5.4 and 5.5, respectively.

We close this discussion with a brief list of open problems that could serve as compelling directions for future research:

1. We conjectured in subsection 5.4.1 that in the random DAG model with $L_k = O(\log(k))$ and fixed $d \geq 3$, reconstruction is impossible for all choices of Boolean processing functions when $\delta \geq \delta_{\mathsf{maj}}$. Naturally, the analogous question for $d = 2$ is also open. Based on the reliable computation literature (see the discussion in subsection 5.4.1), we can conjecture that majority processing functions are optimal for odd $d \geq 3$, and alternating levels of AND and OR processing is optimal for $d = 2$, but it is not obvious which processing functions are optimal for general even $d \geq 4$.

2. We provided some evidence for the previous conjecture in the odd $d \geq 3$ case in part 2 of Proposition 5.1. A potentially simpler open question is to extend the proof of part 2 of Proposition 5.1 in appendix D.1 to show the impossibility of reconstruction using two (or more) vertices in the odd $d \geq 3$ case regardless of the choices of Boolean processing functions.

3. It is unknown whether a result similar to part 2 of Proposition 5.1 holds for even $d \geq 2$. For the $d = 2$ setting, a promising direction is to try and exploit the potential function contraction approach in [280] instead of the TV distance contraction approach in [86, 115].

4. As mentioned in subsection 5.4.2, it is an open problem to find a deterministic polynomial time algorithm to construct deterministic DAGs with sufficiently large d and $L_k = \Theta(\log(k))$ given some δ for which broadcasting is possible. Indeed, the deterministic algorithm in Theorem 5.3 takes quasi-polynomial time.

5. As indicated above, for fixed δ, Theorem 5.3 can only construct deterministic DAGs with sufficiently large d such that broadcasting is possible. However, Corollary 5.1 elucidates that such deterministic DAGs exist for every $d \geq 3$ as long as $\delta < \delta_{\mathsf{maj}}$. It is an open problem to efficiently construct deterministic DAGs with $L_k = \Theta(\log(k))$ for arbitrary $d \geq 3$ and $\delta < \delta_{\mathsf{maj}}$, or $d = 2$ and $\delta < \delta_{\mathsf{andor}}$, such that broadcasting is possible.

6. Recently, the ergodicity of 1D PCA with NAND gates was proved in [127, Theorem 1, Section 2] using a potential (or weight) function contraction approach. Exploiting this idea to prove impossibility of broadcasting in 2D regular grids with all NAND processing functions could be a fruitful direction of future research. In

fact, proving this would establish that broadcasting is impossible in the 2D regular grid model for all symmetric Boolean processing functions. The key difficulty, however, is finding appropriate potential functions in the 2D regular grid setting (which is nontrivial because noise is on the edges rather than on the vertices as in PCA). This difficulty could potentially be circumvented by computationally finding structured potential functions using *sum of squares* and *semidefinite programming* techniques.

7. Much like how the 2D regular grid with NAND processing functions corresponds to the 1D PCA with NAND gates, we can define a 2D 45-degree grid model with 3-input majority processing functions that corresponds to Gray's 1D PCA with 3-input majority gates [108, Example 5].[100] Although Gray's proof sketch of the ergodicity of his 1D PCA with 3-input majority gates in [108, Section 3] shows exponentially fast convergence, it obviously does not account for the boundary effects of 2D grids. It is therefore an interesting future endeavor to study broadcasting in the 2D 45-degree grid model with 3-input majority processing functions.

8. In view of our conjecture that broadcasting should be possible in 3D regular grids,[101] consider the 3D regular grid with all majority processing functions. If the boundary conditions of this 3D regular grid are removed, then a simple projection argument portrays that the resulting 2D PCA (with noise on the edges) uses Toom's NEC rule [274]. Since the standard 2D PCA (with noise on the vertices) that uses Toom's NEC rule is non-ergodic [274], it is an open problem to analogously establish the feasibility of broadcasting in the 3D regular grid with majority processing functions by modifying the simple version of Toom's proof in [92].

■ 5.11 Bibliographical Notes

Chapter 5 and appendix D are based primarily on the manuscript [185], and partly on the earlier draft [184, Theorems 3 and 4] (which will be extended into the forthcoming manuscript [183]). The work in [185] will also be published in part at the Proceedings of the IEEE International Symposium on Information Theory (ISIT) 2019 [186].

[100] A 2D 45-degree grid model has $L_k = 2k + 1$ vertices at each level $k \in \mathbb{N} \cup \{0\}$, and every vertex has three outgoing edges that have 45-degree separation between them. Furthermore, all vertices in the 2D 45-degree grid model that are not on or one step away from the boundary have three incoming edges.

[101] The vertex set of 3D regular grids is the intersection of the 3D integer lattice and the 3D nonnegative orthant.

Chapter 6

Conclusion and Future Directions

We conclude our study of information contraction and decomposition by providing a high-level overview of our main contributions. We began by introducing SDPIs for f-divergences in chapter 2, where we proved various properties of contraction coefficients of source-channel pairs, and notably, derived linear bounds on contraction coefficients of source-channel pairs in terms of maximal correlation. Then, we extended the notion of SDPIs for KL divergence in chapter 3 by developing sufficient conditions for less noisy domination by q-ary symmetric channels and illustrating the relationship between such domination and logarithmic Sobolev inequalities. Furthermore, we also established equivalent characterizations of the less noisy preorder using non-linear operator convex f-divergences in chapter 3. In chapter 4, we elucidated the geometry of SDPIs for χ^2-divergence by expounding the elegant modal decompositions of bivariate distributions. Specifically, we showed that maximal correlation functions are meaningful feature functions that decompose the information contained in categorical bivariate data, proposed the sample extended ACE algorithm for feature extraction and dimensionality reduction, and analyzed the sample complexity of this algorithm. (We also studied the peripherally related problem of reliable communication through permutation channels at the end of chapter 4.) Lastly, we investigated the discrete probability problem of broadcasting on bounded indegree DAGs in chapter 5, which corresponded to analyzing the contraction of TV distance in specific Bayesian networks. In particular, we proved the existence of bounded indegree DAGs with logarithmic layer size where broadcasting is possible using the probabilistic method, constructed deterministic DAGs where broadcasting is possible using regular bipartite lossless expander graphs, and established the impossibility of broadcasting in certain 2D regular grids.

As is also perhaps evident from our exposition, a guiding precept of this dissertation has been to consider topics and problems with rich histories. Indeed, the extensive literature reviews we presented for several different areas demonstrate this inherent proclivity. For example, section 2.2 in chapter 2 provided a survey of f-divergences and contraction coefficients, subsection 3.1.1 in chapter 3 contained an overview of information theoretic preorders over channels, section 4.5 in chapter 4 described various statistical ideas and techniques that are closely related to our feature extraction mechanism, and section 5.1 and the discussion in section 5.4 in chapter 5 explained the relevant literature on broadcasting and related fields such as Ising models, PCA, and reliable

computation and storage using noisy circuits. These meticulous collations of references for the aforementioned areas are collectively another one of our main contributions in this dissertation.

While each chapter in this dissertation closes with its own individual conclusion and copious directions of future research, we suggest some further avenues of future research in the ensuing two sections.

■ 6.1 SDPIs for f-Divergences over Bayesian Networks

As mentioned at the outset of chapter 5, a tight recursive bound on the the contraction coefficient for KL divergence η_{KL} (of channels) in Bayesian network settings was first developed by Evans and Schulman in [85] to prove impossibility results for reliable computation using noisy circuits. This bound was distilled in [231, Theorem 5], where a close connection to a certain site percolation process on the network was also established (see subsection 5.4.4 in chapter 5). Furthermore, an analogous result for the Dobrushin contraction coefficient η_{TV} was proved in [231, Theorem 8]. Hitherto, η_{KL} and η_{TV} are the only known cases where such recursive bounds have been proven. The proof of the η_{KL} case relies crucially on the chain rule for mutual information, cf. [85,231], while the proof of the η_{TV} case exploits Goldstein's simultaneously maximal coupling representation of the TV distance between two joint distributions, cf. [105, 231]. Neither of these two properties are shared by general f-divergences. Therefore, an open problem in the field of SDPIs is to establish a recursive bound on η_f (of channels) for any f-divergence over Bayesian networks.

Surprisingly, it turns out that such recursive bounds hold for non-linear operator convex f-divergences (despite these f-divergences not satisfying the chain rule or having known maximal coupling representations). Indeed, such bounds follow trivially as a consequence of the well-known result that $\eta_{\mathsf{KL}}(W) = \eta_f(W)$ for any channel $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ and every non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$ (cf. Proposition 2.6 in chapter 2). However, the broader problem of understanding the contraction of general f-divergences along Bayesian networks remains open.

■ 6.2 Potential Function Approach to Broadcasting and Related Problems

On a related but separate front, as conveyed by chapter 5, many commonly used models of broadcasting and PCA correspond to discrete-time Markov chains (with possibly uncountable state spaces). The ergodicity of these Markov chains, i.e. whether the distribution over the states of the Markov chain weakly converges over time, is one of the main questions of interest when studying these models. For instance, the impossibility of broadcasting is intuitively implied by the ergodicity (and sufficiently fast convergence rate) of the Markov chain defined by the underlying Bayesian network.

It is well-known in probability theory that ergodicity and related properties of discrete-time and time homogeneous Markov chains (with countable state spaces) can be analyzed using tools from the intimately related fields of *martingale theory*, *Lya-*

punov theory, and *potential theory*, cf. [36, Chapter 5]. For example, *Foster's theorem* characterizes the positive recurrence of Markov chains via the existence of *Lyapunov functions* with certain properties. It is also closely related to martingale convergence based criteria for recurrence of Markov chains. Indeed, such martingale arguments typically proceed by constructing a martingale from the Markov chain under consideration. One standard approach of doing this is to apply a *harmonic function*, which is an eigenfunction of the Markov chain's conditional expectation operator with eigenvalue 1, to the random variables defining the Markov chain—this is a specialization of the so called *Dynkin martingale* (or *Lévy's martingale*). The study of harmonic functions in classical analysis is known as potential theory, and variants of harmonic functions turn out to be the desired Lyapunov functions of Foster's theorem.

We refer to the use of martingale or Lyapunov function based arguments that apply general potential functions to Markov chains to study their ergodicity as the *potential function approach*. Recent developments in the PCA and reliable computation literatures have illustrated the effectiveness of the potential function approach in establishing impossibility results [127, 280, 281]. Indeed, as we outlined in section 5.10 of chapter 5, the authors of [127] prove that the 1D PCA with noisy NOR gates is ergodic by designing a potential function (or a variant of a Lyapunov function) for which the potentials of the states of the automaton over time form a supermartingale. Similarly, the author of [280] devises an appropriate potential function to characterize the noise threshold above which reliable computation is impossible using formulae with 2-input noisy gates. In fact, the potential function approach is currently perhaps the most promising approach to establishing similar noise thresholds for reliable computation using formulae with d-input noisy gates for general even $d \geq 4$. Despite the importance of this approach in proving impossibility results, there is no known systematic method for constructing "good" potential functions. Hence, an important future direction is to develop systematic ways of generating potential functions to prove ergodicity and other impossibility results in different models of broadcasting, reliable computation, and PCA. (Note that this is a much broader objective than the specific suggestions involving the potential function approach in section 5.10.)

In closing, we remark that research on broadcasting, reliable computation, PCA, and related areas is becoming particularly germane to our current times. For instance, new technologies such as quantum computation are genuinely requiring a better grasp of the behavior of noisy circuits. Furthermore, such research is also opening up avenues to better understand more fundamental concepts such as the notion of *quantum nonlocality* in physics, which is known to have deep connections with reliable computation, cf. [253]. Therefore, further work that develops sharper insights about information contraction in Bayesian networks, such as broadcasting DAGs and reliable computation networks, will potentially be a very fruitful future research endeavor.

Appendix A

Proofs from Chapter 2

A.1 Proof of Proposition 2.2

Proof. This proof is outlined in [11], and presented in [180, Theorem 3.2.4] for the $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ case. We provide it here for completeness.

Suppose the marginal pmfs of X and Y satisfy $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. We first show that the largest singular value of the DTM B is unity. Consider the matrix:

$$\begin{aligned} M &= \mathsf{diag}\Big(\sqrt{P_Y}\Big)^{-1} B^T B \, \mathsf{diag}\Big(\sqrt{P_Y}\Big) \\ &= \mathsf{diag}(P_Y)^{-1} W^T \mathsf{diag}(P_X)\, W \\ &= VW \end{aligned}$$

where $V = \mathsf{diag}(P_Y)^{-1} W^T \mathsf{diag}(P_X) \in \mathcal{P}_{\mathcal{X}|\mathcal{Y}}$ is the row stochastic reverse transition probability matrix of conditional pmfs $P_{X|Y}$. Observe that M has the same set of eigenvalues as the Gramian of the DTM $B^T B$, because we are simply using a similarity transformation to define it. As $B^T B$ is positive semidefinite, the eigenvalues of M and $B^T B$ are non-negative real numbers by the *spectral theorem* (see [129, Section 2.5]). Moreover, since V and W are both row stochastic, their product $M = VW$ is also row stochastic. Hence, the largest eigenvalue of M and $B^T B$ is unity by the *Perron-Frobenius theorem* (see [129, Chapter 8]). It follows that the largest singular value of B is also unity. Notice further that $\sqrt{P_X}$ and $\sqrt{P_Y}$ are the left and right singular vectors of B, respectively, corresponding to the singular value of unity. Indeed, we have:

$$\begin{aligned} \sqrt{P_X} B &= \sqrt{P_X}\, \mathsf{diag}\Big(\sqrt{P_X}\Big) W \mathsf{diag}\Big(\sqrt{P_Y}\Big)^{-1} = \sqrt{P_Y}, \\ B \sqrt{P_Y}^T &= \mathsf{diag}\Big(\sqrt{P_X}\Big) W \mathsf{diag}\Big(\sqrt{P_Y}\Big)^{-1} \sqrt{P_Y} = \sqrt{P_X}^T. \end{aligned}$$

Next, starting from Definition 2.3, let $f \in \mathbb{R}^{|\mathcal{X}|}$ and $g \in \mathbb{R}^{|\mathcal{Y}|}$ be the column vectors representing the range of the functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, respectively. Note that we can express the expectations in Definition 2.3 in terms of B, P_X, P_Y, and the vectors f and g:

$$\mathbb{E}\left[f(X)g(Y)\right] = \Big(\mathsf{diag}\Big(\sqrt{P_X}\Big)\, f\Big)^T B \left(\mathsf{diag}\Big(\sqrt{P_Y}\Big)\, g\right),$$

$$\begin{aligned}
\mathbb{E}\left[f(X)\right] &= \sqrt{P_X}\left(\mathsf{diag}\left(\sqrt{P_X}\right) f\right), \\
\mathbb{E}\left[g(Y)\right] &= \sqrt{P_Y}\left(\mathsf{diag}\left(\sqrt{P_Y}\right) g\right), \\
\mathbb{E}\left[f^2(X)\right] &= \left\|\mathsf{diag}\left(\sqrt{P_X}\right) f\right\|_2^2, \\
\mathbb{E}\left[g^2(Y)\right] &= \left\|\mathsf{diag}\left(\sqrt{P_Y}\right) g\right\|_2^2.
\end{aligned}$$

Letting $a = \mathsf{diag}(\sqrt{P_X})\, f$ and $b = \mathsf{diag}(\sqrt{P_Y})\, g$, we have from Definition 2.3:

$$\rho_{\mathsf{max}}(X;Y) = \max_{\substack{a\in\mathbb{R}^{|\mathcal{X}|},\, b\in\mathbb{R}^{|\mathcal{Y}|}: \\ \sqrt{P_X}a=\sqrt{P_Y}b=0 \\ \|a\|_2^2=\|b\|_2^2=1}} a^T B b$$

where the optimization is over all $a \in \mathbb{R}^{|\mathcal{X}|}$ and $b \in \mathbb{R}^{|\mathcal{Y}|}$ because $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Since a and b are orthogonal to the left and right singular vectors corresponding to the maximum singular value of unity of B, respectively, this maximization produces the second largest singular value of B using an alternative version (see [235, Lemma 2]) of the Courant-Fischer-Weyl min-max theorem (see Theorem C.1 in appendix C.1 or [129, Theorems 4.2.6 and 7.3.8]). This proves that $\rho_{\mathsf{max}}(X;Y)$ is the second largest singular value of the DTM when $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$.

We finally argue that one can assume $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ without loss of generality. When P_X or P_Y have zero entries, X and Y only take values in the support sets $\mathrm{supp}(P_X) \subseteq \mathcal{X}$ and $\mathrm{supp}\,(P_Y) \subseteq \mathcal{Y}$ respectively, which means that $P_X \in \mathcal{P}^{\circ}_{\mathrm{supp}(P_X)}$ and $P_Y \in \mathcal{P}^{\circ}_{\mathrm{supp}(P_Y)}$. Let B denote the "true" DTM of dimension $|\mathcal{X}| \times |\mathcal{Y}|$ corresponding to the pmf $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$, and B_{supp} denote the "support" DTM of dimension $|\,\mathrm{supp}(P_X)| \times |\,\mathrm{supp}(P_Y)|$ corresponding to the pmf $P_{X,Y}$ on $\mathrm{supp}(P_X) \times \mathrm{supp}(P_Y)$. Clearly, B can be constructed from B_{supp} by inserting zero vectors into the rows and columns associated with the zero probability letters in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Hence, B and B_{supp} have the same non-zero singular values (counting multiplicity), which implies that they have the same second largest singular value. This completes the proof. ■

■ A.2 Proof of Proposition 2.3

Proof.

Part 1: The normalization of contraction coefficients is evident from the non-negativity of f-divergences and their DPIs (2.17). We remark that in the case of $\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\mathsf{max}}(X;Y)^2$ (where we use (2.37)), $0 \le \rho_{\mathsf{max}}(X;Y) \le 1$ is Rényi's third axiom in defining maximal correlation [236].

Part 2: We provide a simple proof of this well-known property. Assume without loss of generality that $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ by ignoring any zero probability letters of $\mathcal{X}$. If the resulting $|\mathcal{X}| = 1$, then X is a constant *a.s.*, and the result follows trivially. So, we may also assume that $|\mathcal{X}| \ge 2$. Let $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denote the row stochastic transition probability matrix of the channel $P_{Y|X}$. Since W is unit rank (with all its columns

equal to P_Y) if and only if X and Y are independent (which means $P_{Y|X=x} = P_Y$ for every $x \in \mathcal{X}$), it suffices to show that W is unit rank if and only if $\eta_f(P_X, P_{Y|X}) = 0$.

To prove the forward direction, note that if W is unit rank, all its rows are equal to P_Y and we have $R_X W = P_Y$ for all $R_X \in \mathcal{P}_{\mathcal{X}}$. Hence, $\eta_f(P_X, P_{Y|X}) = 0$ using Definition 2.2, because $D_f(R_X W||P_X W) = 0$ for all input pmfs $R_X \in \mathcal{P}_{\mathcal{X}}$.

To prove the converse direction, we employ a slight variant of the argument in [180] that was used to prove the $\eta_{\mathsf{KL}}(P_X, P_{Y|X})$ case. For any $x \in \mathcal{X}$ and $\delta \in (0,1)$, consider $R_X = (1-\delta)\Delta_x + \delta \mathbf{u} \in \mathcal{P}_{\mathcal{X}}^{\circ}$, where δ is chosen such that $R_X \neq P_X$. Then, since $\eta_f(P_X, P_{Y|X}) = 0$ and $0 < D_f(R_X||P_X) < +\infty$, we have $D_f(R_X W||P_X W) = 0$ as f is strictly convex at unity. This implies that $(1-\delta)P_{Y|X=x} + \delta \mathbf{u} W = R_X W = P_X W = P_Y$. Letting $\delta \to 0$ shows that every row of W is equal to P_Y. Hence, W has unit rank.[102]

The converse direction can also be proved as follows. If $\eta_f(P_X, P_{Y|X}) = 0$, then $D_f(R_X W||P_X W) = 0$ for every $R_X \in \mathcal{P}_{\mathcal{X}}$ such that $D_f(R_X||P_X) < +\infty$ using Definition 2.2. So, $R_X W = P_X W$ for every $R_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ as f is strictly convex at unity (and $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ by assumption). This means that every $J_X \in (\mathbb{R}^{|\mathcal{X}|})^*$ satisfying $J_X \mathbf{1} = 0$ and $\|J_X\|_2 = 1$ belongs to the left nullspace of W. (This is because we can obtain any such J_X by defining $J_X = c\,(R_X - P_X)$ for some appropriate choice of $R_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $c \in \mathbb{R}\backslash\{0\}$, where the latter ensures that $\|J_X\|_2 = 1$.) Hence, W^T has nullity $|\mathcal{X}| - 1$, and W is therefore unit rank.

Finally, we also note that in the $\eta_{\chi^2}(P_X, P_{Y|X})$ case, this property of maximal correlation is Rényi's fourth axiom in [236].

Part 3: This part follows immediately from the ensuing lemmata.

Lemma A.1 (Decomposability and Maximal Correlation [5, 289]). *The joint pmf $P_{X,Y}$ is decomposable if and only if $\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\mathsf{max}}(X;Y)^2 = 1$.*

Proof. Although this result was proved in [5, 289], we provide a proof here for completeness. Suppose $P_{X,Y}$ is decomposable and there exist functions $h : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ such that $h(X) = g(Y)$ *a.s.* and $\mathbb{VAR}(h(X)) > 0$. Then, we may assume without loss of generality that $\mathbb{E}[h(X)] = 0$ and $\mathbb{E}[h^2(X)] = 1$, which implies that $\rho_{\mathsf{max}}(X;Y) = 1$ using Definition 2.3. So, we have $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$ by (2.37).

Conversely, suppose $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$, or equivalently, $\rho_{\mathsf{max}}(X;Y) = 1$ (by (2.37)). Let $h : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ be the functions that achieve $\rho_{\mathsf{max}}(X;Y)$—these functions exist when $\mathcal{X}$ and $\mathcal{Y}$ are finite because Definition 2.3 extremizes a continuous objective function over compact sets. Clearly, $h(X)$ and $g(Y)$ are zero mean, unit variance, and have Pearson correlation coefficient 1. This implies that $h(X) = g(Y)$ *a.s.* via a straightforward (and well-known) Cauchy-Schwarz argument. Therefore, $P_{X,Y}$ is decomposable. ■

Lemma A.2 (Simultaneous Extremality). *If f is strictly convex, twice differentiable at unity with $f''(1) > 0$, and $f(0) < \infty$, then $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$ if and only if $\eta_f(P_X, P_{Y|X}) = 1$.*

[102] We cannot simply execute the argument using $R_X = \Delta_x$ because $f(0)$ could be infinity.

Proof. The forward direction follows trivially from parts 1 and 7. To prove the converse direction, suppose $\eta_f(P_X, P_{Y|X}) = 1$. Consider the sequence of input pmfs $\{R_X^{(n)} \in \mathcal{P}_{\mathcal{X}} : 0 < D_f(R_X^{(n)}||P_X) < +\infty, n \in \mathbb{N}\}$ that achieves $\eta_f(P_X, P_{Y|X})$ in the limit:

$$\lim_{n\to\infty} \frac{D_f(R_X^{(n)}W||P_Y)}{D_f(R_X^{(n)}||P_X)} = 1$$

where $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denotes the row stochastic transition probability matrix corresponding to the channel $P_{Y|X}$. Using the sequential compactness of $\mathcal{P}_{\mathcal{X}}$, we may assume that $R_X^{(n)} \to R_X$ for some $R_X \in \mathcal{P}_{\mathcal{X}}$ as $n \to \infty$ (in the ℓ^2-norm sense) by passing to a subsequence if necessary. This leads to two possibilities:

Case 1: Suppose $R_X = P_X$. In this case, the proof of Theorem 2.1 in appendix A.3 illustrates that $\eta_f(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$. Thus, $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$.

Case 2: Suppose $R_X \neq P_X$. Since $f(0) < \infty$, f is strictly convex, and $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, we have $0 < D_f(R_X||P_X) < +\infty$. Hence, we get:

$$\lim_{n\to\infty} \frac{D_f(R_X^{(n)}W||P_Y)}{D_f(R_X^{(n)}||P_X)} = \frac{D_f(R_XW||P_Y)}{D_f(R_X||P_X)} = 1$$

using the continuity of f (which follows from its convexity). Now observe that since f is strictly convex and $0 < D_f(R_XW||P_XW) = D_f(R_X||P_X) < +\infty$, Y is a sufficient statistic of X for performing inference about the pair (R_X, P_X) (cf. [174, Theorem 14] or subsection 2.2.1), which in turn implies that (cf. [174, Theorem 14] or subsection 2.2.1):

$$0 < \chi^2(R_XW||P_XW) = \chi^2(R_X||P_X) < +\infty\,.$$

Therefore, $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$ using (2.36). ■

Lastly, we note that the η_{KL} case of this result was proved in [5].

Part 4: This is proven in [234, Proposition III.3].

Part 5: This is proven in [234, Theorem III.9]. We also note that two proofs of the tensorization property of η_{KL} can be found in [11], and a proof of the tensorization property of η_{χ^2} can be found in [289].

Part 6: To prove the first part, let $P_{U,X,Y}$ denote the joint pmf of (U, X, Y), and $S \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$ and $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denote the row stochastic transition probability matrices corresponding to the channels $P_{X|U}$ and $P_{Y|X}$, respectively. Then, $SW \in \mathcal{P}_{\mathcal{Y}|\mathcal{U}}$ is the row stochastic transition probability matrix corresponding to the channel $P_{Y|U}$ using the Markov property. Observe that for every pmf $R_U \in \mathcal{P}_{\mathcal{U}}\backslash\{P_U\}$:

$$D_f(R_USW||P_USW) \leq \eta_f(P_X, P_{Y|X})\,\eta_f(P_U, P_{X|U})\,D_f(R_U||P_U)$$

where $P_Y = P_USW$, $P_X = P_US$, and we use the SDPI (2.27) twice. Hence, we have:

$$\eta_f(P_U, P_{Y|U}) \leq \eta_f(P_U, P_{X|U})\,\eta_f(P_X, P_{Y|X})$$

using Definition 2.2.

The η_{χ^2} specialization of this result corresponds to the sub-multiplicativity property of the second largest singular value of the DTM. Such a sub-multiplicativity property also holds for the ith largest singular value of the DTM, cf. [149, Theorem 2], and is useful for distributed source and channel coding applications [149]. Moreover, the result in [149, Theorem 2] is also proved in [75, Theorem 3], where the relation to principal inertia components and maximal correlation is expounded.

To prove the second part, observe that for fixed $P_{X,Y}$, and every $P_{U|X}$ such that $U \to X \to Y$ form a Markov chain and $\eta_f(P_U, P_{X|U}) > 0$ (which requires that X is not a constant $a.s.$), we have:

$$\frac{\eta_f(P_U, P_{Y|U})}{\eta_f(P_U, P_{X|U})} \leq \eta_f(P_X, P_{Y|X}) \tag{A.1}$$

using the sub-multiplicativity property established above. Let $U = X$ $a.s.$ so that $P_{U|X} \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ is the identity matrix. Then, $\eta_f(P_U, P_{X|U}) = 1$ and $\eta_f(P_U, P_{Y|U}) = \eta_f(P_X, P_{Y|X})$ using Definition 2.2. Therefore, equality can be achieved in (A.1), and the proof is complete.

We remark that the η_{χ^2} case of this result is presented in [16, Lemma 6], where the authors also prove that the optimal channel $P_{U|X}$ can be taken as $P_{Y|X}$ (so that U is a copy of Y) instead of the identity matrix (where $U = X$ $a.s.$).

Part 7: Following the remark after [189, Theorem 5], we prove this result via the technique used to prove the η_{KL} case in [189, Theorem 5].

Let $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ denote the row stochastic matrix of the channel $P_{Y|X}$, and B denote the DTM of the joint pmf $P_{X,Y}$. Let us define a trajectory of spherically perturbed pmfs of the form (2.22):

$$R_X^{(\epsilon)} = P_X + \epsilon \, K_X \, \mathsf{diag}\Big(\sqrt{P_X}\Big)$$

where $K_X \in \mathcal{S} \triangleq \{x \in (\mathbb{R}^{|\mathcal{X}|})^* : \sqrt{P_X} x^T = 0, \|x\|_2 = 1\}$ is a spherical perturbation vector. When these pmfs pass through the channel W, we get the output trajectory:

$$R_X^{(\epsilon)} W = P_Y + \epsilon \, K_X B \, \mathsf{diag}\Big(\sqrt{P_Y}\Big) \tag{A.2}$$

where B maps input spherical perturbations to output spherical perturbations [139]. Now, starting from Definition 2.2, we have:

$$\begin{aligned}
\eta_f(P_X, P_{Y|X}) &= \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}}: \\ 0 < D_f(R_X||P_X) < +\infty}} \frac{D_f(R_X W||P_X W)}{D_f(R_X||P_X)} \\
&\geq \liminf_{\epsilon \to 0} \sup_{K_X \in \mathcal{S}} \frac{\|K_X B\|_2^2 + o(1)}{\|K_X\|_2^2 + o(1)} \\
&\geq \sup_{K_X \in \mathcal{S}} \liminf_{\epsilon \to 0} \frac{\|K_X B\|_2^2 + o(1)}{1 + o(1)}
\end{aligned}$$

$$= \eta_{\chi^2}(P_X, P_{Y|X})$$
$$= \rho_{\mathsf{max}}(X;Y)^2$$

where the second inequality follows from (2.25) after restricting the supremum over all pmfs of the form (2.22) (where $\epsilon \neq 0$ is some sufficiently small fixed value) and then letting $\epsilon \to 0$, the third inequality follows from the minimax inequality, and the final two equalities follow from (2.39) and (2.37), respectively. This completes the proof.

We remark that the $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$ assumptions, while useful for defining the aforementioned trajectories of pmfs, are not essential for this result. For the special case of η_{KL}, this result was first proved in [5], and then again in [147] and [189, Theorem 5]—the latter two proofs both use perturbation arguments with different flavors. ■

■ A.3 Proof of Theorem 2.1

Proof. We begin by defining the function $\tau : (0, \infty) \to [0, 1]$:

$$\tau(\delta) \triangleq \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}}: \\ 0 < D_f(R_X||P_X) \leq \delta}} \frac{D_f(R_X W||P_X W)}{D_f(R_X||P_X)}$$

so that what we seek to prove is:

$$\lim_{n\to\infty} \tau(\delta_n) = \eta_{\chi^2}(P_X, P_{Y|X})$$

for any decreasing sequence $\{\delta_n > 0 : n \in \mathbb{N}\}$ such that $\lim_{n\to\infty} \delta_n = 0$. Note that the limit on the left hand side exists because as $\delta_n \to 0$, the supremum in $\tau(\delta_n)$ is non-increasing and bounded below by 0.

We first prove that $\lim_{n\to\infty} \tau(\delta_n) \geq \eta_{\chi^2}(P_X, P_{Y|X})$. To this end, consider a trajectory of spherically perturbed pmfs of the form (2.22):

$$R_X^{(n)} = P_X + \epsilon_n K_X \,\mathsf{diag}\Big(\sqrt{P_X}\Big)$$

where $K_X \in \mathcal{S} = \{x \in (\mathbb{R}^{|\mathcal{X}|})^* : \sqrt{P_X} x^T = 0, \|x\|_2 = 1\}$ is a spherical perturbation vector. The associated trajectory of output pmfs after passing through W is given by (A.2):

$$R_X^{(n)} W = P_Y + \epsilon_n K_X B \,\mathsf{diag}\Big(\sqrt{P_Y}\Big)$$

where B denotes the DTM corresponding to $P_{X,Y}$. We ensure that the scalars $\{\epsilon_n \neq 0 : n \in \mathbb{N}\}$ that define our trajectory satisfy $\lim_{n\to\infty} \epsilon_n = 0$ and are sufficiently small such that:

$$D_f(R_X^{(n)}||P_X) = \frac{f''(1)}{2}\epsilon_n^2 \|K_X\|_2^2 + o(\epsilon_n^2) \leq \delta_n$$

where we use (2.25) (and the fact that $f''(1)$ exists and is strictly positive). By definition of τ, we have:

$$\sup_{K_X \in \mathcal{S}} \frac{D_f(R_X^{(n)} W||P_X W)}{D_f(R_X^{(n)}||P_X)} \leq \tau(\delta_n)$$

$$\lim_{n\to\infty} \sup_{K_X\in\mathcal{S}} \frac{\frac{f''(1)}{2}\epsilon_n^2 \|K_X B\|_2^2 + o(\epsilon_n^2)}{\frac{f''(1)}{2}\epsilon_n^2 \|K_X\|_2^2 + o(\epsilon_n^2)} \leq \lim_{n\to\infty} \tau(\delta_n)$$

$$\lim_{n\to\infty} \sup_{K_X\in\mathcal{S}} \frac{\|K_X B\|_2^2 + o(1)}{1+o(1)} \leq \lim_{n\to\infty} \tau(\delta_n)$$

$$\eta_{\chi^2}(P_X, P_{Y|X}) \leq \lim_{n\to\infty} \tau(\delta_n)$$

where the second inequality uses (2.25) for both the numerator and denominator, and the final inequality uses the singular value characterization of $\eta_{\chi^2}(P_X, P_{Y|X})$ in (2.39).

We next prove that $\lim_{n\to\infty}\tau(\delta_n) \leq \eta_{\chi^2}(P_X, P_{Y|X})$. Observe that for each $n\in\mathbb{N}$, there exists a pmf $R_X^{(n)} \in \mathcal{P}_{\mathcal{X}}$ satisfying two properties:

1. $0 < D_f(R_X^{(n)}||P_X) \leq \delta_n$
2. $0 \leq \tau(\delta_n) - \dfrac{D_f(R_X^{(n)}W||P_XW)}{D_f(R_X^{(n)}||P_X)} \leq \dfrac{1}{2^n}$

where the first property holds because $R_X \mapsto D_f(R_X||P_X)$ is a continuous map for fixed $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ (which follows from the convexity of f), and the second property holds because $\tau(\delta_n)$ is defined as a supremum. Since $\tau(\delta_n)$ converges as $n\to\infty$, we have:[103]

$$\lim_{n\to\infty} \frac{D_f(R_X^{(n)}W||P_XW)}{D_f(R_X^{(n)}||P_X)} = \lim_{n\to\infty}\tau(\delta_n)\,. \tag{A.3}$$

Using the sequential compactness of $\mathcal{P}_{\mathcal{X}}$, we can assume that $R_X^{(n)}$ converges as $n\to\infty$ (in the ℓ^2-norm sense) by passing to a subsequence if necessary. Since $D_f(R_X^{(n)}||P_X)\to 0$ as $n\to\infty$, we have that $\lim_{n\to\infty} R_X^{(n)} = P_X$ due to the continuity of $R_X \mapsto D_f(R_X||P_X)$ for fixed $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ and the fact that an f-divergence (where f is strictly convex at unity) is zero if and only if its input pmfs are equal. Let us define the spherical perturbation vectors $\{K_X^{(n)} \in \mathcal{S} : n\in\mathbb{N}\}$ using the relation:

$$R_X^{(n)} = P_X + \epsilon_n\, K_X^{(n)} \mathsf{diag}\Big(\sqrt{P_X}\Big)$$

where $\{\epsilon_n \neq 0 : n\in\mathbb{N}\}$ provide the appropriate scalings, and $\lim_{n\to\infty}\epsilon_n = 0$ (since $\lim_{n\to\infty} R_X^{(n)} = P_X$). The corresponding output pmfs are of the form (A.2) mutatis mutandis, and we can approximate the ratio between output and input f-divergences as before using (2.25):

$$\frac{D_f(R_X^{(n)}W||P_XW)}{D_f(R_X^{(n)}||P_X)} = \frac{\frac{f''(1)}{2}\epsilon_n^2 \left\|K_X^{(n)}B\right\|_2^2 + o(\epsilon_n^2)}{\frac{f''(1)}{2}\epsilon_n^2 \left\|K_X^{(n)}\right\|_2^2 + o(\epsilon_n^2)}$$

[103] Here, we use the fact that if two sequences $\{a_n\in\mathbb{R} : n\in\mathbb{N}\}$ and $\{b_n\in\mathbb{R} : n\in\mathbb{N}\}$ satisfy $\lim_{n\to\infty}|a_n - b_n| = 0$ and $\lim_{n\to\infty} b_n = b \in \mathbb{R}$, then $\lim_{n\to\infty} a_n = b$.

$$= \frac{\left\|K_X^{(n)} B\right\|_2^2 + o(1)}{1 + o(1)} .$$

Using the sequential compactness of $\mathcal{S}$, we may assume that $\lim_{n\to\infty} K_X^{(n)} = K_X^\star \in \mathcal{S}$ by passing to a subsequence if necessary. Hence, letting $n \to \infty$, we get:

$$\lim_{n\to\infty} \tau(\delta_n) = \|K_X^\star B\|_2^2 \leq \eta_{\chi^2}(P_X, P_{Y|X})$$

where the equality follows from (A.3) and the continuity of the map $\left(\mathbb{R}^{|\mathcal{X}|}\right)^* \ni x \mapsto \|xB\|_2^2$, and the inequality follows from (2.39). This completes the proof. ■

■ A.4 Proof of Corollary 2.1

Proof. The convex function $f : (0,\infty) \to \mathbb{R}$, $f(t) = t\log(t)$ is clearly strictly convex and thrice differentiable at unity with $f(1) = 0$, $f'(1) = 1$, $f''(1) = 1 > 0$, and $f'''(1) = -1$. Moreover, the function $g : (0,\infty) \to \mathbb{R}$, $g(x) = \frac{f(x)-f(0)}{x} = \log(x)$ is clearly concave (where $f(0) = \lim_{t\to 0^+} f(t) = 0$). So, to prove Corollary 2.1 using Theorem 2.2, it suffices to show that f satisfies (2.80) for every $t \in (0,\infty)$ (cf. [101]):

$$(f(t) - f'(1)(t-1))\left(1 - \frac{f'''(1)}{3f''(1)}(t-1)\right) \geq \frac{f''(1)}{2}(t-1)^2$$

which simplifies to:

$$2t(t+2)\log(t) - (5t+1)(t-1) \geq 0 .$$

Define $h : (0,\infty) \to \mathbb{R}$, $h(t) = 2t(t+2)\log(t) - (5t+1)(t-1)$ and observe that:

$$h'(t) = 4(t+1)\log(t) - 8(t-1)$$
$$h''(t) = 4\log(t) + \frac{4}{t} - 4 \geq 0$$

where the non-negativity of the second derivative follows from the well-known inequality:

$$\forall x > 0, \quad x\log(x) \geq x - 1 .$$

Since h is convex (as its second derivative is non-negative) and $h(1) = h'(1) = 0$, $t = 1$ is a global minimizer of h and $h(t) \geq 0$ for every $t \in (0,\infty)$ as required.

Finally, we can verify that the constant in Corollary 2.1 is:

$$\frac{f'(1) + f(0)}{f''(1)\min\limits_{x\in\mathcal{X}} P_X(x)} = \frac{1}{\min\limits_{x\in\mathcal{X}} P_X(x)}$$

which completes the proof. ■

■ A.5 Proof of (2.83)

Proof. Two proofs for (2.83) are provided in our conference paper [189, Lemma 6]. We present the one with a convex analysis flavor. It involves recognizing that KL divergence is a Bregman divergence associated with the negative Shannon entropy function, and then exploiting the strong convexity of the negative Shannon entropy function to bound KL divergence. Let $H_{\mathsf{neg}} : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$ be the negative *Shannon entropy* function, which is defined as:

$$\forall Q_X \in \mathcal{P}_{\mathcal{X}}, \;\; H_{\mathsf{neg}}(Q_X) \triangleq \sum_{x \in \mathcal{X}} Q_X(x) \log(Q_X(x)) \, .$$

Since the *Bregman divergence* corresponding to H_{neg} is the KL divergence, cf. [21], we have for all $S_X \in \mathcal{P}_{\mathcal{X}}$ and $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$:

$$D(S_X||Q_X) = H_{\mathsf{neg}}(S_X) - H_{\mathsf{neg}}(Q_X) - J_X \nabla H_{\mathsf{neg}}(Q_X)$$

where $J_X = S_X - Q_X$ is an additive perturbation vector, and $\nabla H_{\mathsf{neg}} : \mathcal{P}_{\mathcal{X}}^{\circ} \to \mathbb{R}^{|\mathcal{X}|}$ is the gradient of H_{neg}. Moreover, as H_{neg} is twice continuously differentiable, we have:

$$\forall Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}, \;\; \nabla^2 H_{\mathsf{neg}}(Q_X) = \mathsf{diag}(Q_X)^{-1} \succeq_{\mathsf{PSD}} I$$

where $\nabla^2 H_{\mathsf{neg}} : \mathcal{P}_{\mathcal{X}}^{\circ} \to \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ denotes the Hessian matrix of H_{neg}, and $I \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ denotes the identity matrix. (Note that $\mathsf{diag}(Q_X)^{-1} - I$ is positive semidefinite because it is a diagonal matrix with non-negative diagonal entries.) Recall from [34, Chapter 9] that a twice continuously differentiable convex function $f : S \to R$ with open domain $S \subseteq \mathbb{R}^n$ is called *strongly convex* if there exists $m > 0$ such that for all $x \in S$, $\nabla^2 f(x) \succeq mI$. This means that H_{neg} is strongly convex on $\mathcal{P}_{\mathcal{X}}^{\circ}$. A consequence of this strong convexity is the following quadratic lower bound [34, Chapter 9]:

$$\begin{aligned} & H_{\mathsf{neg}}(S_X) \geq H_{\mathsf{neg}}(Q_X) + J_X \nabla H_{\mathsf{neg}}(Q_X) + \frac{1}{2} \|J_X\|_2^2 \\ \Leftrightarrow \quad & D(S_X||Q_X) \geq \frac{1}{2} \|J_X\|_2^2 \end{aligned} \tag{A.4}$$

for every $S_X \in \mathcal{P}_{\mathcal{X}}$ and $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, where we allow $S_X \in \mathcal{P}_X \backslash \mathcal{P}_X^{\circ}$ due to the continuity of H_{neg}. This is precisely what we get if we loosen (2.76) in the proof of Lemma 2.2 using $\|J_X\|_1 \geq \|J_X\|_2$ and (2.75). Finally, we have for every $S_X \in \mathcal{P}_{\mathcal{X}}$ and $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$:

$$D(S_X||Q_X) \geq \frac{1}{2} \|J_X\|_2^2 \geq \frac{\min\limits_{x \in \mathcal{X}} Q_X(x)}{2} \chi^2(S_X||Q_X)$$

where the second inequality follows from (2.9). This trivially holds for all $Q_X \in \mathcal{P}_{\mathcal{X}} \backslash \mathcal{P}_{\mathcal{X}}^{\circ}$ as well. ■

Appendix B

Supplementary Results and Proofs from Chapter 3

■ B.1 Basics of Majorization Theory

Since we use some majorization arguments in our analysis, we first briefly introduce the notion of *group majorization* over row vectors in $(\mathbb{R}^q)^*$ (with $q \in \mathbb{N}$) in this appendix. Given a group $\mathcal{G} \subseteq \mathbb{R}^{q\times q}$ of matrices (with the operation of matrix multiplication), we may define a preorder called $\mathcal{G}$-majorization over row vectors in $(\mathbb{R}^q)^*$. For two row vectors $x, y \in (\mathbb{R}^q)^*$, we say that x $\mathcal{G}$-majorizes y if $y \in \mathsf{conv}(\{xG : G \in \mathcal{G}\})$, where $\{xG : G \in \mathcal{G}\}$ is the orbit of x under the group $\mathcal{G}$. Group majorization intuitively captures a notion of "spread" of vectors. So, x $\mathcal{G}$-majorizes y when x is more spread out than y with respect to $\mathcal{G}$. We refer readers to [195, Chapter 14, Section C] and the references therein for a thorough treatment of group majorization. If we let $\mathcal{G}$ be the symmetric group of all permutation matrices in $\mathbb{R}^{q\times q}$, then $\mathcal{G}$-majorization corresponds to traditional majorization of vectors in $(\mathbb{R}^q)^*$ as introduced in [121]. The next proposition collects some results about traditional majorization.

Proposition B.1 (Majorization [121,195]). *Given two row vectors $x = (x_1, \ldots, x_q) \in (\mathbb{R}^q)^*$ and $y = (y_1, \ldots, y_q) \in (\mathbb{R}^q)^*$, let $x_{(1)} \leq \cdots \leq x_{(q)}$ and $y_{(1)} \leq \cdots \leq y_{(q)}$ denote the re-orderings of x and y in ascending order. Then, the following are equivalent:*

1. *x majorizes y, or equivalently, y resides in the convex hull of all permutations of x.*
2. *$y = xD$ for some doubly stochastic matrix $D \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$.*
3. *The entries of x and y satisfy:*

$$\sum_{i=1}^{k} x_{(i)} \leq \sum_{i=1}^{k} y_{(i)}\,, \quad \textit{for } k = 1, \ldots, q-1\,,$$
$$\textit{and} \quad \sum_{i=1}^{q} x_{(i)} = \sum_{i=1}^{q} y_{(i)}\,.$$

When these conditions are true, we write $x \succeq_{\mathsf{maj}} y$.

In the context of subsection 3.1.2, given an Abelian group $(\mathcal{X}, \oplus)$ of order q, another useful notion of $\mathcal{G}$-majorization can be obtained by letting $\mathcal{G} = \{P_z \in \mathbb{R}^{q\times q} : z \in \mathcal{X}\}$ be the group of permutation matrices defined in (3.9) that is isomorphic to $(\mathcal{X}, \oplus)$. For such choice of $\mathcal{G}$, we write $x \succeq_{\mathcal{X}} y$ when x $\mathcal{G}$-majorizes (or $\mathcal{X}$-majorizes) y for any two row vectors $x, y \in (\mathbb{R}^q)^*$. We will only require one fact about such group majorization, which we present in the next proposition.

Proposition B.2 (Group Majorization). *Given two row vectors $x, y \in (\mathbb{R}^q)^*$, $x \succeq_{\mathcal{X}} y$ if and only if there exists $\lambda \in \mathcal{P}_q$ such that $y = x\,\mathsf{circ}_{\mathcal{X}}(\lambda)$.*

Proof. Observe that:

$$\begin{aligned} x \succeq_{\mathcal{X}} y &\Leftrightarrow y \in \mathsf{conv}(\{xP_z : z \in \mathcal{X}\}) \\ &\Leftrightarrow y = \lambda\,\mathsf{circ}_{\mathcal{X}}(x) \;\text{ for some }\; \lambda \in \mathcal{P}_q \\ &\Leftrightarrow y = x\,\mathsf{circ}_{\mathcal{X}}(\lambda) \;\text{ for some }\; \lambda \in \mathcal{P}_q \end{aligned}$$

where the second step follows from (3.14), and the final step follows from the commutativity of $\mathcal{X}$-circular convolution. ∎

Proposition B.2 parallels the equivalence between parts 1 and 2 of Proposition B.1, because $\mathsf{circ}_{\mathcal{X}}(\lambda)$ is a doubly stochastic matrix for every pmf $\lambda \in \mathcal{P}_q$. In closing our discussion of group majorization, we mention a well-known special case of the version of group majorization in Proposition B.2. When $(\mathcal{X}, \oplus)$ is the cyclic Abelian group $\mathbb{Z}/q\mathbb{Z}$ of integers with addition modulo q, $\mathcal{G} = \{I_q, P_q, P_q^2, \ldots, P_q^{q-1}\}$ is the group of all cyclic permutation matrices in $\mathbb{R}^{q\times q}$, where $P_q \in \mathbb{R}^{q\times q}$ is defined in (3.15). The corresponding notion of $\mathcal{G}$-majorization is known as *cyclic majorization*, cf. [102].

We next introduce a variant of the standard notion of majorization (presented in Proposition B.1) known as *weak majorization*. As we will see, weak majorization is indeed "weaker" than the traditional majorization preorder in Proposition B.1, which is sometimes referred to as *strong majorization*. We will exploit weak majorization to derive singular value stability results in appendix C.1.

Given two row vectors $x = (x_1, \ldots, x_q) \in (\mathbb{R}^q)^*$ and $y = (y_1, \ldots, y_q) \in (\mathbb{R}^q)^*$, let $x_{[1]} \geq \cdots \geq x_{[q]}$ and $y_{[1]} \geq \cdots \geq y_{[q]}$ denote the re-orderings of x and y in descending order. We say that x weakly majorizes y if for every $k \in \{1, \ldots, q\}$, cf. [195]:

$$\sum_{i=1}^{k} x_{[i]} \geq \sum_{i=1}^{k} y_{[i]}\,. \tag{B.1}$$

It is worth comparing this definition to part 3 of Proposition B.1. In particular, when the inequality corresponding to $k = q$ in (B.1) is an equality, it is straightforward to verify that the weak majorization preorder reduces to the standard strong majorization preorder. While there are equivalent characterizations of weak majorization analogous to Proposition B.1, cf. [195], an indispensable characterization of weak majorization is via *Karamata's inequality* (or the *majorization inequality*).

Proposition B.3 (Karamata's Inequality [195]). *Given two row vectors $x, y \in (\mathbb{R}^q)^*$, x weakly majorizes y if and only if for every convex non-decreasing function $f : \mathbb{R} \to \mathbb{R}$, we have:*

$$\sum_{i=1}^{q} f(x_i) \geq \sum_{i=1}^{q} f(y_i)\,.$$

We will use Proposition B.3 to prove a class of singular value stability inequalities in appendix C.1. In closing this appendix, we remark that an analogue of Proposition B.3 for strong majorization is expounded in [195].

■ B.2 Elements of Operator Monotonicity and Operator Convexity

In this appendix, we present some preliminaries on operator monotone and operator convex functions that will be useful in our analysis. For any non-empty (finite or infinite) open or closed interval $I \subseteq \mathbb{R}$, let $\mathbb{C}^{n\times n}_{\mathsf{Herm}}(I)$ denote the set of all $n \times n$ Hermitian matrices with all eigenvalues in I, where $\mathbb{C}^{n\times n}_{\mathsf{Herm}} = \mathbb{C}^{n\times n}_{\mathsf{Herm}}(\mathbb{R})$ is the set of all Hermitian matrices. Given a function $f : I \to \mathbb{R}$, we can extend it to a function $f : \mathbb{C}^{n\times n}_{\mathsf{Herm}}(I) \to \mathbb{C}^{n\times n}_{\mathsf{Herm}}$ as follows [27, Chapter V.1]:

$$\forall A \in \mathbb{C}^{n\times n}_{\mathsf{Herm}}(I), \quad f(A) \triangleq U\,\mathsf{diag}(f(\lambda_1), \ldots, f(\lambda_n))\,U^H \tag{B.2}$$

where $A = U\,\mathsf{diag}(\lambda_1, \ldots, \lambda_n)\,U^H$ is the spectral decomposition of A with real eigenvalues $\lambda_1, \ldots, \lambda_n \in I$, and $U \in \mathcal{V}_n(\mathbb{C}^n)$ is a unitary matrix. We say that f is *operator monotone* if for every $n \in \mathbb{N}$, and every pair of matrices $A, B \in \mathbb{C}^{n\times n}_{\mathsf{Herm}}(I)$:

$$A \succeq_{\mathsf{PSD}} B \;\;\Rightarrow\;\; f(A) \succeq_{\mathsf{PSD}} f(B) \tag{B.3}$$

where $\succeq_{\mathsf{PSD}}$ denotes the Löwner partial order [27, Chapter V.1]. Similarly, we say that f is *operator convex* if for every $n \in \mathbb{N}$, every pair of matrices $A, B \in \mathbb{C}^{n\times n}_{\mathsf{Herm}}(I)$, and every $\lambda \in [0, 1]$, cf. [27, Chapter V.1]:[104]

$$\lambda f(A) + (1-\lambda) f(B) \succeq_{\mathsf{PSD}} f(\lambda A + (1-\lambda)B)\,. \tag{B.4}$$

Note that an operator monotone, respectively convex, function $f : I \to \mathbb{R}$ is clearly monotone, respectively convex, and its translated affine transformations $g : \{c + x : x \in I\} \to \mathbb{R}$, $g(t) = af(t-c) + b$ are also operator monotone, respectively convex, for every $a \geq 0$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$.

However, it is not clear from these definitions that nontrivial operator monotone and operator convex functions exist. To remedy this, the celebrated *Löwner-Heinz theorem* exhibits several nontrivial examples of such functions [124, 178]. We next present the relevant aspects of this result that we will utilize in this thesis, cf. [40, Theorem 2.6], [128, Section 6.6, Problem 17], [27, Theorems V.2.5 and V.2.10, Exercises V.2.11 and V.2.13]. (Note that we have applied an affine transformation to part 2 below for convenience.)

[104] It is straightforward to verify that $A, B \in \mathbb{C}^{n\times n}_{\mathsf{Herm}}(I)$ implies that $\lambda A + (1-\lambda)B \in \mathbb{C}^{n\times n}_{\mathsf{Herm}}(I)$ for all $\lambda \in [0, 1]$.

Theorem B.1 (Löwner-Heinz theorem [40, 124, 178]). *The following are true:*

1. *For every $p \in [0,1]$, the function $f : [0,\infty) \to \mathbb{R}$, $f(t) = t^p$ is operator monotone.*
2. *For every $\alpha \in (0,1) \cup (1,2]$, the function $f : (0,\infty) \to \mathbb{R}$, $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$ is operator convex.*
3. *The function $f : (0,\infty) \to \mathbb{R}$, $f(t) = t\log(t)$ is operator convex.*

Parts 1 and 2 of Theorem B.1 illustrate that operator monotonicity and operator convexity are very closely related to each other, and we refer readers to the various results in [27, Chapter V] for concrete statements.

A striking property of operator monotone and operator convex functions is that they are characterized by certain *integral representations*—see *Löwner's theorems* in [27, Chapter V.4, Problem V.5.5]. These representations are based on deep results from complex analysis concerning the theory of *Pick-Herglotz-Nevanlinna functions*. The ensuing lemma presents one such integral representation for operator convex functions which follows from [46, Equation (7)] and the associated references.

Lemma B.1 (Löwner's Integral Representation [46]). *For every operator convex function $f : (0,\infty) \to \mathbb{R}$ with $f(1) = 0$, there exist constants $a \in \mathbb{R}$ and $b \geq 0$, and a finite positive measure μ on $(1,\infty)$ (with its Borel σ-algebra) such that:*

$$\forall t > 0, \quad f(t) = a(t-1) + b(t-1)^2 + \int_{(1,\infty)} \frac{(t-1)(\omega t - \omega - 1)}{t + \omega - 1}\, d\mu(\omega)\,. \tag{B.5}$$

We remark that our f is related to g in [46] by $f(t) = g(t-1)$, and we "normalize" f so that $f(1) = 0$ to ensure that it can be used to define an f-divergence. Furthermore, as noted in [46], the converse also holds, i.e. functions of the form (B.5) are operator convex.

Lemma B.1 can be exploited to derive extremely useful integral characterizations of operator convex f-divergences. Indeed, the next lemma distills such a characterization from [46, p.33] and presents it in a more transparent form.

Lemma B.2 (Integral Representation of f-Divergences [46, p.33]). *Consider any f-divergence such that $f : (0,\infty) \to \mathbb{R}$ is operator convex and satisfies $f(1) = 0$. Then, there exists a constant $b \geq 0$ and a finite positive measure τ on $(0,1)$ (with its Borel σ-algebra) such that for every $R_X, P_X \in \mathcal{P}_\mathcal{X}$:*

$$D_f(R_X||P_X) = b\,\chi^2(R_X||P_X) + \int_{(0,1)} \frac{1+\lambda^2}{\lambda(1-\lambda)}\, \mathsf{LC}_\lambda(R_X||P_X)\, d\tau(\lambda)\,.$$

Proof. Fix any two pmfs $R_X, P_X \in \mathcal{P}_\mathcal{X}$, and suppose that the random variable X has pmf P_X. Then, since f exhibits the integral representation (B.5) in Lemma B.1, we substitute $t = R_X(X)/P_X(X)$ into (B.5) and take expectations to get:

$$\mathbb{E}\left[f\left(\frac{R_X(X)}{P_X(X)}\right)\right] = b\,\mathbb{E}\left[\left(\frac{R_X(X)}{P_X(X)} - 1\right)^2\right] + \int_{(1,\infty)} \mathbb{E}\left[\frac{\left(\frac{R_X(X)}{P_X(X)} - 1\right)\left(\omega\frac{R_X(X)}{P_X(X)} - \omega - 1\right)}{\frac{R_X(X)}{P_X(X)} + \omega - 1}\right] d\mu(\omega)$$

where the first term on the right hand side of (B.5) vanishes after taking expectations (see the affine invariance property in subsection 2.2.1). This implies that:

$$D_f(R_X||P_X) = b\,\chi^2(R_X||P_X) + \int_{(1,\infty)} \mathbb{E}\left[\frac{(1+\omega^2)\left(\frac{R_X(X)}{P_X(X)}-1\right)^2}{\omega\left(\frac{R_X(X)}{P_X(X)}+\omega-1\right)}\right] d\mu(\omega)$$

where the left hand side follows from Definition 2.1, the χ^2-divergence term follows from the definition in subsection 2.2.1, and the last term follows from the affine invariance property in subsection 2.2.1 and the relation:

$$\forall t,\omega>0,\quad \frac{(t-1)(\omega t-\omega-1)}{t+\omega-1} = \frac{(1+\omega^2)(t-1)^2}{\omega(t+\omega-1)} - \frac{t-1}{\omega}\,.$$

Next, observe that the change of variables $\omega = \frac{1}{\lambda}$ yields:

$$D_f(R_X||P_X) = b\,\chi^2(R_X||P_X) + \int_{(0,1)} \mathbb{E}\left[\frac{(1+\lambda^2)\left(\frac{R_X(X)}{P_X(X)}-1\right)^2}{\left(\lambda\frac{R_X(X)}{P_X(X)}+1-\lambda\right)}\right] d\tau(\lambda)$$

for some finite positive measure τ on $(0,1)$. Finally, recognizing that the integrand on the right hand side is a scaled Vincze-Le Cam divergence (see subsection 2.2.1), some straightforward algebra produces the desired integral representation. ■

Lemma B.2 is used in [46, p.33] (in a slightly different form) to prove Proposition 2.6, cf. [46, Theorem 1]. Furthermore, [234, p.3363] also distills the key idea in [46, p.33] and presents an alternative integral representation (in terms of Vincze-Le Cam and χ^2-divergences) analogous to Lemma B.2. However, the representation in [234, p.3363] only holds for operator convex functions f where $f(0)$ is finite, while Lemma B.2 holds for infinite $f(0)$ as well.

■ B.3 Proof of Proposition 3.4

Proof.

Part 1: This is obvious from (3.19).

Part 2: Since the DFT matrix jointly diagonalizes all circulant matrices, it diagonalizes every W_δ for $\delta\in\mathbb{R}$ (using part 1). The corresponding eigenvalues are all real because W_δ is symmetric. To explicitly compute these eigenvalues, we refer to [129, Problem 2.2.P10]. Observe that for any row vector $x=(x_0,\ldots,x_{q-1})\in(\mathbb{R}^q)^*$, the corresponding circulant matrix satisfies:

$$\begin{aligned}\mathsf{circ}_{\mathbb{Z}/q\mathbb{Z}}(x) &= \sum_{k=0}^{q-1} x_k P_q^k = F_q\left(\sum_{k=0}^{q-1} x_k D_q^k\right) F_q^H \\ &= F_q\,\mathsf{diag}(\sqrt{q}\,xF_q)\,F_q^H\end{aligned}$$

where the first equality follows from (3.12) for the group $\mathbb{Z}/q\mathbb{Z}$ [129, Section 0.9.6], $P_q = F_q D_q F_q^H \in \mathbb{R}^{q\times q}$ is defined in (3.15), and:

$$D_q = \mathsf{diag}\left(\left(1, \exp\left(\frac{2\pi i}{q}\right), \exp\left(\frac{4\pi i}{q}\right), \ldots, \exp\left(\frac{2(q-1)\pi i}{q}\right)\right)\right).$$

Hence, we have:

$$\begin{aligned}\lambda_j(W_\delta) &= \sum_{k=1}^{q} (w_\delta)_k \exp\left(\frac{2\pi(j-1)(k-1)i}{q}\right) \\ &= \begin{cases} 1 & , \quad j = 1 \\ 1 - \delta - \frac{\delta}{q-1} & , \quad j \in \{2, \ldots, q\} \end{cases}\end{aligned}$$

where $w_\delta = (1-\delta, \delta/(q-1), \ldots, \delta/(q-1))$.

Part 3: This is also obvious from (3.19)—recall that a square stochastic matrix is doubly stochastic if and only if its stationary distribution is uniform [129, Section 8.7].

Part 4: For $\delta \neq \frac{q-1}{q}$, we can verify that $W_\tau W_\delta = I_q$ when $\tau = -\delta/(1-\delta-\frac{\delta}{q-1})$ by direct computation:

$$\begin{aligned}[W_\tau W_\delta]_{j,j} &= (1-\tau)(1-\delta) + (q-1)\left(\frac{\tau}{q-1}\right)\left(\frac{\delta}{q-1}\right) \\ &= 1\,, \quad \text{for } \; j \in \{1, \ldots, q\}\,, \\ [W_\tau W_\delta]_{j,k} &= \frac{\delta(1-\tau)}{q-1} + \frac{\tau(1-\delta)}{q-1} + (q-2)\frac{\tau\delta}{(q-1)^2} \\ &= 0\,, \quad \text{for } \; j \neq k \; \text{ and } \; j,k \in \{1, \ldots, q\}\,.\end{aligned}$$

The $\delta = \frac{q-1}{q}$ case follows from (3.19).

Part 5: The set $\{W_\delta : \delta \in \mathbb{R} \text{ and } \delta \neq \frac{q-1}{q}\}$ is closed under matrix multiplication. Indeed, for any $\epsilon, \delta \in \mathbb{R}\backslash\{\frac{q-1}{q}\}$, we can straightforwardly verify that $W_\epsilon W_\delta = W_\tau$ with $\tau = \epsilon + \delta - \epsilon\delta - \frac{\epsilon\delta}{q-1}$. Moreover, $\tau \neq \frac{q-1}{q}$ because W_τ is invertible (since W_ϵ and W_δ are invertible using part 4). The set also includes the identity matrix as $W_0 = I_q$, and multiplicative inverses (using part 4). Finally, the associativity of matrix multiplication and the commutativity of circulant matrices proves that $\{W_\delta : \delta \in \mathbb{R} \text{ and } \delta \neq \frac{q-1}{q}\}$ is an Abelian group. ■

■ B.4 Proof of Theorem 3.2

Let $X_1, X_2, Y_1, Y_2, Z_1, Z_2$ be discrete random variables with finite alphabets, and let $P_{Y_i|X_i}$ and $P_{Z_i|X_i}$ for $i = 1, 2$ be discrete channels. We use the notation $P_{Y_1|X_1} \otimes P_{Y_2|X_2}$ to represent the tensor product channel from X_1^2 to Y_1^2. (In particular, the stochastic matrix of $P_{Y_1|X_1} \otimes P_{Y_2|X_2}$ is the Kronecker product of the stochastic matrices of $P_{Y_1|X_1}$ and $P_{Y_2|X_2}$.) The next lemma presents the tensorization property of the less noisy preorder [231, Proposition 16], [268, Proposition 5].

Lemma B.3 (Tensorization of Less Noisy [231, 268]). *If $P_{Z_i|X_i} \succeq_{\mathsf{ln}} P_{Y_i|X_i}$ for $i = 1, 2$, then $P_{Z_1|X_1} \otimes P_{Z_2|X_2} \succeq_{\mathsf{ln}} P_{Y_1|X_1} \otimes P_{Y_2|X_2}$.*

We now prove Theorem 3.2 using Theorem 3.1, Lemma B.3, and the relation (3.22).

Proof of Theorem 3.2. We follow the proof strategy in [231]. First, observe that $\eta_j = \eta_{\mathsf{KL}}(P_{Y_j|X_j})$ for all $j \in \{1, \ldots, n\}$ due to Proposition 2.6 in chapter 2 (cf. [46, Theorem 1]). Hence, using (3.22) (cf. [231, Proposition 15]), we have that the erasure channel $P_{Z_j|X_j} = E_{1-\eta_j}$ with erasure probability $1 - \eta_j$ is less noisy than $P_{Y_j|X_j}$ for every $j \in \{1, \ldots, n\}$. Next, define the memoryless channel:

$$P_{Z_1^n|X_1^n} = \prod_{j=1}^{n} P_{Z_j|X_j}$$

which is the tensor product of the erasure channels $P_{Z_j|X_j}$ over all $j \in \{1, \ldots, n\}$. Then, Lemma B.3 yields that $P_{Z_1^n|X_1^n}$ is less noisy than $P_{Y_1^n|X_1^n}$.

To prove the f-divergence version of Samorodnitsky's SDPI, fix any pair of input distributions $P_{X_1^n}$ and $Q_{X_1^n}$. Then, using Theorem 3.1, we have:

$$D_f(P_{Y_1^n}||Q_{Y_1^n}) \le D_f(P_{Z_1^n}||Q_{Z_1^n}) \tag{B.6}$$

where $P_{Z_1^n}$ and $Q_{Z_1^n}$ are the output distributions after passing $P_{X_1^n}$ and $Q_{X_1^n}$ through the channel $P_{Z_1^n|X_1^n}$, respectively. Now notice that the output Z_1^n of the product erasure channel can be equivalently represented as (X_S, S), where the random subset S represents the indices that are not erased and X_S represents the values at these indices. (Note that $Z_1^n = (\mathsf{e}, \ldots, \mathsf{e})$ corresponds to $S = \varnothing$.) Hence, we can write:

$$\begin{aligned} D_f(P_{Z_1^n}||Q_{Z_1^n}) &= \sum_{T \subseteq \{1,\ldots,n\}} P_S(T) \sum_{x_T} Q_{X_T}(x_T) f\left(\frac{P_S(T)P_{X_T}(x_T)}{P_S(T)Q_{X_T}(x_T)}\right) \\ &= \sum_{T \subseteq \{1,\ldots,n\}} P_S(T)\, D_f(P_{X_T}||Q_{X_T}) \end{aligned}$$

where we use the fact that S is independent of X_1^n and we employ the conventions: $D_f(P_{X_\varnothing}||Q_{X_\varnothing}) = 0$, and $P_S(T)D_f(P_{X_T}||Q_{X_T}) = 0$ if $P_S(T) = 0$. Together with (B.6), this establishes the f-divergence version of Samorodnitsky's SDPI.

To prove the mutual f-information version of Samorodnitsky's SDPI, fix any joint distribution P_{U,X_1^n} and consider the Markov chain $U \to X_1^n \to (Y_1^n, Z_1^n)$ such that Y_1^n and Z_1^n are conditionally independent given X_1^n. This defines a joint distribution P_{U,X_1^n,Y_1^n,Z_1^n} using the channel conditional distributions $P_{Y_1^n|X_1^n}$ and $P_{Z_1^n|X_1^n}$. Given any $U = u$, it follows from (B.6) that:

$$D_f(P_{Y_1^n|U=u}||P_{Y_1^n}) \le D_f(P_{Z_1^n|U=u}||P_{Z_1^n})$$

where $P_{Y_1^n|U=u}$ and $P_{Z_1^n|U=u}$ are the two output distributions corresponding to the input distribution $P_{X_1^n|U=u}$ (due to the Markov relation). Using (2.19) from chapter 2, taking

expectations with respect to P_U yields:

$$I_f(U;Y_1^n) \le I_f(U;Z_1^n)\,. \tag{B.7}$$

Furthermore, using the equivalence between Z_1^n and (X_S, S), we have (as before) that:

$$\begin{aligned} I_f(U;Z_1^n) &= I_f(U;X_S,S) \\ &= D_f(P_{X_S|U,S}P_UP_S||P_{X_S|S}P_UP_S) \\ &= \sum_{T\subseteq\{1,\ldots,n\}} P_S(T) \sum_u P_U(u) \sum_{x_T} P_{X_T}(x_T)\, f\left(\frac{P_{X_T|U}(x_T|u)P_U(u)P_S(T)}{P_{X_T}(x_T)P_U(u)P_S(T)}\right) \\ &= \sum_{T\subseteq\{1,\ldots,n\}} P_S(T)\, D_f(P_{U,X_T}||P_UP_{X_T}) \\ &= \sum_{T\subseteq\{1,\ldots,n\}} P_S(T)\, I_f(U;X_T) \end{aligned}$$

where we use (2.18) from chapter 2 and the fact that S is independent of (U, X_1^n), and we employ the conventions: $I_f(U;X_\varnothing) = 0$, and $P_S(T)I_f(U;X_T) = 0$ if $P_S(T) = 0$. Together with (B.7), this establishes the mutual f-information version of Samorodnitsky's SDPI.

Lastly, the case $\eta_j = \eta$ for all $j \in \{1,\ldots,n\}$ follows from substituting the expression for $P_S(T)$ into the mutual f-information version of Samorodnitsky's SDPI, and then performing some straightforward manipulations. This completes the proof. ■

We remark that under the assumptions of Theorem 3.2, if we additionally have $\eta_j = \eta$ for all $j \in \{1,\ldots,n\}$, then our generalized Samorodnitsky's SDPI is indeed tighter than the tensorized SDPI in (3.27):

$$I_f(U;Y_1^n) \le \sum_{T\subseteq\{1,\ldots,n\}} \eta^{|T|}(1-\eta)^{n-|T|} I_f(U;X_T) \le (1-(1-\eta)^n)\, I_f(U;X_1^n)\,. \tag{B.8}$$

This is because $I_f(U;Z_1^n) \le (1-(1-\eta)^n)\, I_f(U;X_1^n)$ using (3.27) for the product erasure channel $P_{Z_1^n|X_1^n}$.

■ B.5 Alternative Proof of Lemma 3.1

Our alternative proof of Lemma 3.1 requires the following lemma.

Lemma B.4 (Characterization of $\mathcal{X}$-Circulant Matrices). *A matrix $A \in \mathbb{R}^{q\times q}$ is $\mathcal{X}$-circulant if and only if it equals its P_x-conjugate for each $x \in \mathcal{X}$:*

$$\forall x \in \mathcal{X},\ \ A = P_xAP_x^T$$

where the matrices $\{P_x \in \mathbb{R}^{q\times q} : x \in \mathcal{X}\}$ are defined in (3.9).

Proof. The forward direction is obvious because $\mathcal{X}$-circulant matrices commute. To prove the converse direction, recall that $\mathcal{X}$-circulant matrices are jointly diagonalized by a unitary "Fourier" matrix of characters $F \in \mathcal{V}_q(\mathbb{C}^q)$. Moreover, every (real) matrix of the form FDF^H for some diagonal matrix $D \in \mathbb{C}^{q\times q}$ is $\mathcal{X}$-circulant (using (3.12) and the fact that the diagonal matrices $\{F^H P_x F \in \mathbb{C}^{q\times q} : x \in \mathcal{X}\}$ form a basis for all complex diagonal matrices). Now observe that $A = P_x A P_x^T$ for every $x \in \mathcal{X}$ implies that A commutes with every (complex) $\mathcal{X}$-circulant matrix (using (3.12)). This means that $F^H AF$ commutes with every diagonal matrix, and is therefore itself diagonal. Hence, A is an $\mathcal{X}$-circulant matrix. This completes the proof. ■

A corollary of Lemma B.4 is that A^{-1} is $\mathcal{X}$-circulant if A is a non-singular $\mathcal{X}$-circulant matrix. We next prove Lemma 3.1.

Proof. As explained in the earlier proof of Lemma 3.1, it is sufficient to prove that $W \succeq_{\mathsf{deg}} V$ implies $V = W\mathsf{circ}_{\mathcal{X}}(z)$ for some $z \in \mathcal{P}_q$. By Definition 3.1, if $W \succeq_{\mathsf{deg}} V$, then we have for some $R \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$:

$$\begin{aligned} V &= WR \\ \forall x \in \mathcal{X},\ \ P_x V P_x^T &= P_x W P_x^T P_x R P_x^T \\ \forall x \in \mathcal{X},\ \ V &= W P_x R P_x^T \end{aligned}$$

where the third equality holds due to Lemma B.4 because V and W are $\mathcal{X}$-circulant matrices. Hence, if R degrades W to V, then so does its P_x-conjugate for every $x \in \mathcal{X}$. Averaging over all $x \in \mathcal{X}$ gives:

$$V = W\left(\frac{1}{q}\sum_{x\in\mathcal{X}} P_x R P_x^T\right).$$

The stochastic matrix $R' = \frac{1}{q}\sum_{x\in\mathcal{X}} P_x R P_x^T \in \mathbb{R}_{\mathsf{sto}}^{q\times q}$ equals its P_x-conjugate for all $x \in \mathcal{X}$, since $\{P_x \in \mathbb{R}^{q\times q} : x \in \mathcal{X}\}$ is an Abelian group (that is isomorphic to $(\mathcal{X}, \oplus)$). Therefore, R' is $\mathcal{X}$-circulant by Lemma B.4. This completes the proof. ■

■ B.6 Proof of Lemma 3.2

Proof. We provide a proof based on that of [129, Theorem 7.7.3(a)], which handles the invertible A case. If A is the zero matrix, then the lemma trivially holds. So, we assume that A is non-zero. Furthermore, notice that if $A \succeq_{\mathsf{PSD}} B$, then $\mathcal{K}(A) \subseteq \mathcal{K}(B)$. Indeed, for any $x \in \mathcal{K}(A)$, $0 = x^T Ax \geq x^T Bx \geq 0$ since $A \succeq_{\mathsf{PSD}} B \succeq_{\mathsf{PSD}} 0$, which implies that $x \in \mathcal{K}(B)$. So, we will also assume that $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ (which is equivalent to $\mathcal{K}(A) \subseteq \mathcal{K}(B)$ by taking orthogonal complements), and prove that:

$$A \succeq_{\mathsf{PSD}} B \quad \Leftrightarrow \quad \rho\Big(A^{\dagger}B\Big) \leq 1\,. \tag{B.9}$$

To this end, we first establish that:

$$A \succeq_{\mathsf{PSD}} B \quad \Leftrightarrow \quad P \succeq_{\mathsf{PSD}} \left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger}. \tag{B.10}$$

This is a consequence of the following argument:

$$\begin{aligned} A \succeq_{\mathsf{PSD}} B \quad &\Rightarrow \quad \left(A^{\frac{1}{2}}\right)^{\dagger} A \left(A^{\frac{1}{2}}\right)^{\dagger} \succeq_{\mathsf{PSD}} \left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger} \\ &\Leftrightarrow \quad P \succeq_{\mathsf{PSD}} \left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger} \\ &\Rightarrow \quad PAP^T \succeq_{\mathsf{PSD}} PBP^T \\ &\Rightarrow \quad A \succeq_{\mathsf{PSD}} B \end{aligned}$$

where $P \triangleq A^{\frac{1}{2}}(A^{\frac{1}{2}})^{\dagger} = AA^{\dagger} = P^T$ is the orthogonal projection matrix onto $\mathcal{R}(A^{\frac{1}{2}}) = \mathcal{R}(A)$ [106, Section 5.5.4], the second equivalence follows from the facts that $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$ and $P^2 = P$ (idempotency), the third implication easily follows from the first implication, and the final implication holds because $PAP^T = A$ and $PBP^T = B$ (since $\mathcal{R}(B) \subseteq \mathcal{R}(A)$).

Next, observe that:

$$\begin{aligned} P \succeq_{\mathsf{PSD}} \left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger} \quad &\Leftrightarrow \quad \rho\left(\left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger}\right) \leq \rho(P) = 1 \\ &\Leftrightarrow \quad \rho\left(A^{\dagger}BP^T\right) \leq 1 \\ &\Leftrightarrow \quad \rho\left(A^{\dagger}B\right) \leq 1 \end{aligned} \tag{B.11}$$

where the spectral radii of P and $(A^{\frac{1}{2}})^{\dagger}B(A^{\frac{1}{2}})^{\dagger}$ equal their largest eigenvalues, respectively, because all their eigenvalues are non-negative. The forward direction of the first equivalence follows from the Courant-Fischer-Weyl min-max theorem in [129, Theorem 4.2.6], and the converse direction holds because $(A^{\frac{1}{2}})^{\dagger}B(A^{\frac{1}{2}})^{\dagger} \in \mathbb{R}^{q\times q}_{\succeq 0}$ and $\rho((A^{\frac{1}{2}})^{\dagger}B(A^{\frac{1}{2}})^{\dagger}) \leq 1$ imply that:

$$\begin{aligned} I_q \succeq_{\mathsf{PSD}} \left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger} \quad &\Rightarrow \quad PP^T \succeq_{\mathsf{PSD}} P\left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger} P^T \\ &\Rightarrow \quad P \succeq_{\mathsf{PSD}} \left(A^{\frac{1}{2}}\right)^{\dagger} B \left(A^{\frac{1}{2}}\right)^{\dagger}. \end{aligned}$$

The second equivalence holds because $(A^{\frac{1}{2}})^{\dagger}B(A^{\frac{1}{2}})^{\dagger}$ and $A^{\dagger}BP^T$ share the same eigenvalues. Indeed, both matrices have nullspaces containing $\mathcal{K}(A)$, and every other eigenvector $x \in \mathcal{R}(A)$ of $(A^{\frac{1}{2}})^{\dagger}B(A^{\frac{1}{2}})^{\dagger}$ has a unique corresponding eigenvector $(A^{\frac{1}{2}})^{\dagger}x$ of $A^{\dagger}BP^T$ with the same eigenvalue. The final equivalence holds because $A^{\dagger}BP^T = A^{\dagger}B$ (since $\mathcal{R}(B) \subseteq \mathcal{R}(A)$).

Therefore, combining (B.10) and (B.11) yields (B.9), which completes the proof. ■

■ B.7 Alternative Proof of Part 1 (Circle Condition) of Proposition 3.9

Proof. We provide an alternative Fourier analytic proof of the circle condition in part 1 of Proposition 3.9. Since all $\mathcal{X}$-circulant matrices are jointly diagonalized by a unitary "Fourier" matrix of characters $F \in \mathcal{V}_q(\mathbb{C}^q)$, we have:

$$W = F\,\mathsf{diag}(\lambda_w)\,F^H$$
$$V = F\,\mathsf{diag}(\lambda_v)\,F^H$$

where $\lambda_w \in \mathbb{C}^q$ and $\lambda_v \in \mathbb{C}^q$ are the eigenvalues of W and V, respectively. This gives us:

$$WW^T = F\,\mathsf{diag}\Big(|\lambda_w|^2\Big)\,F^H$$
$$VV^T = F\,\mathsf{diag}\Big(|\lambda_v|^2\Big)\,F^H$$

where $|\lambda_w|^2 \in \mathbb{R}^q$ and $|\lambda_v|^2 \in \mathbb{R}^q$ denote vectors that are the entry-wise squared magnitudes of λ_w and λ_v, respectively. Since $W \succeq_{\mathsf{ln}} V$, letting $P_X = \mathbf{u}$ (which means that $P_X W = P_X V = \mathbf{u}$) in part 2 of Proposition 3.8 gives:

$$WW^T \succeq_{\mathsf{PSD}} VV^T$$
$$\mathsf{diag}\Big(|\lambda_w|^2\Big) \succeq_{\mathsf{PSD}} \mathsf{diag}\Big(|\lambda_v|^2\Big)$$
$$\|\lambda_w\|_2^2 \geq \|\lambda_v\|_2^2$$
$$\|w\|_2^2 \geq \|v\|_2^2$$
$$\|w - \mathbf{u}\|_2^2 \geq \|v - \mathbf{u}\|_2^2$$

where the second statement is a non-singular $*$-congruence using F, the fourth inequality follows from an analog of the standard *Parseval-Plancherel theorem*: $q\,\|w\|_2^2 = \|W\|_{\mathsf{Fro}}^2 = \|\mathsf{diag}(\lambda_w)\|_{\mathsf{Fro}}^2 = \|\lambda_w\|_2^2$ and $q\,\|v\|_2^2 = \|\lambda_v\|_2^2$ (due to (3.14) and the unitary invariance of the Frobenius norm $\|\cdot\|_{\mathsf{Fro}}$), and the final inequality holds because $\|w\|_2^2 = \|w - \mathbf{u}\|_2^2 + \|\mathbf{u}\|_2^2$ and $\|v\|_2^2 = \|v - \mathbf{u}\|_2^2 + \|\mathbf{u}\|_2^2$ as $w - \mathbf{u}$ and $v - \mathbf{u}$ are orthogonal to $\mathbf{u}$. This completes the proof. ■

We remark that the circle condition is not sufficient to deduce whether $W \succeq_{\mathsf{ln}} V$. For example, if $(\mathcal{X}, \oplus)$ is a cyclic Abelian group and $F = F_q$ is the DFT matrix, then there exist pmfs $w, v \in \mathcal{P}_q$ such that $\|w\|_2 \geq \|v\|_2$ but the corresponding Fourier coefficients (or eigenvalues of W and V) do not satisfy $|\lambda_w| \geq |\lambda_v|$ entry-wise.

■ B.8 Proof of Proposition 3.12

Proof.

Part 1: We first recall from [69, Appendix, Theorem A.1] that the Markov chain

$\mathbf{1u} \in \mathbb{R}^{q\times q}_{\mathsf{sto}}$ with uniform stationary distribution $\pi = \mathbf{u} \in \mathcal{P}_q$ has LSI constant:

$$\alpha(\mathbf{1u}) = \inf_{\substack{f\in\mathcal{L}^2(\mathcal{X},\mathbf{u}):\\ \|f\|_{\mathbf{u}}=1\\ D(f^2\mathbf{u}||\mathbf{u})\neq 0}} \frac{\mathcal{E}_{\mathsf{std}}(f,f)}{D(f^2\mathbf{u}||\mathbf{u})} = \begin{cases} \frac{1}{2} & , \quad q=2 \\ \frac{1-\frac{2}{q}}{\log(q-1)} & , \quad q>2 \end{cases}.$$

Now using (3.35), $\alpha(W_\delta) = \frac{q\delta}{q-1}\alpha(\mathbf{1u})$, which proves part 1.

Part 2: Observe that $W_\delta W_\delta^* = W_\delta W_\delta^T = W_\delta^2 = W_{\delta'}$, where the first equality holds because W_δ has uniform stationary pmf, and $\delta' = \delta(2 - \frac{q\delta}{q-1})$ using the proof of part 5 of Proposition 3.4. As a result, the discrete LSI constant $\alpha(W_\delta W_\delta^*) = \alpha(W_{\delta'})$, which we can calculate using part 1 of this proposition.

Part 3: It is well-known in the literature that $\rho_{\mathsf{max}}(X;Y)$ equals the second largest singular value of the DTM $W_\delta = \mathsf{diag}(\sqrt{\mathbf{u}})\, W_\delta\, \mathsf{diag}(\sqrt{\mathbf{u}})^{-1}$ (see e.g. Proposition 2.2 in chapter 2). Hence, from part 2 of Proposition 3.4, we have $\rho_{\mathsf{max}}(X;Y) = \left|1 - \delta - \frac{\delta}{q-1}\right|$.

Part 4: First recall that the Dobrushin contraction coefficient (for TV distance) of W_δ can be computed using the two-point characterization (2.49) in chapter 2:

$$\eta_{\mathsf{TV}}(W_\delta) = \frac{1}{2}\max_{x,x'\in[q]} \left\| w_\delta P_q^x - w_\delta P_q^{x'} \right\|_1 = \left|1 - \delta - \frac{\delta}{q-1}\right|$$

where w_δ is the noise pmf of W_δ for $\delta \in [0,1]$ (see (3.18)), $P_q \in \mathbb{R}^{q\times q}$ is defined in (3.15), and we use (2.4) to represent TV distance. Moreover, using part 7 of Proposition 2.3, Definition 2.5, and part 7 of Proposition 2.5, we have:

$$\rho_{\mathsf{max}}(X;Y)^2 \leq \eta_{\mathsf{KL}}(W_\delta) \leq \eta_{\mathsf{TV}}(W_\delta) \tag{B.12}$$

where the joint distribution of (X,Y) is given in part 3 of this proposition. Hence, the value of $\eta_{\mathsf{TV}}(W_\delta)$ and part 3 establish part 4. This completes the proof. ■

■ B.9 Auxiliary Result: Domination Factor Function

Proposition B.4 (Properties of Domination Factor Function). *Given a channel $V \in \mathbb{R}^{q\times r}_{\mathsf{sto}}$ that is strictly positive entry-wise, its domination factor function $\mu_V : (0, \frac{q-1}{q}) \to \mathbb{R}$ (defined in (3.65)) is continuous, convex, and strictly increasing. Moreover, we have $\lim_{\delta\to\frac{q-1}{q}} \mu_V(\delta) = +\infty$.*

Proof. We first prove that μ_V is finite on $(0, \frac{q-1}{q})$. For any $P_X, Q_X \in \mathcal{P}_q$ and any $\delta \in (0, \frac{q-1}{q})$, we have:

$$D(P_XV||Q_XV) \leq \chi^2(P_XV||Q_XV) \leq \frac{\|(P_X - Q_X)V\|_2^2}{\nu} \leq \frac{\|P_X - Q_X\|_2^2\, \|V\|_{\mathsf{op}}^2}{\nu}$$

where the first inequality follows from Lemma 2.3 in chapter 2, and $\nu = \min\{[V]_{i,j} : i \in \{1,\ldots,q\},\, j \in \{1,\ldots,r\}\}$. Furthermore, for any $P_X, Q_X \in \mathcal{P}_q$ and any $\delta \in (0, \frac{q-1}{q})$,

we also have:

$$D(P_X W_\delta || Q_X W_\delta) \geq \frac{1}{2} \left\| (P_X - Q_X) W_\delta \right\|_2^2 \geq \frac{1}{2} \left\| P_X - Q_X \right\|_2^2 \left(1 - \delta - \frac{\delta}{q-1} \right)^2$$

where the first inequality follows from e.g. (A.4) in appendix A.5, and the second inequality follows from part 2 of Proposition 3.4. Hence, we get:

$$\forall \delta \in \left(0, \frac{q-1}{q} \right), \;\; \mu_V(\delta) \leq \frac{2 \left\| V \right\|_{\mathsf{op}}^2}{\nu \left(1 - \delta - \frac{\delta}{q-1} \right)^2} . \tag{B.13}$$

To prove that μ_V is strictly increasing, observe that $W_{\delta'} \succeq_{\mathsf{deg}} W_\delta$ for $0 < \delta' < \delta < \frac{q-1}{q}$, because $W_\delta = W_{\delta'} W_p$ with:

$$\begin{aligned} p &= \delta - \frac{\delta'}{1 - \delta' - \frac{\delta'}{q-1}} + \frac{\delta \delta'}{1 - \delta' - \frac{\delta'}{q-1}} + \frac{\frac{\delta \delta'}{q-1}}{1 - \delta' - \frac{\delta'}{q-1}} \\ &= \frac{\delta - \delta'}{1 - \delta' - \frac{\delta'}{q-1}} \in \left(0, \frac{q-1}{q} \right) \end{aligned}$$

where we use part 4 of Proposition 3.4, the proof of part 5 of Proposition 3.4 in appendix B.3, and the fact that $W_p = W_{\delta'}^{-1} W_\delta$. As a result, we have for every $P_X, Q_X \in \mathcal{P}_q$:

$$D(P_X W_\delta || Q_X W_\delta) \leq \eta_{\mathsf{KL}}(W_p) \, D(P_X W_{\delta'} || Q_X W_{\delta'})$$

using the SDPI for KL divergence, where part 4 of Proposition 3.12 reveals that $\eta_{\mathsf{KL}}(W_p) \in (0,1)$ since $p \in (0, \frac{q-1}{q})$. Hence, we have for $0 < \delta' < \delta < \frac{q-1}{q}$:

$$\mu_V(\delta') \leq \eta_{\mathsf{KL}}(W_p) \, \mu_V(\delta) \tag{B.14}$$

using (3.65), and the fact that $0 < D(P_X W_{\delta'} || Q_X W_{\delta'}) < +\infty$ if and only if $0 < D(P_X W_\delta || Q_X W_\delta) < +\infty$. This implies that μ_V is strictly increasing.

We next establish that μ_V is convex and continuous. For any fixed $P_X, Q_X \in \mathcal{P}_q$ such that $P_X \neq Q_X$, consider the function $\delta \mapsto D(P_X V || Q_X V) \, / D(P_X W_\delta || Q_X W_\delta)$ with domain $(0, \frac{q-1}{q})$. This function is convex, because $\delta \mapsto D(P_X W_\delta || Q_X W_\delta)$ is convex by the convexity of KL divergence, and the reciprocal of a non-negative convex function is convex. Therefore, μ_V is convex since (3.65) defines it as a pointwise supremum of a collection of convex functions. Furthermore, we note that μ_V is also continuous since a convex function is continuous on the interior of its domain.

Finally, observe that:

$$\liminf_{\delta \to \frac{q-1}{q}} \mu_V(\delta) \geq \sup_{\substack{P_X, Q_X \in \mathcal{P}_q \\ P_X \neq Q_X}} \liminf_{\delta \to \frac{q-1}{q}} \frac{D(P_X V || Q_X V)}{D(P_X W_\delta || Q_X W_\delta)}$$

$$= \sup_{\substack{P_X, Q_X \in \mathcal{P}_q \\ P_X \neq Q_X}} \frac{D(P_X V || Q_X V)}{\limsup_{\delta \to \frac{q-1}{q}} D(P_X W_\delta || Q_X W_\delta)}$$

$$= +\infty$$

where the first inequality follows from the minimax inequality and (3.65) (note that $0 < D(P_X W_\delta || Q_X W_\delta) < +\infty$ for $P_X \neq Q_X$ and δ close to $\frac{q-1}{q}$), and the final equality holds because $P_X W_{(q-1)/q} = \mathbf{u}$ for every $P_X \in \mathcal{P}_q$. This completes the proof. ■

■ B.10 Auxiliary Result: Löwner Domination Lemma

Lemma B.5 (Gramian Löwner Domination implies Symmetric Part Löwner Domination). *Given $A \in \mathbb{R}^{q \times q}_{\succeq 0}$ and $B \in \mathbb{R}^{q \times q}$ that is normal, we have:*

$$A^2 = AA^T \succeq_{\mathsf{PSD}} BB^T \quad \Rightarrow \quad A = \frac{A + A^T}{2} \succeq_{\mathsf{PSD}} \frac{B + B^T}{2}.$$

Proof. Since $AA^T \succeq_{\mathsf{PSD}} BB^T \succeq_{\mathsf{PSD}} 0$, using part 1 of Theorem B.1 (the Löwner-Heinz theorem) in appendix B.2 with $p = \frac{1}{2}$, we get (cf. (3.84)):

$$A = \left(AA^T\right)^{\frac{1}{2}} \succeq_{\mathsf{PSD}} \left(BB^T\right)^{\frac{1}{2}} \succeq_{\mathsf{PSD}} 0$$

where the first equality holds because $A \in \mathbb{R}^{q \times q}_{\succeq 0}$. It suffices to now prove that:

$$\left(BB^T\right)^{\frac{1}{2}} \succeq_{\mathsf{PSD}} \frac{B + B^T}{2}$$

as the transitive property of $\succeq_{\mathsf{PSD}}$ will produce $A \succeq_{\mathsf{PSD}} (B + B^T)/2$. Since B is normal, $B = UDU^H$ by the complex spectral theorem [17, Theorem 7.9], where $U \in \mathcal{V}_q(\mathbb{C}^q)$ is a unitary matrix and $D \in \mathbb{C}^{q \times q}$ is a complex diagonal matrix. Using this unitary diagonalization, we have:

$$U|D|U^H = \left(BB^T\right)^{\frac{1}{2}} \succeq_{\mathsf{PSD}} \frac{B + B^T}{2} = U \operatorname{Re}\{D\} U^H$$

since $|D| \succeq_{\mathsf{PSD}} \operatorname{Re}\{D\}$, where $|D|$ and $\operatorname{Re}\{D\}$ denote the element-wise magnitude and real part of D, respectively. This completes the proof. ■

Appendix C

Supplementary Results and Proofs from Chapter 4

C.1 Basics of Matrix Perturbation Theory

Matrix perturbation theory studies how perturbations of a matrix affect different decompositions such as its spectral decomposition or its SVD. We will survey some basic singular value stability inequalities in this appendix; much of our discussion is based on the reference texts [128], [270], and [264], and a survey of perturbation theory for the SVD in [263]. To present these inequalities, we first introduce some relevant background.

Fix any $m, n \in \mathbb{N}$ and consider the vector space of all real $m \times n$ matrices $\mathbb{R}^{m\times n}$. A rather useful, but not very widely known, unitarily invariant norm on this space is the (p,k)-*norm*. For any matrix $A \in \mathbb{R}^{m\times n}$, let $\sigma_i(A)$ denote the ith largest singular value of A for $i \in \{1,\dots,\min\{m,n\}\}$:

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A) \geq 0\,, \tag{C.1}$$

$\{\psi_1^A,\dots,\psi_n^A\} \subseteq \mathbb{R}^n$ denote the orthonormal basis of right (or input) singular vectors of A, and $\{\phi_1^A,\dots,\phi_m^A\} \subseteq \mathbb{R}^m$ denote orthonormal basis of left (or output) singular vectors of A, where for every $i \in \{1,\dots,\min\{m,n\}\}$:

$$A\psi_i^A = \sigma_i(A)\phi_i^A \quad \text{and} \quad A^T\phi_i^A = \sigma_i(A)\psi_i^A\,. \tag{C.2}$$

Then, the (p,k)-norm of A for $p \in [1,\infty]$ and $k \in \{1,\dots,\min\{m,n\}\}$ is defined as:

$$\|A\|_{(p,k)} \triangleq \left(\sum_{i=1}^{k} \sigma_i(A)^p\right)^{\frac{1}{p}} \text{ for } p \in [1,\infty), \quad \text{and} \quad \|A\|_{(\infty,k)} \triangleq \sigma_1(A) \tag{C.3}$$

where the norm does not depend on k when $p = \infty$, cf. [171], [27, Equation (IV.47), p.95].[105] The (p,k)-norms generalize various other well-known matrix norms. In particular, the (∞,k)-norms are all the *operator norm* (or *spectral norm* or *induced* ℓ^2*-norm*):

$$\forall A \in \mathbb{R}^{m\times n}, \;\; \|A\|_{\mathsf{op}} \triangleq \|A\|_{(\infty,k)} = \sigma_1(A)\,, \tag{C.4}$$

[105] The only difficulty in proving that this is a valid norm is in checking the triangle inequality. This follows from Proposition C.3 and the (reverse) Minkowski inequality.

the $(2, \min\{m,n\})$-norm is the *Frobenius norm* (or *Hilbert-Schmidt norm*):

$$\forall A \in \mathbb{R}^{m\times n},\ \ \|A\|_{\mathsf{Fro}} \triangleq \|A\|_{(2,\min\{m,n\})} = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}[A]_{i,j}^2}\,, \tag{C.5}$$

the $(1, \min\{m,n\})$-norm is the *nuclear norm* (or *trace norm*), the $(p, \min\{m,n\})$-norms are the *Schatten ℓ^p-norms*, and the $(1,k)$-norms are the *Ky Fan k-norms*.

As further background, we next present the *Courant-Fischer-Weyl min-max theorem* for singular values, cf. [128, Theorem 3.1.2], [129, Theorems 4.2.6 and 7.3.8], [270, Theorem 1.3.2 and Exercise 1.3.21], since spectral stability inequalities are often derived from such variational characterizations of spectra (eigenvalues or singular values).

Theorem C.1 (Courant-Fischer-Weyl Min-max Theorem [128,129,270]). *For any matrix $A \in \mathbb{R}^{m\times n}$, the kth largest singular value of A for any $k \in \{1,\ldots,\min\{m,n\}\}$ is given by:*

$$\begin{aligned}\sigma_k(A) &= \min_{\substack{V\subseteq\mathbb{R}^n:\\ \dim(V)=n-k+1}} \max_{\substack{x\in V:\\ \|x\|_2=1}} \|Ax\|_2 \\ &= \max_{\substack{V\subseteq\mathbb{R}^n:\\ \dim(V)=k}} \min_{\substack{x\in V:\\ \|x\|_2=1}} \|Ax\|_2\end{aligned}$$

where V denotes a linear subspace of $\mathbb{R}^n$, and $\dim(V)$ denotes the dimension of V.

We note that for the largest singular value, Theorem C.1 simplifies to the well-known expression:

$$\sigma_1(A) = \max_{\substack{x\in\mathbb{R}^n:\\ \|x\|_2=1}} \|Ax\|_2\,. \tag{C.6}$$

Using the variational characterization in Theorem C.1, we can prove (perhaps) the most basic singular value stability result known as the *Weyl inequality* [129, Corollary 7.3.5(a)], [263, Theorem 1]. This inequality portrays that perturbations of the singular values of a matrix are bounded by the operator norm of the perturbation matrix.

Proposition C.1 (Weyl Inequality [129,263]). *For any two matrices $A, B \in \mathbb{R}^{m\times n}$, we have:*

$$\max_{k\in\{1,\ldots,\min\{m,n\}\}} |\sigma_k(A) - \sigma_k(B)| \le \|A-B\|_{\mathsf{op}}\,.$$

Proof. We provide a proof for completeness. Consider a matrix $B \in \mathbb{R}^{m\times n}$ which is (additively) perturbed by a matrix $A - B \in \mathbb{R}^{m\times n}$ to produce the matrix $A \in \mathbb{R}^{m\times n}$. Observe that for any $k \in \{1,\ldots,\min\{m,n\}\}$ and any $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, we have:

$$\begin{aligned} Ax &= Bx + (A-B)x \\ \|Ax\|_2 &\le \|Bx\|_2 + \|(A-B)x\|_2 \\ \sigma_k(A) &\le \sigma_k(B) + \sigma_1(A-B)\end{aligned}$$

$$\sigma_k(A) - \sigma_k(B) \leq \|A - B\|_{\mathsf{op}} \tag{C.7}$$

where the second inequality follows from the triangle inequality (with equality if and only if Bx or $(A-B)x$ is a non-negative scalar multiple of the other), and the third inequality follows from first taking the maximum of $\|(A-B)x\|_2$ over all $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$ and then using Theorem C.1. (Note that (C.7) holds with equality if $\psi_1^{A-B} = \psi_k^A = \psi_k^B$ and $\phi_1^{A-B} = \phi_k^B$.) Likewise, we can show that:

$$\sigma_k(B) - \sigma_k(A) \leq \|B - A\|_{\mathsf{op}} = \|A - B\|_{\mathsf{op}} \tag{C.8}$$

because B is a perturbation of A by $B - A$. Combining (C.7) and (C.8) completes the proof. ■

Proposition C.1 is indeed a stability result because it demonstrates that the map $\sigma_k : \mathbb{R}^{m\times n} \to [0,\infty)$ is Lipschitz continuous for every $k \in \{1,\dots,\min\{m,n\}\}$ [270]. Much like the proof of Proposition C.1, a more general variational characterization of singular values analogous to the *Wielandt minimax formula* in [270, Exercise 1.3.3] can be used to derive a more powerful singular value stability result known as the *Lidskii inequality*, cf. [270, Exercise 1.3.22(ii)]. We reproduce a variant of the Lidskii inequality from [128, Theorem 3.4.5] below.

Proposition C.2 (Lidskii Inequality [128], [270]). *For any two matrices $A, B \in \mathbb{R}^{m\times n}$, we have:*

$$\sum_{j=1}^{k} \left|\sigma_{i_j}(A) - \sigma_{i_j}(B)\right| \leq \|A - B\|_{(1,k)}$$

for every $k \in \{1,\dots,\min\{m,n\}\}$ and every set of indices $1 \leq i_1 < i_2 < \cdots < i_k \leq \min\{m,n\}$. Equivalently, using the definition of weak majorization in (B.1) of appendix B.1, we have that the sequence of singular values of the perturbation matrix $A - B$, $(\sigma_1(A-B),\dots,\sigma_{\min\{m,n\}}(A-B))$, weakly majorizes the sequence of absolute singular value differences $(|\sigma_1(A) - \sigma_1(B)|,\dots,|\sigma_{\min\{m,n\}}(A) - \sigma_{\min\{m,n\}}(B)|)$.

It is easy to see that Proposition C.2 also conveys that the truncated sequence of singular values $(\sigma_1(A-B),\dots,\sigma_k(A-B))$ weakly majorizes the truncated sequence of absolute singular value differences $(|\sigma_1(A) - \sigma_1(B)|,\dots,|\sigma_k(A) - \sigma_k(B)|)$ for any $k \in \{1,\dots,\min\{m,n\}\}$. Hence, applying Karamata's inequality in Proposition B.3 of appendix B.1 to this weak majorization result yields:

$$\sum_{i=1}^{k} f(|\sigma_i(A) - \sigma_i(B)|) \leq \sum_{i=1}^{k} f(\sigma_i(A-B)) \tag{C.9}$$

for every convex non-decreasing function $f : \mathbb{R} \to \mathbb{R}$ and every $k \in \{1,\dots,\min\{m,n\}\}$. This begets the following general stability inequality for singular values which we dub the *generalized Wielandt-Hoffman inequality*.

Proposition C.3 (Generalized Wielandt-Hoffman Inequality). *For any two matrices $A, B \in \mathbb{R}^{m\times n}$, we have:*

$$\left(\sum_{i=1}^{k} |\sigma_i(A) - \sigma_i(B)|^p\right)^{\frac{1}{p}} \leq \|A - B\|_{(p,k)}$$

for every $p \in [1, \infty]$ and $k \in \{1, \ldots, \min\{m, n\}\}$.

Proof. The author of [270] delineates how the $k = \min\{m, n\}$ case of this result can be proved using Hölder's inequality and its extremal equality case. We present a shorter alternative proof for general $k \in \{1, \ldots, \min\{m, n\}\}$ based on the majorization observations stemming from [128]. Since the $p = \infty$ case corresponds to the Weyl inequality in Proposition C.1 (regardless of k), it suffices to prove the generalized Wielandt-Hoffman inequality for $p \in [1, \infty)$. To this end, fix any $p \in [1, \infty)$ and any $k \in \{1, \ldots, \min\{m, n\}\}$. Then, using (C.9), we have:

$$\sum_{i=1}^{k} |\sigma_i(A) - \sigma_i(B)|^p \leq \sum_{i=1}^{k} \sigma_i(A - B)^p$$

where we define $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = x^p$. This implies that:

$$\left(\sum_{i=1}^{k} |\sigma_i(A) - \sigma_i(B)|^p\right)^{\frac{1}{p}} \leq \|A - B\|_{(p,k)}$$

which completes the proof. ∎

Proposition C.3 generalizes several singular value stability inequalities such as the Weyl inequality in Proposition C.1 where $p = \infty$ (and the value of k does not matter), the *Mirsky inequality* (which is also known as the *Wielandt-Hoffman inequality*) [263, 264]:

$$\left(\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(B))^2\right)^{\frac{1}{2}} \leq \|A - B\|_{\mathsf{Fro}} \tag{C.10}$$

where $p = 2$ and $k = \min\{m, n\}$, and the *p-Wielandt-Hoffman inequality* [270]:

$$\left(\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(B))^p\right)^{\frac{1}{p}} \leq \|A - B\|_{(p,\min\{m,n\})} \tag{C.11}$$

where $p \in [1, \infty]$ is general and $k = \min\{m, n\}$.

We next derive two stability inequalities for norms using the aforementioned results, which we will exploit in chapter 4. The first of these lemmata upper bounds the Ky Fan k-norm difference between two matrices via the Frobenius norm of their difference.

Lemma C.1 (Ky Fan k-Norm Stability). *For any two matrices $A, B \in \mathbb{R}^{m\times n}$ and every $k \in \{1,\dots,\min\{m,n\}\}$, we have:*

$$\left|\|A\|_{(1,k)} - \|B\|_{(1,k)}\right| \leq \sqrt{k}\,\|A-B\|_{\mathsf{Fro}}\,.$$

Proof. Observe that:

$$\begin{aligned}
\left|\|A\|_{(1,k)} - \|B\|_{(1,k)}\right| &\leq \sum_{i=1}^{k} |\sigma_i(A) - \sigma_i(B)| \\
&\leq \sum_{i=1}^{k} \sigma_i(A-B) \\
&\leq \sqrt{k}\sqrt{\sum_{i=1}^{k} \sigma_i(A-B)^2} \\
&\leq \sqrt{k}\sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i(A-B)^2} \\
&\leq \sqrt{k}\,\|A-B\|_{\mathsf{Fro}}
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from the Lidskii inequality in Proposition C.2, the third inequality follows from the Cauchy-Schwarz inequality,[106] and the final inequality holds by definition of the Frobenius norm. ■

We remark that the scaling factor of $\sqrt{k}$ in Lemma C.1 is essential. Indeed, although the Mirsky inequality in (C.10) upper bounds $\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(B))^2$ using $\|A-B\|_{\mathsf{Fro}}^2$, we cannot upper bound $\sum_{i=1}^{\min\{m,n\}} |\sigma_i(A) - \sigma_i(B)|$ using $\|A-B\|_{\mathsf{Fro}}$ because ℓ^1-norms are greater than or equal to ℓ^2-norms in general.

Our second lemma establishes a certain stability result for the squared $(2,k)$-norm of a matrix.

Lemma C.2 (Squared $(2,k)$-Norm Stability). *For any two matrices $A, B \in \mathbb{R}^{m\times n}$ and every $k \in \{1,\dots,\min\{m,n\}\}$, we have:*

$$\left|\|A\|_{(2,k)}^2 - \left\|A\Psi_{(k)}^B\right\|_{\mathsf{Fro}}^2\right| = \left|\left\|A\Psi_{(k)}^A\right\|_{\mathsf{Fro}}^2 - \left\|A\Psi_{(k)}^B\right\|_{\mathsf{Fro}}^2\right| \leq 4k\,\|A\|_{\mathsf{op}}\,\|A-B\|_{\mathsf{op}}$$

where $\Psi_{(k)}^A \triangleq [\psi_1^A \cdots \psi_k^A] \in \mathcal{V}_k(\mathbb{R}^n)$ denotes the orthonormal k-frame that collects the first k right singular vectors of A.

[106] This is a useful trick when proving probabilistic bounds for matrix norms, cf. [45, Equation (2)].

Proof. The first equality holds because:

$$\left\|A\Psi_{(k)}^{A}\right\|_{\mathsf{Fro}}^{2}=\sum_{i=1}^{k}\sigma_{i}(A)^{2}=\|A\|_{(2,k)}^{2}\ .$$

To prove the second inequality, observe that:

$$\begin{aligned}
\left|\left\|A\Psi_{(k)}^{A}\right\|_{\mathsf{Fro}}^{2}-\left\|A\Psi_{(k)}^{B}\right\|_{\mathsf{Fro}}^{2}\right| &= \left|\sum_{i=1}^{k}\left\|A\psi_{i}^{A}\right\|_{2}^{2}-\left\|A\psi_{i}^{B}\right\|_{2}^{2}\right| \\
&\leq \sum_{i=1}^{k}\left|\left\|A\psi_{i}^{A}\right\|_{2}^{2}-\left\|A\psi_{i}^{B}\right\|_{2}^{2}\right| \\
&= \sum_{i=1}^{k}\left|\left\|A\psi_{i}^{A}\right\|_{2}-\left\|A\psi_{i}^{B}\right\|_{2}\right|\left(\left\|A\psi_{i}^{A}\right\|_{2}+\left\|A\psi_{i}^{B}\right\|_{2}\right) \\
&\leq 2\left\|A\right\|_{\mathsf{op}}\sum_{i=1}^{k}\left|\left\|A\psi_{i}^{A}\right\|_{2}-\left\|A\psi_{i}^{B}\right\|_{2}\right| \qquad \text{(C.12)} \\
&\leq 2\left\|A\right\|_{\mathsf{op}}\sum_{i=1}^{k}\left|\left\|A\psi_{i}^{A}\right\|_{2}-\left\|B\psi_{i}^{B}\right\|_{2}\right|+\left|\left\|B\psi_{i}^{B}\right\|_{2}-\left\|A\psi_{i}^{B}\right\|_{2}\right| \\
&\leq 2\left\|A\right\|_{\mathsf{op}}\sum_{i=1}^{k}\left|\sigma_{i}(A)-\sigma_{i}(B)\right|+\left\|(A-B)\psi_{i}^{B}\right\|_{2} \\
&\leq 2\left\|A\right\|_{\mathsf{op}}\sum_{i=1}^{k}\left\|A-B\right\|_{\mathsf{op}}+\left\|A-B\right\|_{\mathsf{op}} \\
&= 4k\left\|A\right\|_{\mathsf{op}}\left\|A-B\right\|_{\mathsf{op}}
\end{aligned}$$

where the second inequality uses the triangle inequality, the fourth inequality holds because $\|Ax\|_2 \leq \|A\|_{\mathsf{op}}$ for $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$, the fifth inequality uses the triangle inequality, the sixth inequality follows from the reverse triangle inequality and the relations $\|A\psi_i^A\|_2 = \sigma_i(A)$ and $\|B\psi_i^B\|_2 = \sigma_i(B)$, and the seventh inequality follows from the Weyl inequality in Proposition C.1 and the fact that $\|(A-B)x\|_2 \leq \|A-B\|_{\mathsf{op}}$ for $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$. This completes the proof. ■

This concludes our discussion on spectral stability inequalities.

■ C.2 Elements of Large Deviations Theory and Concentration of Measure

In this appendix, we introduce some basic results from *large deviations theory* and the theory of exponential *concentration of measure* inequalities. We begin by presenting the celebrated *Sanov's theorem* from large deviations theory, cf. [65, Theorem 2.1.10 and Exercises 2.1.16, 2.1.18, and 2.1.19] and [59, Theorem 2.1]. Fix a finite alphabet $\mathcal{X}$ such that $2 \leq |\mathcal{X}| < \infty$ and a pmf $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ on $\mathcal{X}$, and consider a sequence of discrete

random variables $\{X_k : k \in \mathbb{N}\}$ that are drawn i.i.d. from P_X. For every $n \in \mathbb{N}$, let $\hat{P}_{X_1^n} \in \mathcal{P}_\mathcal{X}$ denote the (random) *empirical distribution* of the observations X_1^n:

$$\forall x \in \mathcal{X}, \;\; \hat{P}_{X_1^n}(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = x\}\,, \tag{C.13}$$

and $\widehat{\mathbb{E}}_n[\cdot]$ denote the *empirical expectation operator* of the observations X_1^n:

$$\widehat{\mathbb{E}}_n[f(X)] \triangleq \mathbb{E}_{\hat{P}_{X_1^n}}[f(X)] = \sum_{x \in \mathcal{X}} \hat{P}_{X_1^n}(x) f(x) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tag{C.14}$$

for every function $f : \mathcal{X} \to \mathbb{R}$. Sanov's theorem illustrates that the empirical distributions $\{\hat{P}_{X_1^n} : n \in \mathbb{N}\}$ satisfy a *large deviations principle* with *speed* n and *rate function* given by KL divergence $D(\cdot||P_X)$.

Theorem C.2 (Sanov's Theorem [59,65]). *For every Borel subset of pmfs $S \subseteq \mathcal{P}_\mathcal{X}$, we have:*

$$\begin{aligned} -\inf_{Q_X \in S^\circ} D(Q_X||P_X) &\leq \liminf_{n\to\infty} \frac{1}{n} \log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n} \in S\Big)\Big) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n} \in S\Big)\Big) \leq -\inf_{Q_X \in S} D(Q_X||P_X) \end{aligned}$$

where S° denotes the interior of S, and $\mathcal{P}_\mathcal{X}$ inherits its topology from $\big(\mathbb{R}^{|\mathcal{X}|}\big)^*$.[107] *Furthermore, for every subset of pmfs $S \subseteq \mathcal{P}_\mathcal{X}$ that is contained in the closure of its interior,*[108] *we have equality in the above inequalities:*

$$\lim_{n\to\infty} \frac{1}{n} \log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n} \in S\Big)\Big) = -\inf_{Q_X \in S} D(Q_X||P_X)\,.$$

Finally, for every closed and convex subset of pmfs $S \subseteq \mathcal{P}_\mathcal{X}$ with non-empty interior, we have:

$$\lim_{n\to\infty} \frac{1}{n} \log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n} \in S\Big)\Big) = -\min_{Q_X \in S} D(Q_X||P_X) = -D(Q_X^*||P_X)$$

where the minimum is achieved by a unique pmf $Q_X^* \in S$.

Under the setup of Theorem C.2, in the next lemma, we locally approximate the *Chernoff exponent* of the probability that the empirical mean of a function $t : \mathcal{X} \to \mathbb{R}$ deviates from its theoretical mean $\mathbb{E}[t(X)] = \mathbb{E}_{P_X}[t(X)]$. (This approximation can be intuitively understood using the CLT.)

[107] We can construe $\mathcal{P}_\mathcal{X}$ as a metric space because it inherits its metric, e.g. any ℓ^p-norm with $p \in [1, \infty]$, from $\big(\mathbb{R}^{|\mathcal{X}|}\big)^*$, and openness and closedness are defined with respect to this metric on $\mathcal{P}_\mathcal{X}$.

[108] This includes the case where S is an open set.

Lemma C.3 (Local Approximation of Chernoff Exponent). *For any function $t : \mathcal{X} \to \mathbb{R}$ with non-zero mean, $\mathbb{E}[t(X)] \neq 0$, and strictly positive variance, $\mathbb{VAR}(t(X)) > 0$, we have:*

$$-\lim_{\gamma\to 0^+}\lim_{n\to\infty}\frac{1}{\gamma^2 n}\log\left(\mathbb{P}\left(\left|\frac{\widehat{\mathbb{E}}_n[t(X)]}{\mathbb{E}[t(X)]}-1\right|\geq\gamma\right)\right)=\frac{\mathbb{E}[t(X)]^2}{2\,\mathbb{VAR}(t(X))}$$

where the outer limit is one-sided, i.e. $\gamma > 0$, and $\widehat{\mathbb{E}}_n[\cdot]$ is the empirical expectation operator defined in (C.14).

Proof. Since this result does not change when the function $t : \mathcal{X} \to \mathbb{R}$ is negated, we may assume without loss of generality that $\mathbb{E}[t(X)] > 0$. For any $\gamma > 0$, define the disjoint sets:

$$S_\gamma \triangleq \{Q_X \in \mathcal{P}_\mathcal{X} : \mathbb{E}_{Q_X}[t(X)] \geq (1+\gamma)\mathbb{E}[t(X)]\}$$
$$T_\gamma \triangleq \{Q_X \in \mathcal{P}_\mathcal{X} : \mathbb{E}_{Q_X}[t(X)] \leq (1-\gamma)\mathbb{E}[t(X)]\}$$

where $\mathbb{E}_{Q_X}[t(X)] = \sum_{x\in\mathcal{X}} Q_X(x)t(x)$. Furthermore, since we will eventually let $\gamma \to 0$, we can assume that:

$$0 < \gamma < \min\left\{\frac{\max_{x\in\mathcal{X}} t(x)}{\mathbb{E}[t(X)]} - 1, 1 - \frac{\min_{x\in\mathcal{X}} t(x)}{\mathbb{E}[t(X)]}\right\}.$$

This ensures that:

$$\min_{x\in\mathcal{X}} t(x) < (1-\gamma)\mathbb{E}[t(X)] < (1+\gamma)\mathbb{E}[t(X)] < \max_{x\in\mathcal{X}} t(x)$$

where $\min_{x\in\mathcal{X}} t(x) < \max_{x\in\mathcal{X}} t(x)$ because $\mathbb{VAR}(t(X)) > 0$. Hence, S_γ and T_γ are closed and convex sets that have non-empty interior. Using Theorem C.2, we have:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n}\in S_\gamma\Big)\Big) &= -\min_{Q_X\in S_\gamma} D(Q_X||P_X)\\ &= -D(Q_X^1||P_X) \qquad \text{(C.15)}\\ \lim_{n\to\infty}\frac{1}{n}\log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n}\in T_\gamma\Big)\Big) &= -\min_{Q_X\in T_\gamma} D(Q_X||P_X)\\ &= -D(Q_X^2||P_X) \qquad \text{(C.16)}\end{aligned}$$

where the unique minimizing distributions, $Q_X^1 \in S_\gamma$ and $Q_X^2 \in T_\gamma$, respectively, are members of the *exponential family*, cf. [59, Section 2.1], [290, Lecture 14]:

$$\forall x \in \mathcal{X},\;\; Q_X(x;s) = P_X(x)\exp(st(x) - \alpha(s))$$

with *natural parameter* $s \in \mathbb{R}$, *natural (sufficient) statistic* $t : \mathcal{X} \to \mathbb{R}$, *base distribution* P_X, and *log-partition function* (or *cumulant generating function*):

$$\forall s \in \mathbb{R},\;\; \alpha(s) \triangleq \log(\mathbb{E}[\exp(st(X))])\,.$$

The log-partition function is infinitely differentiable, and has first and second derivatives, cf. [290, Lecture 9]:

$$\begin{aligned}\forall s \in \mathbb{R},\ \ \alpha'(s) &= \mathbb{E}_{Q_X(\cdot;s)}[t(X)] \\ \forall s \in \mathbb{R},\ \ \alpha''(s) &= \mathbb{VAR}_{Q_X(\cdot;s)}(t(X)) > 0\end{aligned}$$

which follow from straightforward calculations. Here, the second derivative (or variance) is strictly positive because every element of $\mathcal{X}$ has positive probability mass under $Q_X(\cdot;s)$. The minimizing distributions are $Q_X^1 = Q_X(\cdot;s_1)$ and $Q_X^2 = Q_X(\cdot;s_2)$, where the optimal parameters $s_1 > 0$ and $s_2 < 0$ are chosen to satisfy:

$$\begin{aligned}\alpha'(s_1) &= \mathbb{E}_{Q_X(\cdot;s_1)}[t(X)] = (1+\gamma)\mathbb{E}[t(X)]\,, \\ \alpha'(s_2) &= \mathbb{E}_{Q_X(\cdot;s_2)}[t(X)] = (1-\gamma)\mathbb{E}[t(X)]\,,\end{aligned}$$

respectively (see [59, Example 2.1] or [290, Lecture 14]).

We next prove that:

$$\lim_{\gamma\to 0^+} \frac{D(Q_X^1||P_X)}{\gamma^2} = \lim_{\gamma\to 0^+} \frac{D(Q_X^2||P_X)}{\gamma^2} = \frac{\mathbb{E}[t(X)]^2}{2\mathbb{VAR}(t(X))}\,. \tag{C.17}$$

Consider the function $d : \mathbb{R} \to [0,\infty)$, $d(s) \triangleq D(Q_X(\cdot;s)||P_X)$. It is straightforward to show that:

$$\begin{aligned}\forall s \in \mathbb{R},\ \ d(s) &= s\alpha'(s) - \alpha(s) \\ \forall s \in \mathbb{R},\ \ d'(s) &= s\alpha''(s) \\ \forall s \in \mathbb{R},\ \ d''(s) &= \alpha''(s) + s\alpha'''(s)\end{aligned}$$

which means that $d(0) = d'(0) = 0$, and $d''(0) = \alpha''(0) = \mathbb{VAR}(t(X))$. Hence, by Taylor's theorem:

$$\lim_{s\to 0} \frac{d(s)}{s^2} = \frac{\mathbb{VAR}(t(X))}{2}\,. \tag{C.18}$$

Now set $\alpha'(s) = \mathbb{E}_{Q_X(\cdot;s)}[t(X)] = (1+\tau)\mathbb{E}[t(X)]$ for any $\tau \in \mathbb{R}$. This implies that $s = {\alpha'}^{-1}((1+\tau)\mathbb{E}[t(X)])$, where ${\alpha'}^{-1}$ exists because α' is strictly increasing (since α'' is strictly positive). Next, observe that:

$$\begin{aligned}\lim_{\tau\to 0} \frac{d(s)}{\tau^2} &= \lim_{\tau\to 0} \frac{d(s)}{s^2} \lim_{\tau\to 0} \frac{s^2}{\tau^2} \\ &= \frac{\mathbb{VAR}(t(X))}{2} \left(\lim_{\tau\to 0} \frac{s}{\tau}\right)^2 \\ &= \frac{\mathbb{VAR}(t(X))}{2} \left(\left.\frac{ds}{d\tau}\right|_{\tau=0}\right)^2 \\ &= \frac{\mathbb{VAR}(t(X))}{2} \left(\left.\frac{\mathbb{E}[t(X)]}{\alpha''(s)}\right|_{\tau=0}\right)^2\end{aligned}$$

$$= \frac{\mathbb{E}[t(X)]^2}{2\mathbb{VAR}(t(X))} \tag{C.19}$$

where the second equality follows from (C.18), the fact that $s \to 0$ when $\tau \to 0$ (by the continuity of α'^{-1}), and the continuity of $x \mapsto x^2$, the third equality follows from the definition of derivative and the fact that $\tau = 0$ corresponds to $s = 0$ (as $\alpha'(0) = \mathbb{E}[t(X)]$), the fourth equality holds because $s = \alpha'^{-1}((1+\tau)\mathbb{E}[t(X)])$, and the fifth equality holds because $\tau = 0$ corresponds to $s = 0$ and $\alpha''(0) = \mathbb{VAR}(t(X))$. Lastly, note that setting $\tau = \gamma > 0$ and $s = s_1 > 0$ gives:

$$\lim_{\gamma \to 0^+} \frac{D(Q_X^1 || P_X)}{\gamma^2} = \lim_{\tau \to 0^+} \frac{d(s)}{\tau^2},$$

and setting $\tau = -\gamma < 0$ and $s = s_2 < 0$ gives:

$$\lim_{\gamma \to 0^+} \frac{D(Q_X^2 || P_X)}{\gamma^2} = \lim_{\tau \to 0^-} \frac{d(s)}{\tau^2},$$

which proves (C.17) via (C.19).

Finally, define the set $\Pi_\gamma \triangleq S_\gamma \cup T_\gamma$:

$$\Pi_\gamma = \left\{ Q_X \in \mathcal{P}_\mathcal{X} : \left| \frac{\mathbb{E}_{Q_X}[t(X)]}{\mathbb{E}[t(X)]} - 1 \right| \geq \gamma \right\}$$

and consider $\mathbb{P}(\hat{P}_{X_1^n} \in \Pi_\gamma) = \mathbb{P}(\hat{P}_{X_1^n} \in S_\gamma) + \mathbb{P}(\hat{P}_{X_1^n} \in T_\gamma)$ (since S_γ and T_γ are disjoint). The *Laplace principle* yields, cf. [65, Lemma 1.2.15]:

$$-\lim_{n \to \infty} \frac{1}{n} \log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n} \in \Pi_\gamma\Big)\Big) = \min\Big\{ D(Q_X^1 || P_X), D(Q_X^2 || P_X) \Big\}$$

using (C.15) and (C.16). Using (C.17), this implies that:

$$-\lim_{\gamma \to 0^+} \lim_{n \to \infty} \frac{1}{\gamma^2 n} \log\Big(\mathbb{P}\Big(\hat{P}_{X_1^n} \in \Pi_\gamma\Big)\Big) = \frac{\mathbb{E}[t(X)]^2}{2\mathbb{VAR}(t(X))},$$

which completes the proof. ■

We now switch our focus to presenting some useful exponential concentration of measure inequalities. The most basic such inequality is perhaps *Hoeffding's inequality*, which guarantees concentration using the boundedness of the underlying random variables, cf. [126, Theorems 1 and 2].

Lemma C.4 (Hoeffding's Inequality [126]). *Suppose $X_1, \ldots, X_n$ are independent bounded random variables such that $0 \leq X_i \leq 1$ for all $i \in \{1, \ldots, n\}$. Then, we have:*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_i] \geq \gamma \right) \leq \exp\Big(-2n\gamma^2\Big)$$

and:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \leq -\gamma\right) \leq \exp\left(-2n\gamma^2\right)$$

for every $\gamma \geq 0$.

The next lemma portrays a tighter variant of Lemma C.4 for i.i.d. Bernoulli random variables known as the *Chernoff-Hoeffding bound*, cf. [126, Theorem 1].

Lemma C.5 (Chernoff-Hoeffding Bound [126]). *Suppose* $X_1, \ldots, X_n$ *are i.i.d.* Bernoulli(p) *random variables with* $p \in (0,1)$. *Then, for every (small enough)* $\gamma > 0$, *we have:*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \gamma\right) \leq \exp(-nD(p+\gamma||p))$$

and:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \leq -\gamma\right) \leq \exp(-nD(p-\gamma||p))$$

where $D(\alpha||\beta) \triangleq \alpha\log(\alpha/\beta) + (1-\alpha)\log((1-\alpha)/(1-\beta))$ *for* $\alpha, \beta \in (0,1)$ *denotes the binary KL divergence function.*

While Hoeffding's inequality only uses knowledge of the boundedness of the underlying random variables, some situations demand a finer understanding of the exponents in such tail bounds. To address this need, *Bennett's inequality* and *Bernstein's inequality* provide concentration of measure bounds that incorporate information about the variances of the underlying random variables. In chapter 4, we will require certain generalizations of the standard Bernstein's inequality. So, we present a vector generalization of Bernstein's inequality below, which we reproduce from [38, Theorem 2.4] with slight re-parametrization for convenience.

Lemma C.6 (Vector Bernstein Inequality [38, Theorem 2.4]). *Let* $V_1, \ldots, V_n \in \mathbb{R}^d$ *be independent random vectors such that for some constant* $C > 0$, $\|V_i - \mathbb{E}[V_i]\|_2 \leq C$ *a.s. for all* $i \in \{1, \ldots, n\}$. *Let* $\nu > 0$ *be another constant such that:*

$$\nu \geq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|V_i - \mathbb{E}[V_i]\|_2^2\right].$$

Then, for all $0 \leq t \leq \frac{\nu}{C}$*:*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} V_i - \mathbb{E}[V_i]\right\|_2 \geq t\right) \leq \exp\left(\frac{1}{4} - \frac{nt^2}{8\nu}\right).$$

As noted in [38], this bound does not depend on the dimension $d \in \mathbb{N}$. Finally, we conclude this appendix by presenting a $d_1 \times d_2$ matrix version of Bernstein's inequality (with $d_1, d_2 \in \mathbb{N}$), which we reproduce from [277, Theorem 1.6] with slight re-parametrization for convenience.

Lemma C.7 (Matrix Bernstein Inequality [277, Theorem 1.6]). *Let $Z_1, \ldots, Z_n \in \mathbb{R}^{d_1 \times d_2}$ be independent random matrices such that for some constant $C > 0$, we have $\|Z_i - \mathbb{E}[Z_i]\|_{\mathsf{op}} \leq C$ a.s. for all $i \in \{1, \ldots, n\}$. Let $\nu > 0$ be another constant such that:*

$$\nu \geq \max\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{COV}(Z_i) \right\|_{\mathsf{op}}, \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{COV}\left(Z_i^T\right) \right\|_{\mathsf{op}} \right\}$$

where $\mathbb{COV}(Z) \triangleq \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T\right]$ for any random matrix Z. Then, for all $0 \leq t \leq \frac{\nu}{C}$:

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i] \right\|_{\mathsf{op}} \geq t \right) \leq (d_1 + d_2) \exp\left(-\frac{3nt^2}{8\nu} \right).$$

■ C.3 Representation of Conditional Expectation Operators

Fix any bivariate distribution $P_{X,Y} \in \mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$. Although the SVD of the corresponding DTM $B \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ is clearly significant from a mutual χ^2-information perspective (see subsection 4.2.3), it is still reasonable to wonder why we study this SVD rather than the SVDs of other commonly used representations of $P_{X,Y}$ such as $P_{X,Y}$ itself, or the matrix $B' \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ defined entry-wise as:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \quad [B']_{y,x} \triangleq \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}, \tag{C.20}$$

whose logarithm is the pointwise mutual information or *information density* [47, 117]. We do not address this question in its full generality here. However, we illustrate in this appendix that the DTM B is the only contraction matrix in the class of matrices $\{\mathsf{diag}(\sqrt{P_Y}) V \mathsf{diag}(\sqrt{Q_X})^{-1} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|} : Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}\}$, where $V \in \mathcal{P}_{\mathcal{X}|\mathcal{Y}}$ denotes the row stochastic matrix corresponding to the conditional distribution $P_{X|Y}$.

For convenience, we will present the aforementioned result in the language of conditional expectation operators. (The equivalence between the two versions of the result can be argued using relations similar to (4.6), (4.7), (4.8), and (4.9).) Recall that the conditional expectation operator C maps any function $f : \mathcal{X} \to \mathbb{R}$ to the function $C(f) : \mathcal{Y} \to \mathbb{R}$, cf. (4.5):

$$\forall y \in \mathcal{Y}, \quad (C(f))(y) = \mathbb{E}[f(X)|Y = y] \tag{C.21}$$

where the conditional distribution $P_{X|Y}$ completely characterizes C. In order to make C a well-defined linear operator with an SVD, we must endow its input and output vector spaces of functions with inner products. Let us fix the output Hilbert space of C to be $\mathcal{L}^2(\mathcal{Y}, P_Y)$, where $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. While this produces a "canonical" choice of input Hilbert space, namely $\mathcal{L}^2(\mathcal{X}, P_X)$, where $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ is the marginal pmf of $P_{X,Y}$, let us instead

choose an arbitrary input Hilbert space $\mathcal{L}^2(\mathcal{X}, Q_X)$ for some $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$. We define the corresponding induced operator norm of $C : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ as:

$$\|C\|_{Q_X \to P_Y} \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X) \setminus \{\mathbf{0}\}} \frac{\|C(f)\|_{P_Y}}{\|f\|_{Q_X}} \tag{C.22}$$

where we use $\mathbf{0}$ to represent the everywhere zero function. The next proposition conveys that the only choice of input Hilbert space that makes C a contraction is the canonical choice $\mathcal{L}^2(\mathcal{X}, P_X)$.

Proposition C.4 (Hilbert Spaces of Conditional Expectation Operators). *The minimum operator norm of C over all choices of input Hilbert spaces in $\{\mathcal{L}^2(\mathcal{X}, Q_X) : Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}\}$ is:*

$$\min_{Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}} \|C\|_{Q_X \to P_Y} = \|C\|_{P_X \to P_Y} = 1$$

where the unique minimizer is $Q_X^* = P_X$. Furthermore, for any $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, the gap between $\|C\|_{Q_X \to P_Y}^2$ and the minimum squared operator norm is lower bounded by:

$$\|C\|_{Q_X \to P_Y}^2 - \|C\|_{P_X \to P_Y}^2 = \|C\|_{Q_X \to P_Y}^2 - 1 \geq \chi^2(P_X || Q_X)\,.$$

Proof. Note that for every pmf $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, we have $\mathbf{1} \in \mathcal{L}^2(\mathcal{X}, Q_X)$ with $\|\mathbf{1}\|_{Q_X} = 1$. Similarly, $C(\mathbf{1}) = \mathbf{1} \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ with $\|C(\mathbf{1})\|_{P_Y} = \|\mathbf{1}\|_{P_Y} = 1$. As a result, we get:

$$\forall Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ},\ \ \|C\|_{Q_X \to P_Y} \geq 1\,.$$

However, we know that $Q_X = P_X$ achieves this lower bound, because for every $f \in \mathcal{L}^2(\mathcal{X}, P_X)$:

$$\|C(f)\|_{P_Y}^2 = \mathbb{E}\Big[\mathbb{E}[f(X)|Y]^2\Big] \leq \mathbb{E}\Big[\mathbb{E}\Big[f(X)^2\Big|Y\Big]\Big] = \mathbb{E}\Big[f(X)^2\Big] = \|f\|_{P_X}^2$$

using conditional Jensen's inequality and the tower property. This proves that:

$$\min_{Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}} \|C\|_{Q_X \to P_Y} = \|C\|_{P_X \to P_Y} = 1$$

where $Q_X^* = P_X$ is a valid minimizer.

To prove that $Q_X^* = P_X$ is the unique minimizer, it suffices to establish that:

$$\forall Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ},\ \ \|C\|_{Q_X \to P_Y}^2 \geq 1 + \chi^2(P_X || Q_X)\,.$$

Fix any pmf $Q_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, and consider the adjoint operator $C^* : \mathcal{L}^2(\mathcal{Y}, P_Y) \to \mathcal{L}^2(\mathcal{X}, Q_X)$ of the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$, which is defined by the relation:

$$\begin{aligned}\langle C(f), g\rangle_{P_Y} &= \mathbb{E}_{P_Y}[\mathbb{E}[f(X)|Y]\, g(Y)] \\ &= \mathbb{E}_{P_{X,Y}}[f(X)g(Y)]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{P_X}[f(X)\,\mathbb{E}[g(Y)|X]] \\
&= \mathbb{E}_{Q_X}\left[f(X)\,\mathbb{E}[g(Y)|X]\,\frac{P_X(X)}{Q_X(X)}\right] \\
&= \langle f, C^*(g)\rangle_{Q_X}
\end{aligned}
$$

for all $f \in \mathcal{L}^2(\mathcal{X}, Q_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$, where the conditional expectation $\mathbb{E}[g(Y)|X]$ is taken with respect to the conditional distribution $P_{Y|X}$. In particular, for any function $g \in \mathcal{L}^2(\mathcal{Y}, P_Y)$, the function $C^*(g) \in \mathcal{L}^2(\mathcal{X}, Q_X)$ is given by:

$$
\forall x \in \mathcal{X},\ \ (C^*(g))(x) = \frac{P_X(x)}{Q_X(x)}\,\mathbb{E}[g(Y)|X = x]\,.
$$

Now observe that for $\mathbf{1} \in \mathcal{L}^2(\mathcal{Y}, P_Y)$ with $\|\mathbf{1}\|_{P_Y} = 1$, we have:

$$
\forall x \in \mathcal{X},\ \ (C^*(\mathbf{1}))(x) = \frac{P_X(x)}{Q_X(x)}\,.
$$

This implies that:

$$
\begin{aligned}
\|C\|^2_{Q_X \to P_Y} &= \|C^*\|^2_{P_Y \to Q_X} \\
&\geq \|C^*(\mathbf{1})\|^2_{Q_X} \\
&= \sum_{x \in \mathcal{X}} Q_X(x)\frac{P_X(x)^2}{Q_X(x)^2} \\
&= 1 + \chi^2(P_X||Q_X)
\end{aligned}
$$

where the first equality follows from the definition of the adjoint operator, and the last equality follows from (2.9) in chapter 2. This completes the proof. ■

Proposition C.4 portrays that given the joint pmf $P_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, P_X and P_Y are the only choice of inner products that make the conditional expectation operators $C = \mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, P_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ and $C^* = \mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, P_Y) \to \mathcal{L}^2(\mathcal{X}, P_X)$ (defined by the conditional distributions $P_{X|Y}$ and $P_{Y|X}$, respectively) adjoints and contraction operators. On the other hand, as mentioned earlier, if we are only given the conditional distribution $P_{X|Y} \in \mathcal{P}_{\mathcal{X}|\mathcal{Y}}$ that defines C, we are free to select $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, but we must choose the corresponding marginal pmf $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ for the other inner product to ensure that C is a contraction. Furthermore, the contraction property of C is attractive because it implies that the DPI for χ^2-divergence is satisfied in the sense of (4.17) (or the DPI for KL divergence is satisfied locally).

We remark that the restriction in Proposition C.4 to Hilbert spaces with inner products defined by probability distributions is natural. In general, every inner product on $\mathbb{R}^n$ (with $n \in \mathbb{N}$) can be represented by a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$:

$$
\forall x_1, x_2 \in \mathbb{R}^n,\ \ \langle x_1, x_2\rangle_A = x_1^T A x_2\,. \tag{C.23}
$$

This symmetric positive definite matrix A can be orthogonally diagonalized by the spectral theorem [129, Section 2.5]. For simplicity, we can drop the orthogonal matrices in this diagonalization and only consider diagonal matrices $A \in \mathbb{R}^{n\times n}$ with strictly positive diagonal entries, which correspond to weighted inner products. Furthermore, we restrict the diagonal entries of A to sum to unity to obtain a "well-defined" problem, since allowing arbitrary scaling would make the minimum in Proposition C.4 zero. This yields the class of inner products considered in Proposition C.4.

■ C.4 Proof of Lemma 4.1

Proof.

Part 1: Fix any $0 < \tau < b_{\mathsf{min}}$, and consider the set:

$$\mathcal{M} \triangleq \left\{ M \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|} : M \geq 0 \text{ entry-wise, } \|M\|_{\mathsf{op}} = 1, \text{ and } \|M - B\|_{\mathsf{op}} \leq \tau \right\}.$$

We first show that $\mathcal{M}$ is closed. To this end, take any sequence $\{M_n \in \mathcal{M} : n \in \mathbb{N}\}$ such that $M_n \to M \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ as $n \to \infty$. Then, clearly $M \geq 0$ entry-wise and $\|M\|_{\mathsf{op}} = 1$ (by continuity of the operator norm). Moreover, we have:

$$\begin{aligned} \|M - B\|_{\mathsf{op}} &\leq \|M - M_n\|_{\mathsf{op}} + \|M_n - B\|_{\mathsf{op}} \\ &\leq \lim_{n\to\infty} \|M - M_n\|_{\mathsf{op}} + \tau \\ &\leq \tau \end{aligned}$$

where the first inequality is the triangle inequality, the second inequality follows from using the fact that $M_n \in \mathcal{M}$ and then letting $n \to \infty$, and the final inequality holds because $M_n \to M$ as $n \to \infty$. Hence, $\mathcal{M}$ is closed.

We next establish that $\mathcal{M} \subseteq \mathcal{B}^\circ$ (see part 1 of Theorem 4.2). Notice that for every $M \in \mathcal{M}$ and every $i \in \mathcal{Y}$, $j \in \mathcal{X}$, the (i,j)th element of $M - B$ satisfies:

$$\begin{aligned} |[M]_{i,j} - [B]_{i,j}| &= \left| e_i^T (M - B) e_j \right| \\ &\leq \|e_i\|_2 \, \|M - B\|_{\mathsf{op}} \, \|e_j\|_2 \\ &= \|M - B\|_{\mathsf{op}} \end{aligned} \tag{C.24}$$

using the Cauchy-Schwarz inequality and the definition of the operator norm.[109] Using (C.24), we have $|[M]_{i,j} - [B]_{i,j}| \leq \tau$, which implies that $[M]_{i,j} \geq [B]_{i,j} - \tau \geq b_{\mathsf{min}} - \tau > 0$. (Note that $b_{\mathsf{min}} > 0$ because we have assumed that $P_{X,Y} \in \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}}$, which means that the DTM $B \in \mathcal{B}^\circ$.) Hence, every $M \in \mathcal{M}$ satisfies $M > 0$ entry-wise, and $\mathcal{M} \subseteq \mathcal{B}^\circ$ using part 1 of Theorem 4.2.

Finally, we note that R_τ is the preimage of $\mathcal{M} \subseteq \mathcal{B}^\circ$ under the map $\beta : \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}^\circ$ defined in (4.19). Hence, $R_\tau \subseteq \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}}$. Furthermore, since $\beta : \mathcal{P}^\circ_{\mathcal{X}\times\mathcal{Y}} \to \mathcal{B}^\circ$ is continuous

[109] The inequality in (C.24) portrays that estimating a matrix by operator norm is a stronger condition than estimating it element-wise.

as shown in part 3 of Theorem 4.2, R_τ is closed because $\mathcal{M}$ is closed [239, Corollary, p.87]. Since R_τ is also bounded, it is compact [239, Theorem 2.41].

Part 2: To prove $R_\tau \subseteq S_{4k\tau}$, it suffices to show that:

$$\left| \left\| B\Psi_{(k)}^{\beta(Q_{X,Y})-\sqrt{Q_Y}^T\sqrt{Q_X}} \right\|_{\mathsf{Fro}}^2 - \left\| B\Psi_{(k)} \right\|_{\mathsf{Fro}}^2 \right| \leq 4k \left\| \beta(Q_{X,Y}) - B \right\|_{\mathsf{op}} \tag{C.25}$$

for every $Q_{X,Y} \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. Since the columns of $\Psi_{(k)}^{\beta(Q_{X,Y})-\sqrt{Q_Y}^T\sqrt{Q_X}}$ are the second to $(k+1)$th leading right singular vectors of $\beta(Q_{X,Y})$, and the columns of $\Psi_{(k)}$ are the second to $(k+1)$th leading right singular vectors of B, the proof of Lemma C.2 in appendix C.1 holds verbatim. This yields (C.25) along with the fact that $\|B\|_{\mathsf{op}} = 1$ (cf. part 1 of Theorem 4.1). ■

Proofs from Chapter 5

■ D.1 Proof of Proposition 5.1

Proof. In this proof, we assume familiarity with the development and notation in section 5.5 and the proof of Theorem 5.1.

Part 1: We first prove part 1. Observe that for any $k \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{P}(X_{k,0} \neq X_{0,0}) &= \frac{1}{2}\mathbb{P}\Big(X_{k,0}^{+} = 0\Big) + \frac{1}{2}\mathbb{P}\Big(X_{k,0}^{-} = 1\Big) \\
&= \frac{1}{2}\mathbb{E}\Big[\mathbb{P}\Big(X_{k,0}^{+} = 0\Big|\sigma_k^{+}\Big)\Big] + \frac{1}{2}\mathbb{E}\Big[\mathbb{P}\Big(X_{k,0}^{-} = 1\Big|\sigma_k^{-}\Big)\Big] \\
&= \frac{1}{2}\mathbb{E}\Big[1 - \sigma_k^{+}\Big] + \frac{1}{2}\mathbb{E}\Big[\sigma_k^{-}\Big] \\
&= \frac{1}{2}\left(1 - \mathbb{E}\Big[\sigma_k^{+} - \sigma_k^{-}\Big]\right) \qquad \text{(D.1)}
\end{aligned}$$

where the third equality holds because $X_{k,0} \sim \mathsf{Bernoulli}(\sigma)$ given $\sigma_k = \sigma$. To see this, recall the relation (5.4) from subsection 5.3.1. Using this relation, it is straightforward to verify that X_k is conditionally independent of $X_{0,0}$ given σ_k. Moreover, the conditional distribution $P_{X_k|\sigma_k}$ can be computed using (5.4), and this yields the desired conditional distribution $P_{X_{k,0}|\sigma_k}$ mentioned above. (We omit these calculations because it is intuitively obvious that random bits at level k can be generated by first generating σ_k, then setting a uniformly and randomly chosen subset of vertices in X_k of size $L_k\sigma_k$ to be 1, and finally setting the remaining vertices in X_k to be 0.)

Due to (D.1), it suffices to prove that $\liminf_{k\to\infty} \mathbb{E}[\sigma_k^{+} - \sigma_k^{-}] > 0$. To this end, recall from the proof of Theorem 5.1 that for any sufficiently small $\epsilon = \epsilon(\delta, d) > 0$ (that depends on δ and d) and any $\tau > 0$, there exists $K = K(\epsilon, \tau) \in \mathbb{N}$ (that depends on ϵ and τ) such that for all $k > K$, (5.50) and (5.51) (which are reproduced below) hold:

$$\mathbb{P}(A|E) \geq 1 - \tau \qquad \text{(D.2)}$$

$$\mathbb{P}(B|E) \geq 1 - \tau \qquad \text{(D.3)}$$

where the events are $A = \{\sigma_k^{+} \geq \hat{\sigma} - \epsilon\}$, $B = \{\sigma_k^{-} \leq 1 - \hat{\sigma} + \epsilon\}$, and $E = \{\sigma_K^{+} \geq \hat{\sigma} - \epsilon, \sigma_K^{-} \leq 1 - \hat{\sigma} + \epsilon\}$, respectively. Now notice that for all $k > K$:

$$\mathbb{E}\Big[\sigma_k^{+} - \sigma_k^{-}\Big] = \mathbb{E}\Big[\sigma_k^{+} - \sigma_k^{-}\Big|E\Big]\,\mathbb{P}(E) + \mathbb{E}\Big[\sigma_k^{+} - \sigma_k^{-}\Big|E^c\Big]\,\mathbb{P}(E^c)$$

$$
\begin{aligned}
&\geq \mathbb{E}\left[\sigma_k^+ - \sigma_k^- \middle| E\right] \mathbb{P}(E) \\
&= \mathbb{P}(E) \left(\mathbb{E}\left[\sigma_k^+ \middle| E, A\right] \mathbb{P}(A|E) + \mathbb{E}\left[\sigma_k^+ \middle| E, A^c\right] \mathbb{P}(A^c|E) - \mathbb{E}\left[\sigma_k^- \middle| E\right]\right) \\
&\geq \mathbb{P}(E) \left(\mathbb{E}\left[\sigma_k^+ \middle| E, A\right] \mathbb{P}(A|E) - \mathbb{E}\left[\sigma_k^- \middle| E\right]\right) \\
&= \mathbb{P}(E) \left(\mathbb{E}\left[\sigma_k^+ \middle| E, A\right] \mathbb{P}(A|E)\right. \\
&\qquad\qquad \left. - \mathbb{E}\left[\sigma_k^- \middle| E, B\right] \mathbb{P}(B|E) - \mathbb{E}\left[\sigma_k^- \middle| E, B^c\right] \mathbb{P}(B^c|E)\right) \\
&\geq \mathbb{P}(E) \left(\mathbb{E}\left[\sigma_k^+ \middle| E, A\right] \mathbb{P}(A|E) - \mathbb{E}\left[\sigma_k^- \middle| E, B\right] - \mathbb{P}(B^c|E)\right) \\
&\geq \mathbb{P}(E) \left(\mathbb{E}\left[\sigma_k^+ \middle| E, A\right] (1-\tau) - \mathbb{E}\left[\sigma_k^- \middle| E, B\right] - \tau\right) \\
&\geq \mathbb{P}(E) \left((\hat{\sigma} - \epsilon)(1-\tau) - (1 - \hat{\sigma} + \epsilon) - \tau\right) \\
&= \mathbb{P}(E) \left(\hat{\sigma} - (1-\hat{\sigma}) - 2\epsilon - \tau(1 + \hat{\sigma} - \epsilon)\right) > 0
\end{aligned}
$$

where the second line holds because $\sigma_k^+ \geq \sigma_k^-$ *a.s.* (monotonicity), the fourth line holds because $\sigma_k^+ \geq 0$ *a.s.*, the sixth line holds because $\sigma_k^- \leq 1$ *a.s.*, the seventh line follows from (D.2) and (D.3), the eighth line follows from the definitions of A and B, and the quantity in the ninth line does not depend on k and is strictly positive for sufficiently small ϵ and τ (which now depends on δ and d) because $\hat{\sigma} > 1 - \hat{\sigma}$. Therefore, $\liminf_{k\to\infty} \mathbb{E}[\sigma_k^+ - \sigma_k^-] > 0$, which completes the proof of part 1.

Part 2: We next prove part 2. We begin with a few seemingly unrelated observations that will actually be quite useful later. Recall that $R_k = \inf_{n \geq k} L_n$ for every $k \in \mathbb{N} \cup \{0\}$ and $R_k = O(d^{2k})$. Hence, there exists a constant $\alpha = \alpha(\delta, d) > 0$ (that depends on δ and d) such that for all sufficiently large k (depending on δ and d), we have:

$$
R_k \leq \alpha d^{2k} \,. \tag{D.4}
$$

Let $\beta = \frac{\log(\alpha)}{6\log(d)}$, and define the sequence $\{m(k) \in \mathbb{N} \cup \{0\}\}$ (indexed by k) as:

$$
m = m(k) \triangleq \left\lfloor \frac{\log\left(R_{\lfloor (2k/3) - \beta \rfloor}\right)}{4\log(d)} \right\rfloor \tag{D.5}
$$

where $\frac{2k}{3} \geq \beta$ for all sufficiently large k (depending on δ and d) so that the sequence is eventually well-defined. This sequence satisfies the following conditions:

$$
\lim_{k\to\infty} m(k) = \infty \,, \tag{D.6}
$$

$$
\lim_{k\to\infty} \frac{d^{2m}}{R_{k-m}} = 0 \,. \tag{D.7}
$$

The first limit (D.6) holds because $\lim_{k\to\infty} R_k = \liminf_{k\to\infty} L_k = \infty$ (by assumption), and the second limit (D.7) is true because for all sufficiently large k (depending on δ

and d):

$$\frac{d^{2m}}{R_{k-m}} \leq \frac{\sqrt{R_{\lfloor (2k/3)-\beta \rfloor}}}{R_{k-m}} \leq \frac{\sqrt{R_{\lfloor (2k/3)-\beta \rfloor}}}{R_{\lfloor (2k/3)-\beta \rfloor}} = \frac{1}{\sqrt{R_{\lfloor (2k/3)-\beta \rfloor}}}$$

where the first inequality follows from (D.5), and the second inequality holds because $\{R_k : k \in \mathbb{N} \cup \{0\}\}$ is non-decreasing, and $m \leq \big(\log(\alpha d^{(4k/3)-2\beta})\big)/(4\log(d)) = \frac{k}{3} + \beta$ for all sufficiently large k using (D.4) and (D.5).

We next establish that a small portion of the random DAG G above the vertex $X_{k,0}$ is a directed tree with high probability. To this end, for any sufficiently large $k \in \mathbb{N}$ (depending on δ and d) such that $k - m \geq 0$, let G_k denote the (random) induced subgraph of the random DAG G consisting of all vertices in levels $k-m, \ldots, k$ that have a path to $X_{k,0}$, where $m = m(k)$ is defined in (D.5). (Note that $X_{k,0}$ always has a path to itself.) Moreover, define the event $T_k \triangleq \{G_k \text{ is a directed tree}\}$. Now, for any sufficiently large k (depending on δ and d) such that $d^{2r} \leq R_{k-r} \leq L_{k-r}$ for every $r \in \{1, \ldots, m\}$ (which is feasible due to (D.7), and ensures that the ensuing steps are valid), notice that:

$$\begin{aligned}
\mathbb{P}(T_k) &= \prod_{r=1}^{m} \prod_{s=0}^{d^r-1} \left(1 - \frac{s}{L_{k-r}}\right) \\
&\geq \prod_{r=1}^{m} \left(1 - \frac{1}{L_{k-r}} \sum_{s=0}^{d^r-1} s\right) \\
&= \prod_{r=1}^{m} \left(1 - \frac{d^r(d^r-1)}{2L_{k-r}}\right) \\
&\geq 1 - \frac{1}{2} \sum_{r=1}^{m} \frac{d^r(d^r-1)}{L_{k-r}} \\
&\geq 1 - \frac{1}{2R_{k-m}} \sum_{r=1}^{m} d^{2r} \\
&= 1 - \frac{1}{2R_{k-m}} \left(\frac{d^2(d^{2m}-1)}{d^2-1}\right) \\
&\geq 1 - \left(\frac{d^2}{2(d^2-1)}\right) \frac{d^{2m}}{R_{k-m}}
\end{aligned} \tag{D.8}$$

where the first equality holds because the edges of G are chosen randomly and independently and we must ensure that the parents of every vertex in G_k are distinct, the second and fourth inequalities are straightforward to prove by induction, and the third and sixth equalities follow from arithmetic and geometric series computations, respectively. The bound in (D.8) conveys that $\lim_{k\to\infty} \mathbb{P}(T_k) = 1$ due to (D.7), i.e. G_k is a directed tree with high probability for large k.

We introduce some useful notation for the remainder of this proof. First, condition on any realization of the random DAG G such that the event T_k occurs (for sufficiently

large k such that (D.8) holds). This also fixes the choices of Boolean processing functions at the vertices (which may vary between vertices and be graph dependent). For any vertex $X_{n,j}$ in the tree G_k with $n < k$, let $\tilde{X}_{n,j}$ denote the output of the edge $\mathsf{BSC}(\delta)$ with input $X_{n,j}$ in G_k. (Hence, $\tilde{X}_{n,j}$ is the input of a Boolean processing function at a single vertex in level $n+1$ of G_k.) On the other hand, let $\tilde{X}_{k,0}$ be the output of an independent $\mathsf{BSC}(\delta)$ channel (which is not necessarily in G) with input $X_{k,0}$. Since G_k is a tree, the random variables $\{\tilde{X}_{n,j} : X_{n,j} \text{ is a vertex of } G_k\}$ describe the values at the gates of a *noisy formula* $\tilde{G}_k$, where the Boolean functions in G_k correspond to d-input δ-noisy gates in $\tilde{G}_k$ (and we think of the independent BSC errors as occurring at the gates rather than the edges). Next, in addition to conditioning on G and T_k, we also condition on one of two realizations $X_{k-m} = x_0$ or $X_{k-m} = x_1$ for any $x_0, x_1 \in \{0,1\}^{L_{k-m}}$. In particular, corresponding to any binary random variable Y in $\tilde{G}_k$, define the following 2-tuple in $[0,1]^2$, cf. [86, 115]:

$$\lambda^Y \triangleq \left(\mathbb{P}(Y \neq 0 | X_{k-m} = x_0, G, T_k), \mathbb{P}(Y \neq 1 | X_{k-m} = x_1, G, T_k)\right). \tag{D.9}$$

Lastly, for any constant $a \in [0,1]$, let (cf. [86, 115]):

$$S(a) \triangleq \mathsf{conv}(\{(a,a), (1-a, 1-a), (0,1), (1,0)\}) \subseteq [0,1]^2. \tag{D.10}$$

With these definitions, we can state a version of the pivotal lemma in [86, Lemma 2], which was proved in the $d = 3$ case in [115].

Lemma D.1 (TV Distance Contraction in Noisy Formulae [86, Lemma 2]). *If $d \geq 3$ is odd and $\delta \geq \delta_{\mathsf{maj}}$, then for every possible d-input δ-noisy gate in $\tilde{G}_k$ with inputs $Y_1, \ldots, Y_d$ and output Y, we have:*

$$\lambda^{Y_1}, \ldots, \lambda^{Y_d} \in S(a) \text{ with } a \in \left[0, \frac{1}{2}\right] \quad \Rightarrow \quad \lambda^Y \in S(f(a))$$

where the function $f : [0,1] \to [0,1]$ is defined in (5.19).

We remark that Lemma D.1 differs from [86, Lemma 2] in the definition of the 2-tuple λ^Y for any binary random variable Y in the noisy formula. Since [86, Lemma 2] is used to yield the impossibility results on reliable computation discussed in subsection 5.4.1, [86, Section III] defines λ^Y for this purpose as $\lambda^Y = (\mathbb{P}(Y \neq X | X = 0), \mathbb{P}(Y \neq X | X = 1))$, where X is a single relevant binary input random variable of the noisy formula (and all other inputs are fixed). In contrast, we define λ^Y in (D.9) by conditioning on any two realizations of the random variables X_{k-m}. This ensures that the inputs, say $\tilde{X}_{n,j_1}, \ldots, \tilde{X}_{n,j_d}$ for some $k - m \leq n < k$ and $j_1, \ldots, j_d \in [L_n]$, of every d-input δ-noisy gate in the noisy formula $\tilde{G}_k$ are conditionally independent given X_{k-m}, which is a crucial property required by the proof of [86, Lemma 2]. We omit the proof of Lemma D.1 because it is virtually identical to the proof of [86, Lemma 2] in [86, Sections IV and V]. (The reader can verify that every step in the proofs in [86, Sections IV and V] continues to hold with our definition of λ^Y.)

Lemma D.1 indeed demonstrates a strong data processing inequality style of contraction for TV distance, cf. [279, Equation (1)]. To see this, observe that $(x, y) \in S(a)$ with $a \in [0, \frac{1}{2}]$ if and only if $a \leq ax + (1-a)y \leq 1 - a$ and $a \leq ay + (1-a)x \leq 1 - a$. This implies that $a \leq \frac{x+y}{2} \leq 1 - a$, and hence, $|1 - x - y| \leq 1 - 2a$. Furthermore, for any binary random variable Y in $\tilde{G}_k$, we have using (2.4) (from chapter 2):

$$\begin{aligned}\left\| P_{Y|G,T_k,X_{k-m}=x_1} - P_{Y|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}} = |1 - \mathbb{P}(Y \neq 0|X_{k-m} = x_0, G, T_k) \\ - \mathbb{P}(Y \neq 1|X_{k-m} = x_1, G, T_k)|\end{aligned} \tag{D.11}$$

where $P_{Y|G,T_k,X_{k-m}=x}$ denotes the conditional distribution of Y given $\{X_{k-m} = x, G, T_k\}$ for any $x \in \{0,1\}^{L_{k-m}}$. Thus, if $\lambda^Y \in S(a)$ with $a \in [0, \frac{1}{2}]$, then we get:

$$\left\| P_{Y|G,T_k,X_{k-m}=x_1} - P_{Y|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}} \leq 1 - 2a\,.$$

Now notice that $\lambda^{\tilde{X}_{k-m,j}} \in S(0)$ for every random variable $\tilde{X}_{k-m,j}$ in $\tilde{G}_k$, where $j \in [L_{k-m}]$. As a result, a straightforward induction argument using Lemma D.1 (much like that in the proof in [86, Section III]) yields $\lambda^{\tilde{X}_{k,0}} \in S(f^{(m)}(0))$. This implies that:[110]

$$\left\| P_{\tilde{X}_{k,0}|G,T_k,X_{k-m}=x_1} - P_{\tilde{X}_{k,0}|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}} \leq 1 - 2f^{(m)}(0) = 1 - 2\Big(\delta * g^{(m-1)}(0)\Big) \tag{D.12}$$

where the function $g : [0,1] \to [0,1]$ is given in (5.40) in section 5.5, and the equality follows from (5.20). Moreover, since $\mathbb{P}(\tilde{X}_{k,0} \neq y|G, T_k, X_{k-m} = x) = \delta * \mathbb{P}(X_{k,0} \neq y|G, T_k, X_{k-m} = x)$ for any $y \in \{0,1\}$ and any $x \in \{0,1\}^{L_{k-m}}$, a simple calculation using (D.11) shows that:

$$\begin{aligned}&\left\| P_{\tilde{X}_{k,0}|G,T_k,X_{k-m}=x_1} - P_{\tilde{X}_{k,0}|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}} \\ &\qquad = (1 - 2\delta) \left\| P_{X_{k,0}|G,T_k,X_{k-m}=x_1} - P_{X_{k,0}|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}}\end{aligned}$$

which, using (D.12), produces:

$$\left\| P_{X_{k,0}|G,T_k,X_{k-m}=x_1} - P_{X_{k,0}|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}} \leq \frac{1 - 2\Big(\delta * g^{(m-1)}(0)\Big)}{1 - 2\delta} \tag{D.13}$$

for any $x_0, x_1 \in \{0,1\}^{L_{k-m}}$. The inequality in (D.13) conveys a contraction of the TV distance on the left hand side. Since g has only one fixed point at $\frac{1}{2}$ when $\delta \geq \delta_{\mathsf{maj}}$ (see section 5.5), and (D.6) holds, the *fixed point theorem* (see e.g. [239, Chapter 5, Exercise 22(c)]) gives us $\lim_{k\to\infty} g^{(m-1)}(0) = \frac{1}{2}$, where $g^{(m-1)}(0)$ increases to $\frac{1}{2}$. Hence, the upper bound in (D.13) decreases to 0 as $k \to \infty$. Furthermore, note that (D.13) holds for all choices of Boolean processing functions (which may vary between vertices

[110] The inequality in (D.12) can be perceived as a repeated application of a *tensorized* universal upper bound on the *Dobrushin curve* of any d-input δ-noisy gate, cf. [229, Section II-A].

and be graph dependent), because Lemma D.1 is agnostic to the particular gates used in $\tilde{G}_k$.

Finally, for any fixed realization of the random DAG G such that T_k occurs (for sufficiently large k such that (D.8) holds), observe that:

$$\begin{aligned}
&\left\| P_{X_{k,0}|G,T_k,X_{0,0}=1} - P_{X_{k,0}|G,T_k,X_{0,0}=0} \right\|_{\mathsf{TV}} \\
&\qquad = \eta_{\mathsf{TV}}\left(P_{X_{k,0}|G,T_k,X_0}\right) \\
&\qquad \le \eta_{\mathsf{TV}}\left(P_{X_{k,0}|G,T_k,X_{k-m}}\right) \eta_{\mathsf{TV}}\left(P_{X_{k-m}|G,T_k,X_0}\right) \\
&\qquad \le \max_{x_0,x_1\in\{0,1\}^{L_{k-m}}} \left\| P_{X_{k,0}|G,T_k,X_{k-m}=x_1} - P_{X_{k,0}|G,T_k,X_{k-m}=x_0} \right\|_{\mathsf{TV}} \\
&\qquad \le \frac{1-2\left(\delta * g^{(m-1)}(0)\right)}{1-2\delta}
\end{aligned} \tag{D.14}$$

where $P_{X_{k,0}|G,T_k,X_0}$, $P_{X_{k,0}|G,T_k,X_{k-m}}$, and $P_{X_{k-m}|G,T_k,X_0}$ are transition kernels from X_0 to $X_{k,0}$, from X_{k-m} to $X_{k,0}$, and from X_0 to X_{k-m}, respectively, the first equality follows from the two-point characterization of the Dobrushin contraction coefficient in (2.49) in chapter 2, where $P_{X_{k,0}|G,T_k,X_{0,0}=y}$ denotes the conditional distribution of $X_{k,0}$ given $\{X_{0,0}=y, G, T_k\}$ for any $y\in\{0,1\}$, the second inequality holds because $X_0 \to X_{k-m} \to X_{k,0}$ forms a Markov chain (given G and T_k) and η_{TV} is sub-multiplicative (see (2.52) in chapter 2), the third inequality follows from (2.49), and the last inequality follows from (D.13). Taking conditional expectations with respect to G given T_k in (D.14) yields:

$$\mathbb{E}\left[\left\| P^+_{X_{k,0}|G} - P^-_{X_{k,0}|G} \right\|_{\mathsf{TV}} \middle| T_k\right] \le \frac{1-2\left(\delta * g^{(m-1)}(0)\right)}{1-2\delta}$$

where $P^+_{X_{k,0}|G}$ and $P^-_{X_{k,0}|G}$ inside the conditional expectation correspond to the conditional probability distributions $P_{X_{k,0}|G,T_k,X_{0,0}=1}$ and $P_{X_{k,0}|G,T_k,X_{0,0}=0}$, respectively (as we condition on T_k). Therefore, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\| P^+_{X_{k,0}|G} - P^-_{X_{k,0}|G} \right\|_{\mathsf{TV}}\right] &= \mathbb{E}\left[\left\| P^+_{X_{k,0}|G} - P^-_{X_{k,0}|G} \right\|_{\mathsf{TV}} \middle| T_k\right] \mathbb{P}(T_k) \\
&\quad + \mathbb{E}\left[\left\| P^+_{X_{k,0}|G} - P^-_{X_{k,0}|G} \right\|_{\mathsf{TV}} \middle| T_k^c\right] (1-\mathbb{P}(T_k)) \\
&\le \frac{1-2\left(\delta * g^{(m-1)}(0)\right)}{1-2\delta} + \left(\frac{d^2}{2(d^2-1)}\right)\frac{d^{2m}}{R_{k-m}}
\end{aligned}$$

using the tower property, the fact that TV distance is bounded by 1, and (D.8). Letting $k\to\infty$ establishes the desired result:

$$\lim_{k\to\infty} \mathbb{E}\left[\left\| P^+_{X_{k,0}|G} - P^-_{X_{k,0}|G} \right\|_{\mathsf{TV}}\right] \le \frac{1-2\left(\delta * \lim_{k\to\infty} g^{(m-1)}(0)\right)}{1-2\delta}$$

$$+\left(\frac{d^2}{2(d^2-1)}\right)\lim_{k\to\infty}\frac{d^{2m}}{R_{k-m}}$$
$$=0$$

because $\lim_{k\to\infty} g^{(m-1)}(0) = \frac{1}{2}$ (as noted earlier) and (D.7) holds. This completes the proof. ■

■ D.2 Proof of Corollary 5.1

Proof. This follows from applying the probabilistic method. Fix any $d \geq 3$, any $\delta \in (0, \delta_{\mathsf{maj}})$, and any sequence of level sizes satisfying $L_k \geq C(\delta, d)\log(k)$ for all sufficiently large k. We know from Theorem 5.1 that for the random DAG model with these parameters and majority processing functions, there exist $\epsilon = \epsilon(\delta, d) > 0$ and $K = K(\delta, d) \in \mathbb{N}$ (which depend on δ and d) such that:

$$\forall k \geq K,\ \ \mathbb{P}\Big(\hat{S}_k \neq X_0\Big) \leq \frac{1}{2} - 2\epsilon\,.$$

Now define $P_k(G) \triangleq \mathbb{P}(h^k_{\mathsf{ML}}(X_k, G) \neq X_0|G)$ for $k \in \mathbb{N}\cup\{0\}$ as the conditional probability that the ML decision rule based on the full k-layer state X_k makes an error given the random DAG G, and let E_k for $k \in \mathbb{N}\cup\{0\}$ be the set of all deterministic DAGs $\mathcal{G}$ with indegree d and level sizes $\{L_m : m \in \mathbb{N}\cup\{0\}\}$ such that $P_k(\mathcal{G}) \leq \frac{1}{2} - \epsilon$. Observe that for every $k \geq K$:

$$\begin{aligned}
\frac{1}{2} - 2\epsilon &\geq \mathbb{P}\Big(\hat{S}_k \neq X_0\Big)\\
&= \mathbb{E}\Big[\mathbb{P}\Big(\hat{S}_k \neq X_0\Big|G\Big)\Big]\\
&\geq \mathbb{E}[P_k(G)]\\
&= \mathbb{E}[P_k(G)|G \in E_k]\,\mathbb{P}(G \in E_k) + \mathbb{E}[P_k(G)|G \notin E_k]\,\mathbb{P}(G \notin E_k)\\
&\geq \mathbb{E}[P_k(G)|G \notin E_k]\,\mathbb{P}(G \notin E_k)\\
&\geq \left(\frac{1}{2} - \epsilon\right)\mathbb{P}(G \notin E_k)
\end{aligned}$$

where the second and fourth lines follow from the law of total expectation, the third line holds because the ML decision rule minimizes the probability of error, the fifth line holds because the first term in the previous line is non-negative, and the final line holds because $G \notin E_k$ implies that $P_k(G) > \frac{1}{2} - \epsilon$. Then, we have for every $k \geq K$:

$$\mathbb{P}(G \in E_k) \geq \frac{2\epsilon}{1-2\epsilon} > 0\,.$$

Since $\{E_k : k \in \mathbb{N}\cup\{0\}\}$ form a non-increasing sequence of sets (because $P_k(G)$ is non-decreasing in k), we get via continuity:

$$\mathbb{P}\left(G \in \bigcap_{k\in\mathbb{N}\cup\{0\}} E_k\right) = \lim_{k\to\infty}\mathbb{P}(G \in E_k) \geq \frac{2\epsilon}{1-2\epsilon} > 0$$

which means that there exists a deterministic DAG $\mathcal{G}$ with indegree d, noise level δ, level sizes $\{L_k : k \in \mathbb{N} \cup \{0\}\}$, and majority processing functions such that $P_k(\mathcal{G}) \leq \frac{1}{2} - \epsilon$ for all $k \in \mathbb{N} \cup \{0\}$. This completes the proof. ■

■ D.3 Proof of Proposition 5.2

Proof.

Part 1: We first prove part 1, where we are given a fixed deterministic DAG $\mathcal{G}$. Observe that the BSC along each edge of this DAG produces its output bit by either copying its input bit exactly with probability $1 - 2\delta$, or generating an independent $\mathsf{Bernoulli}(\frac{1}{2})$ output bit with probability 2δ. This is because the BSC's stochastic transition probability matrix can be decomposed as:

$$\underbrace{\begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix}}_{\mathsf{BSC}(\delta) \text{ channel matrix}} = (1-2\delta) \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\text{copy matrix}} + (2\delta) \underbrace{\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}}_{\mathsf{Bernoulli}(\frac{1}{2}) \text{ bit}} . \tag{D.15}$$

We remark that this simple, but useful, idea originates from *Fortuin-Kasteleyn random cluster representations of Ising models* in the study of percolation, cf. [113], and has been exploited in various other discrete probability contexts such as broadcasting on trees [83, p.412], and reliable computation [87, p.570].

Now consider the events:

$$A_k \triangleq \{\text{all } dL_k \text{ edges from level } k-1 \text{ to level } k \text{ generate independent output bits}\}$$

for $k \in \mathbb{N}$, which have probabilities $\mathbb{P}(A_k) = (2\delta)^{dL_k}$ since the BSCs on the edges are independent. These events are mutually independent (once again because the BSCs on the edges are independent). Since the condition on L_k in the proposition statement is equivalent to:

$$(2\delta)^{dL_k} \geq \frac{1}{k} \text{ for all sufficiently large } k\,,$$

we must have:

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \sum_{k=1}^{\infty} (2\delta)^{dL_k} = +\infty\,.$$

The second Borel-Cantelli lemma then tells us that infinitely many of the events $\{A_k : k \in \mathbb{N}\}$ occur almost surely, i.e. $\mathbb{P}(\bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} A_k) = 1$. In particular, if we define $B_m \triangleq \bigcup_{k=1}^{m} A_k$ for $m \in \mathbb{N}$, then by continuity:

$$\lim_{m\to\infty} \mathbb{P}(B_m) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = 1\,. \tag{D.16}$$

Finally, observe that:

$$\lim_{m\to\infty} \mathbb{P}(h^m_{\mathsf{ML}}(X_m, \mathcal{G}) \neq X_0) = \lim_{m\to\infty} \mathbb{P}(h^m_{\mathsf{ML}}(X_m, \mathcal{G}) \neq X_0 | B_m)\, \mathbb{P}(B_m)$$

$$
\begin{aligned}
&\quad + \mathbb{P}(h^m_{\mathsf{ML}}(X_m, \mathcal{G}) \neq X_0 | B_m^c)\, \mathbb{P}(B_m^c) \\
&= \lim_{m\to\infty} \mathbb{P}(h^m_{\mathsf{ML}}(X_m, \mathcal{G}) \neq X_0 | B_m) \\
&= \lim_{m\to\infty} \frac{1}{2}\, \mathbb{P}(h^m_{\mathsf{ML}}(X_m, \mathcal{G}) = 1 | B_m) \\
&\qquad + \frac{1}{2}\, \mathbb{P}(h^m_{\mathsf{ML}}(X_m, \mathcal{G}) = 0 | B_m) \\
&= \frac{1}{2}
\end{aligned} \tag{D.17}
$$

where $h^m_{\mathsf{ML}}(\cdot, \mathcal{G}) : \{0,1\}^{L_m} \to \{0,1\}$ denotes the ML decision rule at level m based on X_m (given knowledge of the DAG $\mathcal{G}$), the second equality uses (D.16), and the third equality holds because $X_{0,0} \sim \mathsf{Bernoulli}(\frac{1}{2})$ is independent of B_m, and X_m is conditionally independent of X_0 given B_m. The condition in (D.17) is equivalent to the TV distance condition in part 1 of the proposition statement; this proves part 1.

Part 2: To prove part 2, notice that part 1 immediately yields:

$$
\lim_{k\to\infty} \left\| P^+_{X_k|G} - P^-_{X_k|G} \right\|_{\mathsf{TV}} = 0 \quad \textit{pointwise}
$$

which completes the proof. ■

■ D.4 Proof of Proposition 5.3

Proof.

Part 1: Fix any noise level $\delta \in (0, \frac{1}{2})$ and any constant $\epsilon \in (0, \frac{1}{4})$. Furthermore, tentatively suppose that $L_k \geq A(\epsilon, \delta)\sqrt{\log(k)}$ for all sufficiently large k, where the constant $A(\epsilon, \delta)$ is defined as:

$$
A(\epsilon, \delta) \triangleq \frac{2}{(1-2\delta)\epsilon\sqrt{1-2\epsilon}}\,. \tag{D.18}
$$

Now consider the deterministic DAG $\mathcal{G}$ such that each vertex at level $k \in \mathbb{N}$ is connected to all L_{k-1} vertices at level $k-1$ and all Boolean processing functions are the majority rule. (Note that when there is only one input, the majority rule behaves like the identity map.) For all $k \in \mathbb{N}$, since X_k is an exchangeable sequence of random variables given σ_0, σ_k is a sufficient statistic of X_k for performing inference about σ_0, where σ_k is defined in (5.2) (cf. subsection 5.3.1). We next prove a useful "one-step broadcasting" lemma involving σ_k's for this model.

Lemma D.2 (One-Step Broadcasting in Unbounded Degree DAG). *Under the aforementioned assumptions, there exists $K = K(\epsilon, \delta) \in \mathbb{N}$ (that depends on ϵ and δ) such that for all $k \geq K$, we have:*

$$
\mathbb{P}\left(\sigma_k \geq \frac{1}{2} + \epsilon \,\middle|\, \sigma_{k-1} \geq \frac{1}{2} + \epsilon \right) \geq 1 - \left(\frac{1}{k-1} \right)^2 .
$$

Proof. Suppose we are given that $\sigma_{k-1} = \sigma \geq \frac{1}{2} + \epsilon$ for any $k \in \mathbb{N}$. Then, $\{X_{k,j} : j \in [L_k]\}$ are conditionally i.i.d. $\mathsf{Bernoulli}(\mathbb{P}(X_{k,0} = 1|\sigma_{k-1} = \sigma))$ and $L_k\sigma_k \sim \mathsf{binomial}(L_k, \mathbb{P}(X_{k,0} = 1|\sigma_{k-1} = \sigma))$, where $\mathbb{P}(X_{k,0} = 1|\sigma_{k-1} = \sigma) = \mathbb{E}[\sigma_k|\sigma_{k-1} = \sigma]$. Furthermore, since $X_{k,0}$ is the majority of the values of $X_{k-1,0}, \ldots, X_{k-1,L_{k-1}-1}$ after passing them through independent $\mathsf{BSC}(\delta)$'s, we have:

$$\begin{aligned}
\mathbb{E}[\sigma_k|\sigma_{k-1} = \sigma] &= \mathbb{P}(X_{k,0} = 1|\sigma_{k-1} = \sigma) \\
&= 1 - \mathbb{P}\left(\sum_{i=1}^{L_{k-1}\sigma} Z_i + \sum_{j=1}^{L_{k-1}(1-\sigma)} Y_j < \frac{L_{k-1}}{2}\right) \\
&\geq 1 - \exp\left(-2L_{k-1}\left(\frac{1}{2} - \sigma * \delta\right)^2\right) \\
&\geq 1 - \exp\left(-2L_{k-1}\left(\frac{1}{2} - \left(\frac{1}{2} + \epsilon\right) * \delta\right)^2\right) \\
&= 1 - \exp\left(-2L_{k-1}\epsilon^2(1-2\delta)^2\right) \qquad \text{(D.19)}
\end{aligned}$$

where Z_i are i.i.d. $\mathsf{Bernoulli}(1-\delta)$, Y_j are i.i.d. $\mathsf{Bernoulli}(\delta)$, $\{Z_i : i \in \{1, \ldots, L_{k-1}\sigma\}\}$ and $\{Y_j : j \in \{1, \ldots, L_{k-1}(1-\sigma)\}\}$ are independent, the first inequality follows from Hoeffding's inequality (see Lemma C.4 in appendix C.2) using the fact that $\sigma * \delta > \frac{1}{2}$ (because $\sigma > \frac{1}{2}$), and the second inequality holds because $\sigma \geq \frac{1}{2} + \epsilon$, which implies that $\sigma * \delta \geq \left(\frac{1}{2} + \epsilon\right) * \delta > \frac{1}{2}$.

Next, observe that there exists $K = K(\epsilon, \delta) \in \mathbb{N}$ (that depends on ϵ and δ) such that for all $k \geq K$, we have:

$$\exp\left(-2L_{k-1}\epsilon^2(1-2\delta)^2\right) \leq \epsilon$$

because $\lim_{k\to\infty} L_k = \infty$ by assumption. So, for any $k \geq K$, this yields the bound:

$$\mathbb{E}[\sigma_k|\sigma_{k-1} = \sigma] \geq 1 - \epsilon > \frac{1}{2} + \epsilon$$

using (D.19) and the fact that $\epsilon < \frac{1}{4}$. As a result, we can apply the Chernoff-Hoeffding bound (see Lemma C.5 in appendix C.2), to σ_k for any $k \geq K$ and get:

$$\mathbb{P}\left(\sigma_k < \frac{1}{2} + \epsilon \,\middle|\, \sigma_{k-1} = \sigma\right) \leq \exp\left(-L_k D\left(\frac{1}{2} + \epsilon \,\middle\|\, \mathbb{E}[\sigma_k|\sigma_{k-1} = \sigma]\right)\right).$$

Notice that:

$$\begin{aligned}
D\left(\frac{1}{2} + \epsilon \,\middle\|\, \mathbb{E}[\sigma_k|\sigma_{k-1} = \sigma]\right) &\geq -H\left(\frac{1}{2} + \epsilon\right) - \left(\frac{1}{2} - \epsilon\right)\log(1 - \mathbb{E}[\sigma_k|\sigma_{k-1} = \sigma]) \\
&\geq L_{k-1}\epsilon^2(1-2\epsilon)(1-2\delta)^2 - H\left(\frac{1}{2} + \epsilon\right)
\end{aligned}$$

where $H(\cdot)$ denotes the binary Shannon entropy function (see e.g. Proposition 5.7), the first inequality holds because $\log(\mathbb{E}[\sigma_k|\sigma_{k-1}=\sigma]) < 0$, and the second inequality follows from (D.19). Hence, we have for any $k \geq K$:

$$\mathbb{P}\left(\sigma_k < \frac{1}{2}+\epsilon \,\middle|\, \sigma_{k-1}=\sigma\right) \leq \exp\left(-L_{k-1}L_k\epsilon^2(1-2\epsilon)(1-2\delta)^2 + L_k H\left(\frac{1}{2}+\epsilon\right)\right)$$

where we can multiply both sides by $\mathbb{P}(\sigma_{k-1}=\sigma)$ and then sum over all $\sigma \geq \frac{1}{2}+\epsilon$ (as in the proof of (5.46) within the proof of Theorem 5.1 in section 5.5) to get:

$$\mathbb{P}\left(\sigma_k < \frac{1}{2}+\epsilon \,\middle|\, \sigma_{k-1} \geq \frac{1}{2}+\epsilon\right) \leq \exp\left(-L_{k-1}L_k\left(\epsilon^2(1-2\epsilon)(1-2\delta)^2 - \frac{1}{L_{k-1}}H\left(\frac{1}{2}+\epsilon\right)\right)\right).$$

Since $\lim_{k\to\infty} L_k = \infty$ by assumption, we can choose $K = K(\epsilon,\delta)$ to be sufficiently large so that for all $k \geq K$, we also have $H(\frac{1}{2}+\epsilon)/L_{k-1} \leq \epsilon^2(1-2\epsilon)(1-2\delta)^2/2$. Thus, for every $k \geq K$:

$$\mathbb{P}\left(\sigma_k < \frac{1}{2}+\epsilon \,\middle|\, \sigma_{k-1} \geq \frac{1}{2}+\epsilon\right) \leq \exp\left(-L_{k-1}L_k\,\frac{\epsilon^2(1-2\epsilon)(1-2\delta)^2}{2}\right). \tag{D.20}$$

Finally, we once again increase $K = K(\epsilon,\delta)$ if necessary to ensure that $L_{k-1} \geq A(\epsilon,\delta)\sqrt{\log(k-1)}$ for every $k \geq K$ (as presumed earlier). This implies that for all $k \geq K$:

$$L_{k-1}L_k \geq A(\epsilon,\delta)^2\sqrt{\log(k-1)\log(k)} \geq A(\epsilon,\delta)^2\log(k-1)$$

which, using (D.18) and (D.20), produces:

$$\mathbb{P}\left(\sigma_k < \frac{1}{2}+\epsilon \,\middle|\, \sigma_{k-1} \geq \frac{1}{2}+\epsilon\right) \leq \exp(-2\log(k-1)) = \left(\frac{1}{k-1}\right)^2$$

for all $k \geq K$. This proves Lemma D.2. ■

Lemma D.2 is an analogue of (5.46) in the proof of Theorem 5.1 in section 5.5. It illustrates that if the proportion of 1's is large in a given layer of $\mathcal{G}$, then it remains large in the next layer of $\mathcal{G}$ with high probability.

To proceed, we specialize Lemma D.2 by arbitrarily selecting a particular value of ϵ, say $\epsilon = \frac{7}{32} \in (0, \frac{1}{4})$. This implies that the constant $A(\epsilon,\delta)$ becomes:

$$A(\delta) = A\left(\frac{7}{32},\delta\right) = \frac{256}{21(1-2\delta)} \tag{D.21}$$

using (D.18). In the proposition statement, it is assumed that $L_k \geq A(\delta)\sqrt{\log(k)}$ for all sufficiently large k. Thus, Lemma D.2 holds with $\epsilon = \frac{7}{32} \in (0, \frac{1}{4})$ under the assumptions of part 1 of Proposition 5.3. At this point, we can execute the proof of part 1 of Theorem

5.1 in section 5.5 mutatis mutandis (with Lemma D.2 playing the pivotal role of (5.46)) to establish part 1 of Proposition 5.3. We omit the details of this proof for brevity.

Part 2: To prove part 2, we use the proof technique of part 1 of Proposition 5.2 in appendix D.3. Recall that the $\mathsf{BSC}(\delta)$ along each edge of the DAG $\mathcal{G}$ produces its output bit by either copying its input bit with probability $1-2\delta$, or generating an independent $\mathsf{Bernoulli}(\frac{1}{2})$ output bit with probability 2δ. As before, consider the mutually independent events:

$$A_k \triangleq \{\text{all } L_{k-1}L_k \text{ edges from level } k-1 \text{ to level } k \text{ generate independent output bits}\}$$

for $k \in \mathbb{N}$, which have probabilities $\mathbb{P}(A_k) = (2\delta)^{L_{k-1}L_k}$. Define the constant $B(\delta)$ as:

$$B(\delta) \triangleq \frac{1}{\sqrt{\log\left(\frac{1}{2\delta}\right)}}. \tag{D.22}$$

Since we assume in the proposition statement that $L_k \leq B(\delta)\sqrt{\log(k)}$ for all sufficiently large k, we have:

$$L_{k-1}L_k \leq \frac{\sqrt{\log(k-1)\log(k)}}{\log\left(\frac{1}{2\delta}\right)} \leq \frac{\log(k)}{\log\left(\frac{1}{2\delta}\right)}$$

for all sufficiently large k, which implies that:

$$(2\delta)^{L_{k-1}L_k} \geq \frac{1}{k}$$

for all sufficiently large k. Hence, we get $\sum_{k=1}^{\infty}\mathbb{P}(A_k) = \sum_{k=1}^{\infty}(2\delta)^{L_{k-1}L_k} = +\infty$, and the second Borel-Cantelli lemma establishes that infinitely many of the events $\{A_k : k \in \mathbb{N}\}$ occur almost surely, or equivalently, $\mathbb{P}(\bigcap_{m=1}^{\infty}\bigcup_{k=m}^{\infty}A_k) = 1$. As a result, we can define $B_m \triangleq \bigcup_{k=1}^{m}A_k$ for $m \in \mathbb{N}$ such that $\lim_{m\to\infty}\mathbb{P}(B_m) = 1$. Therefore, we have (as before):

$$\lim_{m\to\infty}\mathbb{P}(h_{\mathsf{ML}}^m(X_m, \mathcal{G}) \neq X_0) = \frac{1}{2}$$

where $h_{\mathsf{ML}}^m(\cdot, \mathcal{G}) : \{0,1\}^{L_m} \to \{0,1\}$ denotes the ML decision rule at level m based on X_m (given knowledge of the DAG $\mathcal{G}$). This completes the proof. ■

■ D.5 Proof of Proposition 5.5

Proof. Recall that $L_k\sigma_k \sim \mathsf{binomial}(L_k, g(\sigma))$ given $\sigma_{k-1} = \sigma$. This implies via Hoeffding's inequality (see Lemma C.4 in appendix C.2) and (5.41) that for every $k \in \mathbb{N}$ and $\epsilon_k > 0$:

$$\mathbb{P}(|\sigma_k - g(\sigma_{k-1})| > \epsilon_k | \sigma_{k-1} = \sigma) \leq 2\exp\left(-2L_k\epsilon_k^2\right)$$

where we can take expectations with respect to σ_{k-1} to get:

$$\mathbb{P}(|\sigma_k - g(\sigma_{k-1})| > \epsilon_k) \leq 2\exp\left(-2L_k\epsilon_k^2\right). \tag{D.23}$$

Now fix any $\tau > 0$, and choose a sufficiently large integer $K = K(\tau) \in \mathbb{N}$ (that depends on τ) such that:

$$\begin{aligned}\mathbb{P}(\exists k > K,\, |\sigma_k - g(\sigma_{k-1})| > \epsilon_k) &\leq \sum_{k=K+1}^{\infty} \mathbb{P}(|\sigma_k - g(\sigma_{k-1})| > \epsilon_k) \\ &\leq 2 \sum_{k=K+1}^{\infty} \exp\Big(-2L_k \epsilon_k^2\Big) \\ &\leq \tau\end{aligned}$$

where we use the union bound and (D.23), and let $\epsilon_k = \sqrt{\log(k)/L_k}$ (or equivalently, $\exp(-2L_k\epsilon_k^2) = 1/k^2$). This implies that for any $\tau > 0$:

$$\mathbb{P}(\forall k > K,\, |\sigma_k - g(\sigma_{k-1})| \leq \epsilon_k) \geq 1 - \tau\,. \tag{D.24}$$

Since for every $k > K$, $|\sigma_k - g(\sigma_{k-1})| \leq \epsilon_k$, we can recursively obtain the following relation:

$$\forall k \in \mathbb{N}\backslash\{1,\ldots,K\},\quad \Big|\sigma_k - g^{(k-K)}(\sigma_K)\Big| \leq \sum_{m=K+1}^{k} D(\delta,d)^{k-m}\epsilon_m \tag{D.25}$$

where $D(\delta,d)$ denotes the Lipschitz constant of g on $[0,1]$ as defined in (5.45), and $D(\delta,d) \in (0,1)$ since $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$. Since $L_m = \omega(\log(m))$, for any $\epsilon > 0$, we can take $K = K(\epsilon,\tau)$ (which depends on both ϵ and τ) to be sufficiently large so that $\sup_{m>K} \epsilon_m \leq \epsilon(1 - D(\delta,d))$. Now observe that we have:

$$\begin{aligned}\forall k \in \mathbb{N}\backslash\{1,\ldots,K\},\quad \sum_{m=K+1}^{k} D(\delta,d)^{k-m}\epsilon_m &\leq \left(\sup_{m>K} \epsilon_m\right) \sum_{j=0}^{\infty} D(\delta,d)^j \\ &= \left(\sup_{m>K} \epsilon_m\right) \frac{1}{1 - D(\delta,d)} \\ &\leq \epsilon\,.\end{aligned}$$

Moreover, since $g : [0,1] \to [0,1]$ is a contraction when $\delta \in (\delta_{\mathsf{maj}}, \frac{1}{2})$, it has a unique fixed point $\sigma = \frac{1}{2}$, and $\lim_{m\to\infty} g^{(m)}(\sigma_K) = \frac{1}{2}$ almost surely by the fixed point theorem. As a result, for any $\tau > 0$ and any $\epsilon > 0$, there exists $K = K(\epsilon,\tau) \in \mathbb{N}$ such that:

$$\mathbb{P}\Big(\forall k > K,\, \Big|\sigma_k - g^{(k-K)}(\sigma_K)\Big| \leq \epsilon\Big) \geq 1 - \tau$$

which implies, after letting $k \to \infty$, that:

$$\mathbb{P}\left(\frac{1}{2} - \epsilon \leq \liminf_{k\to\infty} \sigma_k \leq \limsup_{k\to\infty} \sigma_k \leq \frac{1}{2} + \epsilon\right) \geq 1 - \tau\,.$$

Lastly, we can first let $\epsilon \to 0$ and employ the continuity of $\mathbb{P}$, and then let $\tau \to 0$ to obtain:

$$\mathbb{P}\left(\lim_{k\to\infty} \sigma_k = \frac{1}{2}\right) = 1\,.$$

This completes the proof. ■

■ D.6 Proof of Proposition 5.6

Proof. This proof is analogous to the proof of Proposition 5.5 in appendix D.5. For every $k \in \mathbb{N}$ and $\epsilon_k > 0$, we have after taking expectations in (5.71) that:

$$\mathbb{P}(|\sigma_{2k} - g(\sigma_{2k-2})| > \epsilon_k) \leq 4 \exp\left(-\frac{\hat{L}_k \epsilon_k^2}{8}\right) \tag{D.26}$$

where $\hat{L}_k = \min\{L_{2k}, L_{2k-1}\}$ for $k \in \mathbb{N}$. Now fix any $\tau > 0$, and choose a sufficiently large integer $K = K(\tau) \in \mathbb{N}$ (that depends on τ) such that:

$$\begin{aligned}\mathbb{P}(\exists k > K,\, |\sigma_{2k} - g(\sigma_{2k-2})| > \epsilon_k) &\leq \sum_{m=K+1}^{\infty} \mathbb{P}(|\sigma_{2m} - g(\sigma_{2m-2})| > \epsilon_m) \\ &\leq 4 \sum_{m=K+1}^{\infty} \exp\left(-\frac{\hat{L}_m \epsilon_m^2}{8}\right) \leq \tau\end{aligned}$$

where we use the union bound and (D.26), and we set $\epsilon_m = 4\big(\log(m)/\hat{L}_m\big)^{1/2}$ (which ensures that $\exp(-\hat{L}_m \epsilon_m^2/8) = 1/m^2$). This implies that for any $\tau > 0$:

$$\mathbb{P}(\forall k > K,\, |\sigma_{2k} - g(\sigma_{2k-2})| \leq \epsilon_k) \geq 1 - \tau\,. \tag{D.27}$$

Since for every $k > K$, $|\sigma_{2k} - g(\sigma_{2k-2})| \leq \epsilon_k$, we can recursively obtain the following relation:

$$\forall k \in \mathbb{N}\backslash\{1, \ldots, K\},\ \left|\sigma_{2k} - g^{(k-K)}(\sigma_{2K})\right| \leq \sum_{m=K+1}^{k} D(\delta)^{k-m} \epsilon_m \tag{D.28}$$

where $D(\delta)$ denotes the Lipschitz constant of g on $[0,1]$ as shown in (5.64), which is in $(0,1)$ since $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$. Since $L_m = \omega(\log(m))$, for any $\epsilon > 0$, we can take $K = K(\epsilon, \tau) \in \mathbb{N}$ (which depends on both ϵ and τ) to be sufficiently large so that $\sup_{m>K} \epsilon_m \leq \epsilon(1 - D(\delta))$. This implies that:

$$\forall k \in \mathbb{N}\backslash\{1, \ldots, K\},\ \sum_{m=K+1}^{k} D(\delta)^{k-m} \epsilon_m \leq \epsilon$$

as shown in the proof of Proposition 5.5. Moreover, since $g : [0,1] \to [0,1]$ is a contraction when $\delta \in (\delta_{\mathsf{andor}}, \frac{1}{2})$, it has a unique fixed point $\sigma = t \in [0,1]$, and $\lim_{m\to\infty} g^{(m)}(\sigma_{2K}) = t$ almost surely by the fixed point theorem. As a result, for any $\tau > 0$ and any $\epsilon > 0$, there exists $K = K(\epsilon, \tau) \in \mathbb{N}$ such that:

$$\mathbb{P}\Big(\forall k > K,\, \left|\sigma_{2k} - g^{(k-K)}(\sigma_{2K})\right| \leq \epsilon\Big) \geq 1 - \tau$$

which implies that:

$$\mathbb{P}\left(\lim_{k\to\infty} \sigma_{2k} = t\right) = 1$$

once again as shown in the proof of Proposition 5.5. This completes the proof. ■

■ D.7 Proof of Corollary 5.2

Proof. Fix any $\epsilon > 0$ and any $d \geq \left(\frac{2}{\epsilon}\right)^5$, and let $\alpha = d^{-6/5}$. To establish the first part of the corollary, it suffices to prove that for every $n \in \mathbb{N}$:

$$\begin{aligned} & n\left(1 - (1-\alpha)^d - \sqrt{2d\alpha H(\alpha)}\right) \geq (1-\epsilon)d\alpha n \\ \Leftrightarrow \quad & 1 - \frac{1 - (1-\alpha)^d - \sqrt{2d\alpha H(\alpha)}}{d\alpha} \leq \epsilon\,. \end{aligned}$$

Indeed, if this is true, then Proposition 5.7 immediately implies the desired lower bound on the probability that B is a d-regular bipartite lossless $(d^{-6/5}, (1-\epsilon)d)$-expander graph. To this end, observe that:

$$\begin{aligned} 1 - \frac{1 - (1-\alpha)^d - \sqrt{2d\alpha H(\alpha)}}{d\alpha} &\leq 1 - \frac{1 - e^{-d\alpha}}{d\alpha} + \frac{\sqrt{2d\alpha H(\alpha)}}{d\alpha} \\ &\leq \frac{d\alpha}{2} + \sqrt{\frac{2H(\alpha)}{d\alpha}} \\ &\leq \frac{d\alpha}{2} + \sqrt{\frac{4\log(2)}{d\sqrt{\alpha}}} \\ &= \frac{1}{2d^{1/5}} + \frac{2\sqrt{\log(2)}}{d^{1/5}} \\ &\leq \frac{2}{d^{1/5}} \\ &\leq \epsilon \end{aligned}$$

where the first inequality follows from the standard bound $(1-\alpha)^d \leq e^{-d\alpha}$ for $\alpha \in (0,1)$ and $d \in \mathbb{N}$, the second inequality follows from the easily verifiable bounds $0 \leq 1 - \frac{1-e^{-x}}{x} \leq \frac{x}{2}$ for $x > 0$, the third inequality follows from the well-known bound $H(\alpha) \leq 2\log(2)\sqrt{\alpha(1-\alpha)} \leq 2\log(2)\sqrt{\alpha}$ for $\alpha \in (0,1)$, the fourth equality follows from substituting $\alpha = d^{-6/5}$, the fifth inequality follows from direct computation, and the final inequality holds because $d \geq \left(\frac{2}{\epsilon}\right)^5$. This proves the first part of the corollary.

The existence of d-regular bipartite lossless $(d^{-6/5}, (1-\epsilon)d)$-expander graphs for every sufficiently large n (depending on d) in the second part of the corollary follows from the first part by invoking the probabilistic method. This completes the proof. ■

Bibliography

[1] E. Abbe, *Community Detection and Stochastic Block Models*, ser. Foundations and Trends in Communications and Information Theory. Hanover, MA, USA: now Publishers Inc., 2018, vol. 14, no. 1-2.

[2] E. Abbe and L. Zheng, "Linear universal decoding for compound channels: an Euclidean geometric approach," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Toronto, ON, Canada, July 6-11 2008, pp. 1098–1102.

[3] E. Abbe and L. Zheng, "A coordinate system for Gaussian networks," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 721–733, February 2012.

[4] J. A. Adell, A. Lekuona, and Y. Yu, "Sharp bounds on the entropy of the Poisson law and related quantities," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2299–2306, May 2010.

[5] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *The Annals of Probability*, vol. 4, no. 6, pp. 925–939, December 1976.

[6] G. Ajjanagadde and Y. Polyanskiy, "Adder MAC and estimates for Rényi entropy," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 29-October 2 2015, pp. 434–441.

[7] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the Second International Symposium on Information Theory (Tsaghkadzor, Armenia, USSR, September 2-8 1971)*, B. N. Petrov and F. Csaki, Eds. Budapest, Hungary: Akadémiai Kiadó, 1973, pp. 267–281.

[8] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[9] S. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. Translations of Mathematical Monographs. Providence, RI, USA: American Mathematical Society, Oxford University Press, 2000, vol. 191.

[10] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between Boolean functions," in *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 2-4 2013, pp. 13–19.

[11] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," April 2013, arXiv:1304.6133 [cs.IT].

[12] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Honolulu, HI, USA, June 29-July 4 2014, pp. 3022–3026.

[13] M. Artin, *Algebra*, 2nd ed., ser. Pearson Modern Classics for Advanced Mathematics Series. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2010.

[14] J. M. Ash, A. E. Gatto, and S. Vági, "A multidimensional Taylor's theorem with minimal hypothesis," *Colloquium Mathematicum*, vol. 60-61, no. 1, pp. 245–252, 1990.

[15] R. B. Ash, *Information Theory*, ser. Interscience Tracts in Pure and Applied Mathematics. New York, NY, USA: John Wiley & Sons, Inc., 1965, no. 19.

[16] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, pp. 1–37, January 2016, article no. 15.

[17] S. Axler, *Linear Algebra Done Right*, 2nd ed., ser. Undergraduate Texts in Mathematics. New York, NY, USA: Springer, 2004.

[18] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer, 2010.

[19] D. Bakry, "Functional inequalities for Markov semigroups," in *Probability Measures on Groups: Recent Directions and Trends*, ser. Proceedings of the CIMPA-TIFR School, Tata Institute of Fundamental Research, Mumbai, India, 2002, S. G. Dani and P. Graczyk, Eds. New Delhi, India: Narosa Publishing House, 2006, pp. 91–147.

[20] A. S. Bandeira, "Ten lectures and forty-two open problems in the mathematics of data science," October 2016, Department of Mathematics, MIT, Cambridge, MA, USA, Lecture Notes 18.S096 (Fall 2015).

[21] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, October 2005.

[22] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the Advances in Neural Information Processing Systems 14 (NIPS)*, Vancouver, BC, Canada, December 3-8 2001, pp. 585–591.

[23] J. Bennett and S. Lanning, "The Netflix prize," in *Proceedings of KDD Cup and Workshop*, San Jose, CA, USA, August 12 2007, pp. 3–6.

[24] J.-P. Benzécri, *L'Analyse des Données, Tôme 2: L'Analyse des Correspondances*. Paris, France: Dunod, 1973, in French.

[25] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Transactions on Information Theory*, vol. IT-19, no. 2, pp. 197–207, March 1973.

[26] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, ser. Classics in Applied Mathematics. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1994, vol. 9.

[27] R. Bhatia, *Matrix Analysis*, ser. Graduate Texts in Mathematics. New York, NY, USA: Springer, 1997, vol. 169.

[28] N. Bhatnagar, J. Vera, E. Vigoda, and D. Weitz, "Reconstruction for colorings on trees," *SIAM Journal on Discrete Mathematics*, vol. 25, no. 2, pp. 809–826, July 2011.

[29] R. Bhattacharya and E. C. Waymire, "Iterated random maps and some classes of Markov processes," in *Stochastic Processes: Theory and Methods*, ser. Handbook of Statistics, D. N. Shanbhag and C. R. Rao, Eds., vol. 19. Amsterdam, Netherlands: North-Holland, Elsevier, 2001, pp. 145–170.

[30] D. Blackwell, "Comparison of experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, CA, USA, July 31-August 12 1950)*, J. Neyman, Ed. Berkeley, CA, USA: University of California Press, 1951, pp. 93–102.

[31] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov, "On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice," *Journal of Statistical Physics*, vol. 79, no. 1-2, pp. 473–482, April 1995.

[32] B. Bollobás, *Random Graphs*, 2nd ed., ser. Cambridge Studies in Advanced Mathematics. Cambridge, United Kingdom: Cambridge University Press, 2001, vol. 73.

[33] S. Borade and L. Zheng, "Euclidean information theory," in *Proceedings of the IEEE International Zurich Seminar on Communications*, Zurich, Switzerland, March 12-14 2008, pp. 14–17.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[35] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, September 1985.

[36] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, ser. Texts in Applied Mathematics, J. E. Marsden, L. Sirovich, and S. S. Antman, Eds. New York, NY, USA: Springer, 1999, vol. 31.

[37] A. Buja, "Theory of bivariate ACE," Department of Statistics, University of Washington, Seattle, WA, USA, Tech. Rep. 74, December 1985.

[38] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7235–7254, November 2011.

[39] M. R. Capalbo, O. Reingold, S. P. Vadhan, and A. Wigderson, "Randomness conductors and constant-degree lossless expanders," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, Montréal, QC, Canada, May 19-21 2002, pp. 659–668.

[40] E. Carlen, "Trace inequalities and quantum entropy: An introductory course," in *Entropy and the Quantum: Arizona School of Analysis with Applications, March 16-20 2009*, ser. Contemporary Mathematics, R. Sims and D. Ueltschi, Eds., vol. 529. Tucson, AZ, USA: American Mathematical Society, 2010, pp. 73–140.

[41] E. Çinlar, *Probability and Stochastics*, ser. Graduate Texts in Mathematics. New York, NY, USA: Springer, February 2011, vol. 261.

[42] D. Chafaï, "Singular values of random matrices," November 2009, Université Paris-Est Marne-la-Vallée, Paris, France, Lecture Notes.

[43] S.-C. Chang and E. J. Weldon Jr., "Coding for T-user multiple-access channels," *IEEE Transactions on Information Theory*, vol. IT-25, no. 6, pp. 684–691, November 1979.

[44] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *The Annals of Mathematical Statistics*, vol. 22, no. 4, pp. 581–586, December 1951.

[45] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, vol. 43, no. 1, pp. 177–214, February 2015.

[46] M.-D. Choi, M. B. Ruskai, and E. Seneta, "Equivalence of certain entropy contraction coefficients," *Linear Algebra and its Applications, Elsevier*, vol. 208-209, pp. 29–36, September 1994.

[47] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, March 1990.

[48] B. S. Cirel'son, "Reliable storage of information in a system of unreliable components with local interactions," in *Locally Interacting Systems and Their Application in Biology*, ser. Lecture Notes in Mathematics, R. L. Dobrushin, V. I. Kryukov, and A. L. Toom, Eds., vol. 653. Berlin, Heidelberg, Germany: Springer, 1978, pp. 15–30.

[49] J. E. Cohen, Y. Iwasa, G. Rautu, M. B. Ruskai, E. Seneta, and G. Zbăganu, "Relative entropy under mappings by stochastic matrices," *Linear Algebra and its Applications, Elsevier*, vol. 179, pp. 211–235, January 1993.

[50] J. E. Cohen, J. H. B. Kemperman, and G. Zbăganu, *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences*. Ann Arbor, MI, USA: Birkhäuser, 1998.

[51] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis, Elsevier*, vol. 21, no. 1, pp. 5–30, July 2006.

[52] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. IT-18, no. 1, pp. 2–14, January 1972.

[53] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[54] I. Csiszár, "Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von Markoffschen ketten," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, Ser. A*, vol. 8, pp. 85–108, January 1963, in German.

[55] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, January 1967.

[56] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, no. 1-4, pp. 191–213, March 1972.

[57] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Transactions on Infomation Theory*, vol. IT-24, no. 3, pp. 339–348, May 1978.

[58] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. New York, NY, USA: Cambridge University Press, 2011.

[59] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*, ser. Foundations and Trends in Communications and Information Theory, S. Verdú, Ed. Hanover, MA, USA: now Publishers Inc., 2004, vol. 1, no. 4.

[60] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[61] G. Dahl, "Majorization polytopes," *Linear Algebra and its Applications, Elsevier*, vol. 297, pp. 157–175, August 1999.

[62] G. Dahl, "Matrix majorization," *Linear Algebra and its Applications, Elsevier*, vol. 288, pp. 53–73, February 1999.

[63] C. Daskalakis, E. Mossel, and S. Roch, "Optimal phylogenetic reconstruction," in *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC)*, Seattle, WA, USA, May 21-23 2006, pp. 159–168.

[64] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1501–1518, November 1991.

[65] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., ser. Stochastic Modelling and Applied Probability. New York, NY, USA: Springer, 1998, vol. 38.

[66] J. W. Demmel, *Applied Numerical Linear Algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1997.

[67] P. Diaconis, *Group Representations in Probability and Statistics*, ser. Lecture Notes-Monograph Series, S. S. Gupta, Ed. Hayward, CA, USA: Institute of Mathematical Statistics, 1988, vol. 11.

[68] P. Diaconis, K. Khare, and L. Saloff-Coste, "Gibbs sampling, exponential families and orthogonal polynomials," *Statistical Science*, vol. 23, no. 2, pp. 151–178, May 2008.

[69] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *The Annals of Applied Probability*, vol. 6, no. 3, pp. 695–750, August 1996.

[70] P. Diaconis and L. Saloff-Coste, "Nash inequalities for finite Markov chains," *Journal of Theoretical Probability*, vol. 9, no. 2, pp. 459–510, April 1996.

[71] S. N. Diggavi and M. Grossglauser, "On transmission over deletion channels," in *Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 3-5 2001, pp. 573–582.

[72] R. L. Dobrushin, "Central limit theorem for nonstationary Markov chains. I," *Theory of Probability and Its Applications*, vol. 1, no. 1, pp. 65–80, 1956.

[73] R. L. Dobrushin and S. I. Ortyukov, "Lower bound for the redundancy of self-correcting arrangements of unreliable functional elements," *Problemy Peredachi Informatsii*, vol. 13, no. 1, pp. 82–89, 1977, in Russian.

[74] S. S. Dragomir and V. Gluščević, "Some inequalities for the Kullback-Leibler and χ^2-distances in information theory and applications," *Tamsui Oxford Journal of Mathematical Sciences*, vol. 17, no. 2, pp. 97–111, 2001.

[75] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011–5038, August 2017.

[76] F. du Pin Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1879–1892, March 2018.

[77] R. Durrett, "Oriented percolation in two dimensions," *The Annals of Probability*, vol. 12, no. 4, pp. 999–1040, November 1984.

[78] G. K. Eagleson, "Polynomial expansions of bivariate distributions," *The Annals of Mathematical Statistics*, vol. 35, no. 3, pp. 1208–1215, September 1964.

[79] A. Edelman, "Random matrix theory," February 2016, Department of Mathematics, MIT, Cambridge, MA, USA, Lecture Notes 18.338.

[80] A. El Gamal, "The capacity of a class of broadcast channels," *IEEE Transactions on Information Theory*, vol. IT-25, no. 2, pp. 166–169, March 1979.

[81] A. El Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge University Press, 2011.

[82] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[83] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman, "Broadcasting on trees and the Ising model," *The Annals of Applied Probability*, vol. 10, no. 2, pp. 410–433, May 2000.

[84] W. Evans and N. Pippenger, "On the maximum tolerable noise for reliable computation by formulas," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1299–1305, May 1998.

[85] W. S. Evans and L. J. Schulman, "Signal propagation and noisy circuits," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2367–2373, November 1999.

[86] W. S. Evans and L. J. Schulman, "On the maximum tolerable noise of k-input gates for reliable computation by formulas," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 3094–3098, November 2003.

[87] T. Feder, "Reliable computation by networks in the presence of noise," *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 569–571, May 1989.

[88] A. A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1491–1498, June 2003.

[89] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York, NY, USA: John Wiley & Sons, Inc., 1968, vol. 1.

[90] N. J. Fine, "Binomial coefficients modulo a prime," *The American Mathematical Monthly*, vol. 54, no. 10, pp. 589–592, December 1947.

[91] S. Friedli and Y. Velenik, *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. New York, NY, USA: Cambridge University Press, 2018.

[92] P. Gács, "A new version of Toom's proof," Department of Computer Science, Boston University, Boston, MA, USA, Tech. Rep. TR 95-009, 1995.

[93] P. Gács, "Reliable cellular automata with self-organization," *Journal of Statistical Physics*, vol. 103, no. 1-2, pp. 45–267, April 2001.

[94] M. Gadouleau and A. Goupil, "Binary codes for packet error and packet loss correction in store and forward," in *Proceedings of the International ITG Conference on Source and Channel Coding (SCC)*, no. 25, Siegen, Germany, January 18-21 2010, pp. 1–6.

[95] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[96] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, December 1941, in German.

[97] Y. Geng, C. Nair, S. Shamai, and Z. V. Wang, "On broadcast channels with binary inputs and symmetric outputs," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 6980–6989, November 2013.

[98] H.-O. Georgii, *Gibbs Measures and Phase Transitions*, 2nd ed., ser. De Gruyter Studies in Mathematics. Berlin, Germany: De Gruyter, 2011, vol. 9.

[99] A. Gerschenfeld and A. Montanari, "Reconstruction for models on random graphs," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Providence, RI, USA, October 20-23 2007, pp. 194–204.

[100] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, December 2002.

[101] G. L. Gilardoni, "On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5377–5386, November 2010.

[102] A. Giovagnoli and H. P. Wynn, "Cyclic majorization and smoothing operators," *Linear Algebra and its Applications, Elsevier*, vol. 239, pp. 215–225, May 1996.

[103] A. Gohari, O. Günlü, and G. Kramer, "Coding for positive rate in the source model key agreement problem," August 2018, arXiv:1709.05174v4 [cs.IT].

[104] A. A. Gohari and V. Anantharam, "Evaluation of Marton's inner bound for the general broadcast channel," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 608–619, February 2012.

[105] S. Goldstein, "Maximal coupling," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 46, no. 2, pp. 193–204, January 1979.

[106] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., ser. Johns Hopkins Studies in the Mathematical Sciences. Baltimore, MD, USA: The Johns Hopkins University Press, 1996.

[107] L. F. Gray, "The positive rates problem for attractive nearest neighbor spin systems on $\mathbb{Z}$," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 61, no. 3, pp. 389–404, September 1982.

[108] L. F. Gray, "The behavior of processes with statistical mechanical properties," in *Percolation Theory and Ergodic Theory of Infinite Particle Systems*, ser. The IMA Volumes in Mathematics and Its Applications, H. Kesten, Ed., vol. 8. New York, NY, USA: Springer, 1987, pp. 131–167.

[109] L. F. Gray, "A reader's guide to Gacs's "positive rates" paper," *Journal of Statistical Physics*, vol. 103, no. 1-2, pp. 1–44, April 2001.

[110] M. Greenacre and T. Hastie, "The geometric interpretation of correspondence analysis," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 437–447, June 1987.

[111] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*. San Diego, CA, USA: Academic Press, March 1984.

[112] R. C. Griffiths, "The canonical correlation coefficients of bivariate gamma distributions," *The Annals of Mathematical Statistics*, vol. 40, no. 4, pp. 1401–1408, August 1969.

[113] G. Grimmett, "Percolation and disordered systems," in *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVI-1996*, ser. Lecture Notes in Mathematics, P. Bernard, Ed., vol. 1665. Berlin, Heidelberg, Germany: Springer, 1997, pp. 153–300.

[114] L. Györfi and I. Vajda, "A class of modified Pearson and Neyman statistics," *Statistics and Decisions*, vol. 19, no. 3, pp. 239–251, January 2001.

[115] B. Hajek and T. Weller, "On the maximum tolerable noise for reliable computation by formulas," *IEEE Transactions on Infomation Theory*, vol. 37, no. 2, pp. 388–391, March 1991.

[116] J. M. Hammersley, "On estimating restricted parameters," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 12, no. 2, pp. 192–240, 1950.

[117] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[118] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6343–6354, November 2015.

[119] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 4th ed. New York, NY, USA: Springer, 2015.

[120] G. H. Hardy, *A Course of Pure Mathematics*, 10th ed. London, UK: Cambridge University Press, 1967.

[121] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 1st ed. London, UK: Cambridge University Press, 1934.

[122] P. Harremoës and I. Vajda, "On pairs of f-divergences and their joint range," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3230–3235, June 2011.

[123] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 25-30 2017, pp. 3130–3134.

[124] E. Heinz, "Beiträge zur störungstheorie der spektralzerlegung," *Mathematische Annalen*, vol. 123, pp. 415–438, 1951, in German.

[125] H. O. Hirschfeld, "A connection between correlation and contingency," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4, pp. 520–524, October 1935.

[126] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.

[127] A. E. Holroyd, I. Marcovici, and J. B. Martin, "Percolation games, probabilistic cellular automata, and the hard-core model," *Probability Theory and Related Fields*, pp. 1–31, 2018.

[128] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. New York, NY, USA: Cambridge University Press, 1991.

[129] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2013.

[130] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441 and 498–520, September 1933.

[131] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, December 1936.

[132] S.-L. Huang, A. Makur, F. Kozynski, and L. Zheng, "Efficient statistics: Extracting information from iid observations," in *Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 1-3 2014, pp. 699–706.

[133] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," in preparation.

[134] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "An information-theoretic view of learning in high dimensions: Universal features, maximal correlations, bottlenecks, and common information," in *Proceedings of the Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, February 11-16 2018, presentation.

[135] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "An information-theoretic approach to universal feature selection in high-dimensional inference," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 25-30 2017, pp. 1336–1340.

[136] S.-L. Huang, C. Suh, and L. Zheng, "Euclidean information theory of networks," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6795–6814, December 2015.

[137] S.-L. Huang, G. W. Wornell, and L. Zheng, "Gaussian universal features, canonical correlations, and common information," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Guangzhou, China, November 25-29 2018.

[138] S.-L. Huang, L. Zhang, and L. Zheng, "An information-theoretic approach to unsupervised feature selection for high-dimensional data," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Kaohsiung, Taiwan, November 6-10 2017, pp. 434–438.

[139] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Cambridge, MA, USA, July 1-6 2012, pp. 1029–1033.

[140] S.-L. Huang and L. Zheng, "The linear information coupling problems," June 2014, arXiv:1406.2834 [cs.IT].

[141] S.-L. Huang and L. Zheng, "A spectrum decomposition to the feature spaces and the application to big data analytics," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, China, June 14-19 2015, pp. 661–665.

[142] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*. New York, NY, USA: Cambridge University Press, 2010.

[143] D. Ioffe, "Extremality of the disordered state for the Ising model on general trees," in *Trees: Workshop in Versailles, June 14-16 1995*, ser. Progress in Probability, B. Chauvin, S. Cohen, and A. Rouault, Eds., vol. 40. Basel, Switzerland: Birkhäuser, 1996, pp. 3–14.

[144] D. Ioffe, "On the extremality of the disordered state for the Ising model on the Bethe lattice," *Letters in Mathematical Physics*, vol. 37, no. 2, pp. 137–143, June 1996.

[145] I. C. F. Ipsen and T. M. Selee, "Ergodicity coefficients defined by vector norms," *SIAM Journal on Matrix Analysis and Applications*, vol. 32, no. 1, pp. 153–200, March 2011.

[146] S. Janson and E. Mossel, "Robust reconstruction on trees is determined by the second eigenvalue," *The Annals of Probability*, vol. 32, no. 3B, pp. 2630–2649, July 2004.

[147] S. Kamath and V. Anantharam, "Non-interactive simulation of joint distributions: The Hirschfeld-Gebelein-Rényi maximal correlation and the hypercontractivity ribbon," in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 1-5 2012, pp. 1057–1064.

[148] S. Kamath, A. Orlitsky, V. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *28th Annual Conference on Learning Theory (COLT), Proceedings of Machine Learning Research (PMLR)*, vol. 40, Paris, France, July 3-6 2015, pp. 1066–1100.

[149] W. Kang and S. Ulukus, "A new data processing inequality and its applications in distributed source and channel coding," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 56–69, January 2011.

[150] R. W. Keener, *Theoretical Statistics: Topics for a Core Course*, ser. Springer Texts in Statistics. New York, NY, USA: Springer, 2010.

[151] H. Kesten and B. P. Stigum, "A limit theorem for multidimensional Galton-Watson processes," *The Annals of Mathematical Statistics*, vol. 37, no. 5, pp. 1211–1223, October 1966.

[152] H. Kim, W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Discovering potential correlations via hypercontractivity," *Entropy*, vol. 19, no. 11, pp. 1–32, November 2017, article no. 586.

[153] D. E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms*, 3rd ed. Reading, MA, USA: Addison-Wesley, 1997, vol. 2.

[154] A. N. Kolmogorov and Y. M. Barzdin, "On the realization of networks in three-dimensional space," in *Selected Works of A. N. Kolmogorov – Volume III: Information Theory and the Theory of Algorithms*, ser. Mathematics and Its Applications (Soviet Series), A. N. Shiryayev, Ed., vol. 27. Dordrecht, Netherlands: Springer, 1993, pp. 194–202, originally published in: Problemy Kibernetiki, no. 19, pp. 261-268, 1967.

[155] A. Kontorovich, "Obtaining measure concentration from Markov contraction," *Markov Processes and Related Fields*, vol. 18, no. 4, pp. 613–638, January 2012.

[156] J. Körner and K. Marton, "Comparison of two noisy channels," in *Topics in Information Theory*, ser. Second Colloquium, Keszthely, Hungary, 1975, I. Csiszár and P. Elias, Eds. Amsterdam, Netherlands: North-Holland, 1977, pp. 411–423.

[157] A. E. Koudou, "Probabilités de Lancaster," *Expositiones Mathematicae*, vol. 14, no. 3, pp. 247–275, 1996, in French.

[158] A. E. Koudou, "Lancaster bivariate probability distributions with Poisson, negative binomial and gamma margins," *Test*, vol. 7, no. 1, pp. 95–110, June 1998.

[159] M. Kovačević and V. Y. F. Tan, "Codes in the space of multisets–Coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, July 2018.

[160] M. Kovačević and D. Vukobratović, "Subset codes for packet networks," *IEEE Communications Letters*, vol. 17, no. 4, pp. 729–732, April 2013.

[161] M. Kovačević and D. Vukobratović, "Perfect codes in the discrete simplex," *Designs, Codes and Cryptography*, vol. 75, no. 1, pp. 81–95, April 2015.

[162] F. Krząkała, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová, "Gibbs states and the set of solutions of random constraint satisfaction problems," *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 104, no. 25, pp. 10 318–10 323, June 2007.

[163] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, March 1951.

[164] G. R. Kumar and T. A. Courtade, "Which Boolean functions are most informative?" in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Istanbul, Turkey, July 7-12 2013, pp. 226–230.

[165] H. O. Lancaster, "The structure of bivariate distributions," *The Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 719–736, September 1958.

[166] H. O. Lancaster, *The Chi-Squared Distribution*. New York, NY, USA: John Wiley & Sons Inc., 1969.

[167] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 1986.

[168] M. Ledoux, "Concentration of measure and logarithmic Sobolev inequalities," in *Séminaire de Probabilités XXXIII*, ser. Lecture Notes in Mathematics, J. Azéma, M. Émery, M. Ledoux, and M. Yor, Eds., vol. 1709. Berlin, Heidelberg, Germany: Springer, 1999, pp. 120–216.

[169] M. Leshno and Y. Spector, "An elementary proof of Blackwell's theorem," *Mathematical Social Sciences, Elsevier*, vol. 25, no. 1, pp. 95–98, December 1992.

[170] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, 1st ed. Providence, RI, USA: American Mathematical Society, 2009.

[171] C.-K. Li and N.-K. Tsing, "Some isometries of rectangular complex matrices," *Linear and Multilinear Algebra*, vol. 23, no. 1, pp. 47–53, 1988.

[172] Y.-C. Li and C.-C. Yeh, "Some equivalent forms of Bernoulli's inequality: A survey," *Applied Mathematics*, vol. 4, no. 7, pp. 1070–1093, 2013.

[173] F. Liese and I. Vajda, *Convex Statistical Distances*, ser. Teubner-Texte Zur Mathematik. Leipzig, Germany: Teubner, 1987, vol. 95.

[174] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[175] T. M. Liggett, "Attractive nearest neighbor spin systems on the integers," *The Annals of Probability*, vol. 6, no. 4, pp. 629–636, August 1978.

[176] R. Linsker, "Self-organization in a perceptual network," *Computer, IEEE*, vol. 21, no. 3, pp. 105–117, March 1988.

[177] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. IT-28, no. 2, pp. 129–137, March 1982.

[178] K. Löwner, "Über monotone matrixfunktionen," *Mathematische Zeitschrift*, vol. 38, no. 1, pp. 177–216, December 1934, in German.

[179] R. Lyons and Y. Peres, *Probability on Trees and Networks*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. New York, NY, USA: Cambridge University Press, 2017, vol. 42.

[180] A. Makur, "A study of local approximations in information theory," Masters Thesis in Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, June 2015.

[181] A. Makur, "Information capacity BSC and BEC permutation channels," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 2-5 2018, pp. 1112–1119.

[182] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 29-October 2 2015, pp. 972–979.

[183] A. Makur, E. Mossel, and Y. Polyanskiy, "Broadcasting on two-dimensional regular grids," in preparation.

[184] A. Makur, E. Mossel, and Y. Polyanskiy, "Broadcasting on bounded degree DAGs," March 2018, arXiv:1803.07527 [cs.IT].

[185] A. Makur, E. Mossel, and Y. Polyanskiy, "Broadcasting on random directed acyclic graphs," November 2018, arXiv:1811.03946 [cs.IT].

[186] A. Makur, E. Mossel, and Y. Polyanskiy, "Broadcasting on random networks," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Paris, France, July 7-12 2019, pp. 1–5.

[187] A. Makur and Y. Polyanskiy, "Less noisy domination by symmetric channels," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 25-30 2017, pp. 2463–2467.

[188] A. Makur and Y. Polyanskiy, "Comparison of channels: Criteria for domination by a symmetric channel," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5704–5725, August 2018.

[189] A. Makur and L. Zheng, "Bounds between contraction coefficients," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 29-October 2 2015, pp. 1422–1429.

[190] A. Makur and L. Zheng, "Polynomial spectral decomposition of conditional expectation operators," in *Proceedings of the 54th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 27-30 2016, pp. 633–640.

[191] A. Makur and L. Zheng, "Polynomial singular value decompositions of a family of source-channel models," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7716–7728, December 2017.

[192] A. Makur and L. Zheng, "Linear bounds between contraction coefficients for f-divergences," July 2018, arXiv:1510.01844v4 [cs.IT].

[193] G. A. Margulis, "Explicit constructions of concentrators," *Problemy Peredachi Informatsii*, vol. 9, no. 4, pp. 71–80, 1973, in Russian.

[194] G. A. Margulis, "Probabilistic characteristics of graphs with large connectivity," *Problemy Peredachi Informatsii*, vol. 10, no. 2, pp. 101–108, 1974, in Russian.

[195] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, 2nd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer, 2011.

[196] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, March 1960.

[197] F. G. Mehler, "Ueber die entwicklung einer function von beliebig vielen variablen nach laplaceschen functionen höherer ordnung," *Journal für die reine und angewandte Mathematik*, vol. 66, pp. 161–176, 1866, in German.

[198] J. Meixner, "Orthogonale polynomsysteme mit einer besonderen gestalt der erzeugenden funktion," *Journal of the London Mathematical Society*, vol. 9, pp. 6–13, January 1934, in German.

[199] R. Melrose, "Functional analysis," May 2017, Department of Mathematics, MIT, Cambridge, MA, USA, Lecture Notes 18.102.

[200] N. Merhav, *Statistical Physics and Information Theory*, ser. Foundations and Trends in Communications and Information Theory. Hanover, MA, USA: now Publishers Inc., 2010, vol. 6, no. 1-2.

[201] J. J. Metzner, "Simplification of packet-symbol decoding with errors, deletions, misordering of packets, and no sequence numbers," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2626–2639, June 2009.

[202] M. Mézard and A. Montanari, "Reconstruction on trees and spin glass transition," *Journal of Statistical Physics*, vol. 124, no. 6, pp. 1317–1350, September 2006.

[203] L. Miclo, "Remarques sur l'hypercontractivité at l'évolution de éntropie pour des chaînes de Markov finies," in *Séminaire de Probabilités XXXI*, ser. Lecture Notes in Mathematics, J. Azéma, M. Yor, and M. Émery, Eds., vol. 1655. Berlin, Heidelberg, Germany: Springer, 1997, pp. 136–167, in French.

[204] M. Mitzenmacher, "Polynomial time low-density parity-check codes with rates very close to the capacity of the q-ary random deletion channel for large q," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5496–5501, December 2006.

[205] A. Montanari, R. Restrepo, and P. Tetali, "Reconstruction and clustering in random constraint satisfaction problems," *SIAM Journal on Discrete Mathematics*, vol. 25, no. 2, pp. 771–808, July 2011.

[206] R. Montenegro and P. Tetali, *Mathematical Aspects of Mixing Times in Markov Chains*, ser. Foundations and Trends in Theoretical Computer Science, M. Sudan, Ed. Hanover, MA, USA: now Publishers Inc., 2006, vol. 1, no. 3.

[207] T. Morimoto, "Markov processes and the H-theorem," *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, March 1963.

[208] C. N. Morris, "Natural exponential families with quadratic variance functions," *The Annals of Statistics*, vol. 10, no. 1, pp. 65–80, March 1982.

[209] C. N. Morris, "Natural exponential families with quadratic variance functions: Statistical theory," *The Annals of Statistics*, vol. 11, no. 2, pp. 515–529, June 1983.

[210] E. Mossel, "Recursive reconstruction on periodic trees," *Random Structures and Algorithms*, vol. 13, no. 1, pp. 81–97, August 1998.

[211] E. Mossel, "Reconstruction on trees: Beating the second eigenvalue," *The Annals of Applied Probability*, vol. 11, no. 1, pp. 285–300, February 2001.

[212] E. Mossel, "On the impossibility of reconstructing ancestral data and phylogenies," *Journal of Computational Biology*, vol. 10, no. 5, pp. 669–676, July 2003.

[213] E. Mossel, "Phase transitions in phylogeny," *Transactions of the American Mathematical Society*, vol. 356, no. 6, pp. 2379–2404, June 2004.

[214] E. Mossel, "Survey: Information flow on trees," in *Graphs, Morphisms and Statistical Physics, DIMACS Workshop, March 19-21 2001*, ser. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, J. Nešetřil and P. Winkler, Eds., vol. 63. New Brunswick, NJ, USA: American Mathematical Society, 2004, pp. 155–170.

[215] E. Mossel, K. Oleszkiewicz, and A. Sen, "On reverse hypercontractivity," *Geometric and Functional Analysis*, vol. 23, no. 3, pp. 1062–1097, June 2013.

[216] C. Nair, "Capacity regions of two new classes of 2-receiver broadcast channels," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Seoul, South Korea, June 28-July 3 2009, pp. 1839–1843.

[217] C. Nair, "An extremal inequality related to hypercontractivity of Gaussian random variables," in *Proceedings of the Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, February 9-14 2014, pp. 1–7.

[218] J. Neyman, "Contribution to the theory of the χ^2 test," in *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, CA, USA, August 13-18 1945 and January 27-29 1946)*, J. Neyman, Ed. Berkeley, CA, USA: University of California Press, 1949, pp. 239–273.

[219] F. Nie, X. Wang, C. Deng, and H. Huang, "Learning a structured optimal bipartite graph for co-clustering," in *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS)*, Long Beach, CA, USA, December 4-9 2017, pp. 4132–4141.

[220] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, January 2014.

[221] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed., ser. Prentice-Hall Signal Processing Series. Upper Saddle River, NJ, USA: Pearson, 2010.

[222] E. Ordentlich and M. J. Weinberger, "A distribution dependent refinement of Pinsker's inequality," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.

[223] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[224] G. K. Pedersen, "Some operator monotone functions," *Proceedings of the American Mathematical Society*, vol. 36, no. 1, pp. 309–310, November 1972.

[225] R. Pemantle and Y. Peres, "The critical Ising model on trees, concave recursions and nonlinear capacity," *The Annals of Probability*, vol. 38, no. 1, pp. 184–206, January 2010.

[226] M. S. Pinsker, "On the complexity of a concentrator," in *Proceedings of the 7th International Teletraffic Congress (ITC)*, Stockholm, Sweden, June 13-20 1973, pp. 318/1–318/4.

[227] Y. Polyanskiy, "Saddle point in the minimax converse for channel coding," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2576–2595, May 2013.

[228] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[229] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 35–55, January 2016.

[230] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," August 2017, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, Lecture Notes 6.441.

[231] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*, ser. The IMA Volumes in Mathematics and its Applications, E. Carlen, M. Madiman, and E. M. Werner, Eds., vol. 161. New York, NY, USA: Springer, 2017, pp. 211–249.

[232] E. C. Posner, "Random coding strategies for minimum entropy," *IEEE Transactions on Information Theory*, vol. IT-21, no. 4, pp. 388–391, July 1975.

[233] D. Qiu, A. Makur, and L. Zheng, "Probabilistic clustering using maximal matrix norm couplings," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 2-5 2018, pp. 1020–1027.

[234] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[235] V. Rakočević and H. K. Wimmer, "A variational characterization of canonical angles between subspaces," *Journal of Geometry*, vol. 78, no. 1, pp. 122–124, 2003.

[236] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[237] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, UK: Cambridge University Press, 2008.

[238] S. Roch, "Toward extracting all phylogenetic information from matrices of evolutionary distances," *Science*, vol. 327, no. 5971, pp. 1376–1379, March 2010.

[239] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., ser. International Series in Pure and Applied Mathematics. New York, NY, USA: McGraw-Hill, 1976.

[240] L. Russo, "On the critical percolation probabilities," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 56, no. 2, pp. 229–237, June 1981.

[241] A. Samorodnitsky, "On the entropy of a noisy function," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5446–5464, October 2016.

[242] O. V. Sarmanov, "Maximal correlation coefficient (non-symmetric case)," *Doklady Akademii Nauk SSSR*, vol. 121, no. 1, pp. 52–55, 1958, in Russian.

[243] I. Sason, "Bounds on f-divergences and related distances," Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel, Irwin and Joan Jacobs Center for Communication and Information Technologies (CCIT) 859, May 2014.

[244] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating f-divergences," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Jerusalem, Israel, April 26-May 1 2015, pp. 1–5.

[245] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, November 2016.

[246] G. Schiebinger, M. J. Wainwright, and B. Yu, "The geometry of kernelized spectral clustering," *The Annals of Statistics*, vol. 43, no. 2, pp. 819–846, April 2015.

[247] T. M. Selee, "Stochastic matrices: Ergodicity coefficients, and applications to ranking," PhD Thesis in Applied Mathematics, North Carolina State University, Raleigh, NC, USA, 2009.

[248] E. Seneta, "Coefficients of ergodicity: Structure and applications," *Advances in Applied Probability*, vol. 11, no. 3, pp. 576–590, September 1979.

[249] E. Seneta, *Non-negative Matrices and Markov Chains*, 2nd ed., ser. Springer Series in Statistics. New York, NY, USA: Springer, 1981.

[250] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 706–715, September 1956.

[251] C. E. Shannon, "A note on a partial ordering for communication channels," *Information and Control*, vol. 1, no. 4, pp. 390–397, December 1958.

[252] S. Sherman, "On a theorem of Hardy, Littlewood, Polya, and Blackwell," *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 37, no. 12, pp. 826–831, December 1951.

[253] N. Shutty, M. Wootters, and P. Hayden, "Noise thresholds for amplification: From quantum foundations to classical fault-tolerant computation," September 2018, arXiv:1809.09748 [cs.IT].

[254] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, July 2017.

[255] Y. G. Sinai, *Theory of Phase Transitions: Rigorous Results*, ser. International Series in Natural Philosophy. Oxford, United Kingdom: Pergamon Press, 1982, vol. 108.

[256] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1710–1722, November 1996.

[257] A. Sly, "Reconstruction of random colourings," *Communications in Mathematical Physics*, vol. 288, no. 3, pp. 943–961, June 2009.

[258] A. Sly, "Reconstruction for the Potts model," *The Annals of Probability*, vol. 39, no. 4, pp. 1365–1406, July 2011.

[259] M. Steel and J. Hein, "Reconstructing pedigrees: A combinatorial perspective," *Journal of Theoretical Biology, Elsevier*, vol. 240, no. 3, pp. 360–367, June 2006.

[260] C. Stein, "Notes on a seminar on theoretical statistics. I. Comparison of experiments," University of Chicago, Tech. Rep., 1951.

[261] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, ser. Princeton Lectures in Analysis. Princeton, NJ, USA: Princeton University Press, 2003, vol. 1.

[262] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, ser. Princeton Lectures in Analysis. Princeton, NJ, USA: Princeton University Press, 2005, vol. 3.

[263] G. W. Stewart, "Perturbation theory for the singular value decomposition," in *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, R. J. Vaccaro, Ed. New York, NY, USA: Elsevier, 1991, pp. 99–109.

[264] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, ser. Computer Science and Scientific Computing. New York, NY, USA: Academic Press, 1990.

[265] C. Stępniak, "Ordering of nonnegative definite matrices with application to comparison of linear models," *Linear Algebra and its Applications, Elsevier*, vol. 70, pp. 67–71, October 1985.

[266] G. Strang, *Linear Algebra and Its Applications*, 4th ed. New York, NY, USA: Cengage Learning, 2005.

[267] F. E. Su, "Methods for quantifying rates of convergence for random walks on groups," PhD Thesis in Mathematics, Harvard University, Cambridge, MA, USA, 1995.

[268] D. Sutter and J. M. Renes, "Universal polar codes for more capable and less noisy channels and sources," April 2014, arXiv:1312.5990v3 [cs.IT].

[269] V. Y. F. Tan, *Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities*, ser. Foundations and Trends in Communications and Information Theory, S. Verdú, Ed. Hanover, MA, USA: now Publishers Inc., 2014, vol. 11, no. 1-2.

[270] T. Tao, *Topics in Random Matrix Theory*, ser. Graduate Studies in Mathematics. Providence, RI, USA: American Mathematical Society, 2010, vol. 132.

[271] E. A. Thompson, *Pedigree Analysis in Human Genetics*, ser. Johns Hopkins Series in Contemporary Medicine and Public Health. Baltimore, MD, USA: Johns Hopkins University Press, 1986.

[272] R. Tibshirani, "Data mining," Spring 2013, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA, Lecture Notes 36-462.

[273] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 22-24 1999, pp. 368–377.

[274] A. L. Toom, "Stable and attractive trajectories in multicomponent systems," in *Multicomponent Random Systems*, ser. Advances in Probability and Related Topics, R. L. Dobrushin and Y. G. Sinai, Eds., vol. 6. New York, NY, USA: Marcel Dekker, 1980, pp. 549–575, translation from Russian.

[275] E. Torgersen, *Comparison of Statistical Experiments*, ser. Encyclopedia of Mathematics and Its Applications. New York, NY, USA: Cambridge University Press, 1991.

[276] E. Torgersen, "Stochastic orders and comparison of experiments," in *Stochastic Orders and Decision Under Risk*, ser. Lecture Notes-Monograph Series, K. Mosler and M. Scarsini, Eds., vol. 19. Hayward, CA, USA: Institute of Mathematical Statistics, 1991, pp. 334–371.

[277] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, August 2012.

[278] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 2009.

[279] F. Unger, "Noise threshold for universality of 2-input gates," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Nice, France, June 24-29 2007, pp. 1901–1905.

[280] F. Unger, "Noise threshold for universality of two-input gates," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3693–3698, August 2008.

[281] F. Unger, "Better gates can make fault-tolerant computation impossible," *Electronic Colloquium on Computational Complexity (ECCC)*, no. 164, pp. 1–17, November 2010.

[282] M. van Dijk, "On a special class of broadcast channels with confidential messages," *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 712–714, March 1997.

[283] S. Verdú, "Total variation distance and the distribution of relative information," in *Proceedings of the Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, February 9-14 2014, pp. 1–3.

[284] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg, Germany: Springer, 2009, vol. 338.

[285] I. Vincze, "On the concept and measure of information contained in an observation," in *Contributions to Probability: A Collection of Papers Dedicated to Eugene Lukacs*, J. Gani and V. K. Rohatgi, Eds. New York, NY, USA: Academic Press, 1981, pp. 207–214.

[286] J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," in *Automata Studies*, ser. Annals of Mathematics Studies, C. E. Shannon and J. McCarthy, Eds., vol. 34. Princeton, NJ, USA: Princeton University Press, 1956, pp. 43–98.

[287] J. M. Walsh, S. Weber, and C. wa Maina, "Optimal rate-delay tradeoffs and delay mitigating codes for multipath routed and network coded networks," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5491–5510, December 2009.

[288] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, January 1960.

[289] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 100–113, January 1975.

[290] G. W. Wornell, "Inference and information," May 2017, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, Lecture Notes 6.437.

[291] A. D. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. IT-21, no. 2, pp. 163–179, March 1975.

[292] Y. Xu and T. Zhang, "Variable shortened-and-punctured Reed–Solomon codes for packet loss protection," *IEEE Transactions on Broadcasting*, vol. 48, no. 3, pp. 237–245, September 2002.

[293] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen, and G. L. Yang, Eds. New York, NY, USA: Springer, 1997, pp. 423–435.

[294] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the Davis-Kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, June 2015.

[295] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," in *Information Theory: New Trends and Open Problems*, ser. CISM International Centre for Mechanical Sciences (Courses and Lectures), G. Longo, Ed., vol. 219. New York, NY, USA: Springer, 1975, pp. 87–123.

[296] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Transactions on Information Theory*, vol. IT-19, no. 3, pp. 275–283, May 1973.