Dissipation of information in channels with input constraints

Yury Polyanskiy and Yihong Wu*

August 12, 2015

Abstract

One of the basic tenets in information theory, the data processing inequality states that output divergence does not exceed the input divergence for any channel. For channels without input constraints, various estimates on the amount of such contraction are known, Dobrushin's coefficient for the total variation being perhaps the most well-known. This work investigates channels with average input cost constraint. It is found that while the contraction coefficient typically equals one (no contraction), the information nevertheless dissipates. A certain non-linear function, the *Dobrushin curve* of the channel, is proposed to quantify the amount of dissipation. Tools for evaluating the Dobrushin curve of additive-noise channels are developed based on coupling arguments. Some basic applications in stochastic control, uniqueness of Gibbs measures and fundamental limits of noisy circuits are discussed.

As an application, it shown that in the chain of n power-constrained relays and Gaussian channels the end-to-end mutual information and maximal squared correlation decay as $\Theta(\frac{\log\log n}{\log n})$, which is in stark contrast with the exponential decay in chains of discrete channels. Similarly, the behavior of noisy circuits (composed of gates with bounded fan-in) and broadcasting of information on trees (of bounded degree) does not experience threshold behavior in the signal-to-noise ratio (SNR). Namely, unlike the case of discrete channels, the probability of bit error stays bounded away from $\frac{1}{2}$ regardless of the SNR.

Contents

1 Introduction **2**
- 1.1 Contraction properties of Markov kernels . . . 4
- 1.2 Exponential decay of information when $\eta_{\mathrm{TV}} < 1$ . . . 5
- 1.3 Organization . . . 7

2 Dobrushin curve of additive-noise channels **8**
- 2.1 Definitions and examples . . . 8
- 2.2 Criterion for $F_{\mathsf{TV}}(t) < t$ . . . 9
- 2.3 Bounds on F_{TV} via coupling . . . 12
- 2.4 By-product: CLT in smoothed total variation . . . 14

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, yp@mit.edu. Y.W. is with the Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL, yihongwu@illinois.edu. The research of Y.P. has been supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370 and by the NSF CAREER award under grant agreement CCF-12-53205. The research of Y.W. has been supported in part by NSF grants IIS-1447879 and CCF-1423088.

3 From total variation to f-divergences **15**
3.1 A parameterized family of f-divergences . . . 15
3.2 Integral representation and contraction of Rényi divergences . . . 17

4 Proof of Theorem 1 **19**
4.1 Convergence in total variation . . . 19
4.2 Special case: finite-alphabet W . . . 22
4.3 Convergence of mutual information . . . 23
4.4 Convergence of correlation coefficients . . . 26
4.5 Achievable schemes . . . 28

5 Applications **30**
5.1 Optimal memoryless control in Gaussian noise . . . 30
5.2 Uniqueness of Gibbs measures . . . 32
5.3 Circuits of noisy gates . . . 34
5.4 Broadcasting on trees . . . 37

A Convergence rate analysis **39**

B Contraction coefficient for mutual information: General case **39**

1 Introduction

Consider the following Markov chain

$$W \to X_1 \to Y_1 \to X_2 \to Y_2 \to \cdots \to X_n \to Y_n\,, \tag{1}$$

where the random variable W is the original message (which is to be estimated on the basis of Y_n only), each $P_{Y_j|X_j}$ is a standard vector-Gaussian channel of dimension d:

$$P_{Y_j|X_j=\mathbf{x}} = \mathcal{N}(\mathbf{x}, \mathbf{I}_d) \tag{2}$$

and each input X_j satisfies a power constraint:

$$\mathbb{E}[\|X_j\|^2] \le dE\,. \tag{3}$$

The goal is to design the transition kernels $P_{X_{j+1}|Y_j}$, which we refer to as processors or encoders, to facilitate the estimation of W at the end of the chain. See Fig. 1 for an illustration.

Intuitively, at each stage some information about the original message W is lost due to the external noise. Furthermore, each processor cannot de-noise completely due to the finite power constraint. Therefore it is reasonable to expect that for very large n we should have

$$P_{W,Y_n} \approx P_W P_{Y_n}\,,$$

that is, W and Y_n become almost independent. We quantify this intuition in terms of the total

variation, Kullback-Leibler (KL) divergence and correlation, namely

$$\mathsf{TV}(P,Q) \triangleq \sup_E |P[E]-Q[E]| = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|, \tag{4}$$

$$D(P\|Q) \triangleq \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P, \tag{5}$$

$$\rho(A,B) \triangleq \frac{\mathbb{E}[AB]-\mathbb{E}[A]\mathbb{E}[B]}{\sqrt{\mathrm{Var}[A]\mathrm{Var}[B]}}, \tag{6}$$

$$I(A;B) \triangleq D(P_{A,B}\|P_A P_B). \tag{7}$$

Our main result is the following theorem, which shows that the information about the original message is eventually lost in both an information-theoretic and an estimation-theoretic sense.

Theorem 1. *Let W, X_j, Y_j for a Markov chain as in (1) – (3). Then*

$$\mathsf{TV}(P_{WY_n}, P_W P_{Y_n}) \le \frac{CdE}{\log n} \to 0, \tag{8}$$

$$I(W;Y_n) \le C' d^2 E \cdot \frac{\log\log n}{\log n} \to 0, \tag{9}$$

$$\sup_{g\in L_2(P_{Y_n})} \rho(W, g(Y_n)) \to 0, \tag{10}$$

where $C, C' > 0$ are some universal constants. Moreover, the right-hand side of (10) is $O(\frac{1}{\sqrt{\log n}})$ if W is finitely valued and $O(\sqrt{\frac{\log\log n}{\log n}})$ if W is sub-Gaussian, respectively.

When W is scalar Gaussian, all estimates of the convergence rates in Theorem 1 are sharp, in the sense that there exists a sequence of power-constrained relay functions such that $\mathsf{TV}(P_{WY_n}, P_W P_{Y_n})) = \Omega(\frac{1}{\log n})$, $I(W;Y_n) = \Omega(\frac{\log\log n}{\log n})$ and $\sup_{g\in L_2(P_{Y_n})} \rho(W, g(Y_n)) = \Omega(\sqrt{\frac{\log\log n}{\log n}})$.

Our interest in the problem has been mainly motivated by the fact that the moment constraint (3) renders the standard tools for estimating convergence rates of information measures inapplicable. Thus a few new ideas are developed in this paper. In order to explain this subtlety, it is perhaps easiest to contrast Theorem 1 and especially (9) with the recent results of Subramanian et al. [Sub12, SVL13] on cascades of AWGN channels. Other applications of our techniques are deferred till Section 5.

In [Sub12, SVL13] an upper estimate on $I(W;Y_n)$ is derived under extra constraints on relay functions. Among these constraints, the most important one is that the average constraint (3) is replaced with a seemingly similar one:

$$\|X_j\|^2 \le dE \qquad \text{a.s.} \tag{11}$$

It turns out, however, that for the analysis of (11) the standard tools (in particular the Dobrushin contraction coefficient) not only recover all the results of [Sub12, SVL13] but in fact simplify and strengthen them. Thus, we start with describing those classical methods in the next section, and describe how to analyze (11) in Section 1.2 to follow.

Added in print: A completely different method (without recoursing to the total variation) for showing (9) has been developed in [PW15, CPW15] based on strong data processing inequalities for mutual information in Gaussian noise.

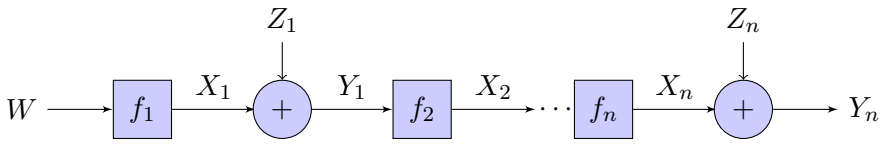


Figure 1: Cascade of AWGN channels with power-constrained relays f_j.

1.1 Contraction properties of Markov kernels

Fix a transition probability kernel (channel) $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ acting between two measurable spaces. We denote by $P_{Y|X} \circ P$ the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P, which is the distribution of the output Y when the input X is distributed according to P, and by $P \times P_{Y|X}$ the joint distribution P_{XY} if $P_X = P$. We also denote by $P_{Z|Y} \circ P_{Y|X}$ the serial composition of channels. Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a convex function with $f(1) = 0$ and let $D_f(P||Q) \triangleq \mathbb{E}_Q[f(\frac{\mathrm{d}P}{\mathrm{d}Q})]$ denote the corresponding f-divergence, cf. [Csi67]. For example taking $f(x) = (x-1)^2$ we obtain the χ^2-divergence:

$$\chi^2(P\|Q) \triangleq \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q - 1\,. \tag{12}$$

For any Q that is not a point mass, define:

$$\eta_f(Q) \triangleq \sup_{P: 0 < D_f(P\|Q) < \infty} \frac{D_f(P_{Y|X} \circ P \| P_{Y|X} \circ Q)}{D_f(P\|Q)}, \tag{13}$$

$$\eta_f \triangleq \sup_Q \eta_f(Q)\,. \tag{14}$$

For $f(x) = |x-1|$, $f(x) = (x-1)^2$ and $f(x) = x \log x$ we will write $\eta_{\mathrm{TV}}(Q), \eta_{\chi^2}(Q)$ and $\eta_{\mathrm{KL}}(Q)$, respectively. In particular, η_{TV} is known as the *Dobrushin's coefficient* of the kernel $P_{Y|X}$, which is one of the main tools for studying ergodicity property of Markov chains as well as Gibbs measures.

General alphabets Dobrushin [Dob56] showed that supremum in the definition of η_{TV} can be restricted to single-point distributions P and Q, thus providing a simple criterion for strong ergodicity of Markov processes. It is well-known, e.g. Sarmanov [Sar58], that $\eta_{\chi^2}(Q)$ is the squared *maximal correlation* coefficient of the joint distribution $P_{XY} = Q \times P_{Y|X}$:

$$S(X;Y) \triangleq \sup_{f,g} \rho(f(X), g(Y)) = \sqrt{\eta_{\chi^2}(Q)}\,. \tag{15}$$

Later [CKZ98, Proposition II.4.10] (see also [CIR+93, Theorem 4.1] for finite alphabets) demonstrated that all other contraction coefficients are upper-bounded by the Dobrushin's coefficient η_{TV}:

$$\eta_f \le \eta_{\mathrm{TV}}, \tag{16}$$

and this inequality is typically strict.[1] In the opposite direction it can be shown, cf. [CKZ98, Proposition II.6.15],

$$\eta_{\chi^2} \le \eta_f\,, \tag{17}$$

[1] E.g. for the binary symmetric channel with crossover probability δ we have $\eta_{\chi^2} = \eta_{\mathrm{KL}} = (1-2\delta)^2 < \eta_{\mathrm{TV}} = |1-2\delta|$.

whenever f is thrice differentiable with $f''(1) > 0$. Moreover, (17) holds with equality for all nonlinear and operator convex f, e.g., for KL divergence and for squared Hellinger distance; see [CRS94, Theorem 1] and [CKZ98, Proposition II.6.13 and Corollary II.6.16]. In particular,

$$\eta_{\chi^2} = \eta_{\mathrm{KL}}, \tag{18}$$

which was first obtained in [AG76] using different methods. Rather naturally, we also have [CKZ98, Proposition II.4.12]:

$$\eta_f = 1 \quad \Longleftrightarrow \quad \eta_{\mathrm{TV}} = 1$$

for any non-linear f.

The fixed-input contraction coefficient $\eta_{\mathrm{KL}}(Q)$ is closely related to the (modified) log-Sobolev inequalities. Indeed, when Q is invariant under $P_{Y|X}$ (i.e. $P_{Y|X} \circ Q = Q$) any initial distribution P converges to Q exponentially fast in terms of $D(P^n_{Y|X} \circ P||Q)$ with exponent upper-bounded by $\eta_{\mathrm{KL}}(Q)$, which in turn can be estimated from log-Sobolev inequalities, e.g. [Led99]. When Q is not invariant, it was shown [DMLM03] that

$$1 - \alpha(Q) \le \eta_{\mathrm{KL}}(Q) \le 1 - C\alpha(Q) \tag{19}$$

holds for some universal constant C, where $\alpha(Q)$ is a modified log-Sobolev (also known as 1-log-Sobolev) constant:

$$\alpha(Q) = \inf_{f \perp 1, \|f\|_2 = 1} \frac{\mathbb{E}\left[f^2(X) \log \frac{f^2(X)}{f^2(X')}\right]}{\mathbb{E}[f^2(X) \log f^2(X)]}, \qquad P_{XX'} = Q \times (P_{X|Y} \circ P_{Y|X}).$$

Finite alphabets Ahlswede and Gács [AG76] have shown

$$\eta_{\chi^2}(Q) < 1 \iff \eta_{\mathrm{KL}}(Q) < 1 \iff \text{graph } \{(x,y) : Q(x) > 0, P_{Y|X}(y|x) > 0\} \text{ is connected.}$$

As a criterion for $\eta_f(Q) < 1$, this is an improvement of (16) only for channels with $\eta_{\mathrm{TV}} = 1$. Furthermore, [AG76] shows

$$\eta_{\chi^2}(Q) \le \eta_{\mathrm{KL}}(Q), \tag{20}$$

with inequality frequently being strict.[2] We note that the main result of [AG76] characterizes $\eta_{\mathrm{KL}}(Q)$ as the maximal ratio of hyper-contractivity of the conditional expectation operator $\mathbb{E}[\cdot|X]$. For finite alphabets, (17) can be strengthened to the following fixed-input version under the same conditions on f (c.f. [Rag14, Theorem 3.3]):

$$\eta_{\chi^2}(Q) \le \eta_f(Q)\,. \tag{21}$$

For connections between η_{KL} and log-Sobolev inequalities on finite alphabets see [Rag13].

1.2 Exponential decay of information when $\eta_{\mathrm{TV}} < 1$

First, it can be shown that (See Appendix B for a proof in the general case. The finite alphabet case has been shown in [AGKN13])

$$\sup \frac{I(U;Y)}{I(U;X)} = \eta_{\mathrm{KL}}(P_X)\,, \tag{22}$$

[2] See [AG76, Theorem 9] and [AGKN13] for examples.

where the supremum is taken over all Markov chains $U \to X \to Y$ with fixed P_{XY} such that $0 < I(U;X) < \infty$. Thus, for an arbitrary Markov chain

$$W \to X_1 \to Y_1 \to X_2 \to Y_2 \to \cdots \to Y_n$$

with equal channels $P_{Y_j|X_j} = P_{Y|X}$ for all j, we have

$$I(W;Y_n) \le \prod_{j=1}^{n} \eta_{\mathrm{KL}}(P_{X_j}) \cdot I(W;X_1) \le (\eta_{\mathrm{KL}})^n \cdot H(W)\,. \tag{23}$$

A similar argument leads to

$$\mathsf{TV}(P_{WY_n}, P_W P_{Y_n}) \le \prod_{j=1}^{n} \eta_{\mathrm{TV}}(P_{X_j}) \le (\eta_{\mathrm{TV}})^n, \tag{24}$$

$$\rho^2(W;Y_n) \le S(W;Y_n) \le \prod_{j=1}^{n} S(X_j;Y_j) \le (\eta_{\chi^2})^n\,. \tag{25}$$

Thus, in the simple case when $\eta_{\mathrm{TV}} < 1$ we have from (16) that when $n \to \infty$, all three information quantities converge to zero exponentially as fast as η_{TV}^n.

Let us now consider the case of [Sub12,SVL13], namely the AWGN channel $P_{Y|X}$ with maximal power constraint (11). First recall that

$$\mathsf{TV}(\mathcal{N}(\mu_1, \mathbf{I}_d), \mathcal{N}(\mu_2, \mathbf{I}_d)) = 1 - 2\mathsf{Q}(|\mu_1 - \mu_2|/2)\,, \tag{26}$$

where $\mathsf{Q}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ is the Gaussian complimentary CDF and $|\cdot|$ denotes the Euclidean norm. Then by Dobrushin's characterization of η_{TV} we get that for any P_{X_j} satisfying (11) we have

$$\eta_{\mathrm{TV}} = \sup_{x_1, x_2} 1 - 2\mathsf{Q}(|x_1 - x_2|/2) = 1 - 2\mathsf{Q}(\sqrt{dE})\,.$$

From (23) this implies

$$I(W;Y_n) \le (1 - 2\mathsf{Q}(\sqrt{dE}))^n \cdot H(W)\,. \tag{27}$$

It turns out (27) is stronger than the main result of [SVL13] and independent of the cardinality of W. Indeed, although [SVL13] did not point this out, the analysis there corresponds to the following upper-bound on η_{TV}

$$\eta_{\mathrm{TV}} \le 1 - \sum_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} P_{Y|X}(y|x) \tag{28}$$

(here we assumed finite alphabet $\mathcal{Y}$ for simplicity). This bound is clearly tight for the case of $|\mathcal{X}| = 2$ but rather loose for larger $|\mathcal{X}|$. Since we calculated η_{TV} exactly, (27) must yield a better bound than that of [SVL13]. However, the estimate (27) relies on the Dobrushin coefficient, which, as will be shown below, breaks down if the power constraints is imposed on average instead of almost surely. To remedy this problem requires developing new tools to complement the Dobrushin coefficient. For the generalization to average power constraint as well as discussions for multi-hop communication, see Proposition 13 and Remark 9 in Section 4.3.

The main part of this paper handles convergence of $I(W;Y_n) \to 0$ in the case (3), for which unfortunately $\eta_{\mathrm{TV}} = \eta_{\mathrm{KL}} = \eta_{\chi^2} = 1$. Indeed, by taking

$$P = (1-t)\delta_0 + t\delta_a\,, \tag{29}$$

$$Q = (1-t)\delta_0 + t\delta_{-a}\,, \tag{30}$$

and performing a straightforward calculation, we find

$$\frac{\mathsf{TV}(P * \mathcal{N}(0,1), Q * \mathcal{N}(0,1))}{\mathsf{TV}(P,Q)} = 1 - 2\mathsf{Q}(a) \xrightarrow{a\to\infty} 1. \tag{31}$$

Therefore, even if one restricts the supremum in (14) to P and Q satisfying the moment constraint (3) (in fact, any constraint on the tails for that matter), choosing $a \to \infty$ and $t \to 0$ accordingly drives the ratio in (31) to one, thus proving $\eta_{\mathrm{TV}} = 1$. This example is instructive: The ratio (31) approaches 1 only when the $\mathsf{TV}(P,Q) \to 0$. Our idea is to get *non-multiplicative* contraction inequalities that still guarantee strict decrease of total variation after convolution.

Similarly, there is no moment condition which can guarantee the strict contraction of the KL divergence or mutual information. For example, it can be shown that

$$\sup \frac{I(U; X+Z)}{I(U;X)} = 1\,,$$

where the supremum is over all Markov chains $U \to X \to X+Z$ with $\mathbb{E}[|X|^2] \le 1$. This suggests that the exponential decay of mutual information in (23) obtained under peak power constraint might fail. Indeed, we will show that under average power constraint, the decay speed of mutual information can be much slower than exponential (see Section 4.5).

1.3 Organization

The rest of the paper is organized as follows. Section 2 proves results on reduction of total variation over additive-noise channels; we call the resulting relation the *Dobrushin curve* of a channel. Section 3 shows how to convert knowledge about total variation to other f-divergences, extending (16). Section 4 shows how to use Dobrushin curve to prove Theorem 1. Finally, Section 5 concludes with applications (other than Theorem 1).

In particular, in Section 5.1 we show that the optimal correlation achieved by non-linear control in the n-stage Gaussian quadratic control problem studied by Lipsa and Martins [LM11] is $\Theta(\sqrt{\frac{\log\log n}{\log n}})$; in contrast, the best linear controller only achieves exponentially small correlation. The inferiority of linear control can be explained from the viewpoint of dissipation of information and contraction of KL divergence. In Section 5.2 we extend Dobrushin's strategy for proving uniqueness of Gibbs measures to unbounded systems with moment constraints on marginal distributions. And in Section 5.3 we apply our technique to proving a lower bound on the probability of error in circuits of noisy gates.

Finally, in Section 5.4 we show that in the question of broadcasting a single bit on a tree of Gaussian channels there is no phase transition. Namely, for arbitrarily low SNR it is possible to build relays satisfying the average power constraint so that given the received values on all leaves at depth d the probability of error of estimating the original bit is bounded away from $1/2$. This is in contrast to the case of trees of binary symmetric channels, studied by Evans-Kenyon-Peres-Schulman [EKPS00], who showed that there there is a phase transition in terms of the strength of the channel noise.

2 Dobrushin curve of additive-noise channels

2.1 Definitions and examples

Let $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ be a probability transition kernel. Then, we define the *Dobrushin curve* of $P_{Y|X}$ as follows:

$$F_{\mathsf{TV}}(t) = \sup\{\mathsf{TV}(P_{Y|X} \circ P, P_{Y|X} \circ Q) : \mathsf{TV}(P, Q) \le t, (P, Q) \in \mathcal{G}\}, \qquad t \in [0, 1] \tag{32}$$

where $\mathcal{G}$ is some (convex) set of pairs of probability measures. The curve $t \mapsto F_{\mathsf{TV}}(t)$ defines the upper boundary of the region

$$\mathcal{F}_{\mathsf{TV}} = \left\{\left(\mathsf{TV}(P_{Y|X} \circ P, P_{Y|X} \circ Q), \mathsf{TV}(P, Q)\right) : (P, Q) \in \mathcal{G}\right\} \subset [0, 1]^2, \tag{33}$$

which is the joint range of the input and output total variations.

We notice the following "data-processing" property of Dobrushin curves: if $F_{\mathsf{TV}1}$ and $F_{\mathsf{TV}2}$ are the Dobrushin curves of channels $P_{Y_1|X_1}$ and $P_{Y_2|X_2}$ (and the respective feasible sets $\mathcal{G}_1$ and $\mathcal{G}_2$), then for any $P_{X_2|Y_1}$ that connects them:

$$X_1 \overset{P_{Y_1|X_1}}{\longrightarrow} Y_1 \longrightarrow X_2 \overset{P_{Y_2|X_2}}{\longrightarrow} Y_2$$

we naturally have for the combined channel

$$F_{\mathsf{TV}}(t) \le F_{\mathsf{TV}2}(F_{\mathsf{TV}1}(t))$$

(the constraint set $\mathcal{G}$ corresponding to $F_{\mathsf{TV}}(t)$ is defined so that $(P, Q) \in \mathcal{G}_1$ and $(P_{X_2|Y_1} \circ P_{Y_1|X_1} \circ P, P_{X_2|Y_1} \circ P_{Y_1|X_1} \circ Q) \in \mathcal{G}_2$). This observation will be central for the analysis of the Markov chain (1). We proceed to computing F_{TV}.

For simplicity, in the sequel we focus our presentation on the following:

1. Consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ with Borel σ-algebra and $d \in \mathbb{N} \cup \{+\infty\}$.
2. There is a norm $|\cdot|$ on $\mathbb{R}^d$.
3. The constraint set $\mathcal{G}$ is defined by some average cost constraint:

$$\mathcal{G}_a \triangleq \{(P, Q) : \mathbb{E}_P[\mathsf{M}(|X|)] + \mathbb{E}_Q[\mathsf{M}(|X|)] \le 2a\}, \tag{34}$$

   where $\mathsf{M} : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing convex cost function[3] with $\mathsf{M}(0) = 0$ and $a \ge 0$.
4. The random transformation $P_{Y|X}$ acts by convolution (on $\mathbb{R}^d$) with noise P_Z:

$$P_{Y|X=x} = P_{Z+x} \qquad x, Y, Z \in \mathbb{R}^d.$$

Remark 1. For any point $(\mathsf{TV}(P, Q), \mathsf{TV}(P * P_Z, Q * P_Z))$ in the region $\mathcal{F}_{\mathsf{TV}}$ and $\lambda \in [0, 1]$, we can achieve the point $(\lambda\mathsf{TV}(P, Q), \lambda\mathsf{TV}(P * P_Z, Q * P_Z))$ by setting $P_\lambda = \lambda P + (1 - \lambda)\delta_0$ and $Q_\lambda = \lambda Q + (1 - \lambda)\delta_0$. This implies that $t \mapsto \frac{F_{\mathsf{TV}}(t)}{t}$ is non-increasing. However, this does not imply that $\mathcal{F}_{\mathsf{TV}}$ is convex or that F_{TV} is concave. Shortly, we will demonstrate that for many noise distribution P_Z the Dobrushin curve F_{TV} is in fact concave.

[3] Our motivating examples are $\mathsf{M}(x) = x^p$ with $p \ge 1$, $\mathsf{M}(x) = \exp(\alpha x) - 1$ and $\mathsf{M}(x) = \exp(\alpha x^2) - 1$ with $\alpha > 0$, which we call p^{th}-moment, sub-exponential and sub-Gaussian constraints, respectively.

Expanding on the previous remark, we can further show relations between F_{TV} computed for different cost values of a in (34).

Proposition 2. *Let $F_{\mathsf{TV}}(t,a)$ be the Dobrushin curve for some channel $P_{Y|X}$ and constraint (34), where $\mathsf{M}(0)=0$. Then for all $\alpha \geq 0$ such that $\alpha t \leq 1$ we have*

$$F_{\mathsf{TV}}(\alpha t, \alpha a) = \alpha F_{\mathsf{TV}}(t,a)\,. \tag{35}$$

In particular, $F_{\mathsf{TV}}(t,a) = tg(a/t)$, where $g(a) \triangleq F_{\mathsf{TV}}(1,a)$ and in the unconstrained case Dobrushin curve is a straight line: $F_{\mathsf{TV}}(t,\infty) = \eta_{\mathrm{TV}} t$.

Proof. Without loss of generality, we may assume $\alpha \leq 1$ (otherwise, apply to $t' = \alpha t$ and $\alpha' = 1/\alpha$). For all $s \in [0,1]$ we have two inequalities

$$F_{\mathsf{TV}}(st, sa) \geq sF_{\mathsf{TV}}(t,a), \tag{36}$$
$$F_{\mathsf{TV}}(s, sa/t) \geq s/tF_{\mathsf{TV}}(t,a). \tag{37}$$

To show the first start with arbitrary $(P,Q) \in \mathcal{G}_a$ such that $\mathsf{TV}(P,Q) = t$ and $\mathsf{TV}(P_{Y|X}\circ P, P_{Y|X}\circ Q) = f$. Then we can construct distributions

$$P_1 = sP + (1-s)\delta_0\,, \quad Q_1 = sQ + (1-s)\delta_0\,,$$

for which

$$\int \mathsf{M}(x)(dP_1 + dQ_1) \leq sa, \quad \mathsf{TV}(P_1,Q_1) = st, \quad \mathsf{TV}(P_{Y|X}\circ P_1, P_{Y|X}\circ Q_1) = sf$$

and thus (36) follows after optimizing over (P,Q). The second inequality follows by considering[4]

$$P_2 = \frac{s}{t}(P-Q)^+ + (1-s)\delta_0, \quad Q_2 = \frac{s}{t}(P-Q)^- + (1-s)\delta_0$$

and a similar argument. Finally, (35) follows from (36) (with $s=\alpha$) and (37) (with $s = t/\alpha$). □

2.2 Criterion for $F_{\mathsf{TV}}(t) < t$

Similar to how Dobrushin's results [Dob56] reduce the computation of η_{TV} to considering the two-point quantity $\mathsf{TV}(P_{Y|X=x}, P_{Y|X=x'})$, our main tool will be the following function $\theta : \mathbb{R}^d \to [0,1]$ defined by

$$\theta(x) \triangleq \mathsf{TV}(P_Z, P_{Z+x})\,, \qquad x \in \mathbb{R}^d. \tag{38}$$

Some simple properties of θ (general case) are as follows:

- $\theta(0) = 0$, $\lim_{x\to\infty}\theta(x) = 1$.
- $\theta(x) = \theta(-x)$.
- If P_Z is compactly supported then $\theta(x) = 1$ when $|x|$ is sufficiently large.
- θ is lower-semicontinuous (since total variation is weakly lower-semicontinuous).

[4] Measures $(P-Q)^+$ and $(P-Q)^-$ denote the two pieces of Jordan decomposition of measure $(P-Q)$.

- If P_Z has a density f_Z, then

$$\theta(x) = \int_{\mathbb{R}^d} |f_Z(z-x) - f_Z(z)| \mathrm{d}z.$$

  and θ is continuous on $\mathbb{R}$, which follows from the denseness of compactly-supported continuous functions in $L_1(\mathbb{R}^d)$.

Further properties of θ in dimension $d = 1$ include:

- θ is continuous at 0 if and only if Z has a density with respect to the Lebesgue measure. To see this, decompose $P_Z = \mu_a + \mu_s$ into absolutely continuous and singular parts (with respect to the Lebesgue measure). By [Pro04, Theorem 10], $\liminf_{h\to 0} \mathsf{TV}(P_Z, P_{Z+h}) = 0$ if and only if P_Z is absolutely continuous. By the previous remark we have

$$\limsup_{x\to 0} \theta(x) = \mu_s(\mathbb{R}).$$

- If P_Z has a non-increasing density supported on $\mathbb{R}_+$, then $\theta(x)$ is a concave, non-decreasing function on $\mathbb{R}_+$ given by

$$\theta(x) = \mathbb{P}\left[Z \le x\right], \qquad x \ge 0\,. \tag{39}$$

- If P_Z has a symmetric density which is non-increasing on $\mathbb{R}_+$, then $\theta(x)$ is a concave, non-decreasing function on $\mathbb{R}_+$ given by

$$\theta(x) = \mathbb{P}\left[|Z| \le x/2\right], \qquad x \ge 0 \tag{40}$$

- In general, θ need not be monotonic on $\mathbb{R}^+$ (e.g. P_Z is discrete or has a multimodal density such as a Gaussian mixture).

The following result gives a necessary and sufficient condition for the total variation to strictly contract on an additive-noise channel, which essentially means that the noise distribution is almost mutually singular to a translate of itself. Intuitively, it means that if the noise is too weak (e.g., when the noise has a compact support or has a singular distribution), then one can send one bit error-free if the signal magnitude is sufficiently large.

Theorem 3. *Define*

$$\eta(A) = \sup_{x:|x|\le A} \theta(x).$$

The following are equivalent

1. $\eta(A) = 1$ *for some* $A > 0$.
2. $F_{\mathsf{TV}}(t) = t$ *in some neighborhood of* 0.
3. $F_{\mathsf{TV}}(t) = t$ *for some* $t > 0$.

Remark 2. It is possible to have $\eta(A) = 1$ with $\theta(x) < 1$ on $[-A, A]$. For example, let

$$P_Z = \frac{1}{2}\sum_{k\ge 1} 2^{-k}\delta_k + \frac{1}{2}\sum_{k\ge 1} 2^{-k} U(2k-1, 2k).$$

where $U(a, b)$ denotes the uniform distribution on (a, b).

Proof. The equivalence of 2 and 3 follows from Remark 1.

For $1 \Rightarrow 2$, choosing $P = (1-t)\delta_0 + t\delta_x$ and $Q = \delta_0$, we have $\mathsf{TV}(P * P_Z, Q * P_Z) = t\theta(x)$. Optimizing over $x \in [-A, A]$ yields $F_{\mathsf{TV}}(t) = t$, provided that $t\mathsf{M}(A) \le a$.

Before proceeding further, we notice that for any channel $P_{Y|X}$ with Dobrushin coefficient η_{TV} and any measure ν on $\mathcal{X}$ such that $\int \mathrm{d}\nu = 0$ we have

$$\mathsf{TV}(P_{Y|X} \circ \nu, 0) \le \eta_{\mathrm{TV}} \mathsf{TV}(\nu, 0)\,,$$

where here and below the total variation distance defined in (4) naturally extended to non-probability measures as follows:

$$\mathsf{TV}(\nu, \mu) = \frac{1}{2}\int |\mathrm{d}\nu - d\mu|\,.$$

Next, by representing $\nu = \nu^+ - \nu^-$ and playing with scaling ν^+ or ν^- we get the result of [CIR+93, Lemma 3.2]:

$$\mathsf{TV}(P_{Y|X} \circ \nu, 0) \le \eta_{\mathrm{TV}} \mathsf{TV}(\nu, 0) + \frac{1-\eta_{\mathrm{TV}}}{2}\left|\int \mathrm{d}\nu\right|$$

Now we prove $3 \Rightarrow 1$. Fix arbitrary $(P, Q) \in \mathcal{G}_a$ and choose large $A > 0$. Let P_1, Q_1 be restrictions of P and Q to the closed ball

$$B \triangleq \{x : |x| \le A\}$$

and $P_2 = P - P_1, Q_2 = Q - Q_1$. By [CIR+93, Lemma 3.2] we have then:

$$\mathsf{TV}(P_1 * P_Z, Q_1 * P_Z) \le \eta \mathsf{TV}(P_1, Q_1) + \frac{1-\eta}{2}\,|P(B) - Q(B)|\,, \qquad \eta \triangleq \eta(A)\,.$$

Since $(P, Q) \in \mathcal{G}_a$, applying Markov's inequality yields $P(B^c) + Q(B^c) \le \frac{2a}{\mathsf{M}(A)}$ and thus

$$\mathsf{TV}(P_2, Q_2) \le \frac{a}{\mathsf{M}(A)}.$$

Also, since $P(\mathcal{X}) - Q(\mathcal{X}) = 0$, we have

$$|P(B) - Q(B)| = |P(B^c) - Q(B^c)| \le \frac{2a}{\mathsf{M}(A)}.$$

Putting it all together and using triangle inequality, we have

$$\begin{aligned}
\mathsf{TV}(P * P_Z, Q * P_Z) &\le \mathsf{TV}(P_1 * P_Z, Q_1 * P_Z) + \mathsf{TV}(P_2 * P_Z, Q_2 * P_Z) \\
&\le \mathsf{TV}(P_1 * P_Z, Q_1 * P_Z) + \mathsf{TV}(P_2, Q_2) \\
&\le \eta \mathsf{TV}(P_1, Q_1) + \frac{1-\eta}{2}\,|P(B) - Q(B)| + \mathsf{TV}(P_2, Q_2) \\
&= \eta \mathsf{TV}(P, Q) + (1-\eta)\left(\frac{1}{2}\,|P(B) - Q(B)| + \mathsf{TV}(P_2, Q_2)\right) \\
&\le \eta \mathsf{TV}(P, Q) + (1-\eta)\frac{2a}{\mathsf{M}(A)}\,,
\end{aligned}$$

where the equality step follows from the crucial fact that $\mathsf{TV}(P, Q) = \mathsf{TV}(P_1, Q_1) + \mathsf{TV}(P_2, Q_2)$, due to the disjointedness of supports.

By the arbitrariness of (P, Q), we have shown that for every $A > 0$ and t,

$$F_{\mathsf{TV}}(t) \le \eta(A)t + (1 - \eta(A))\frac{2a}{\mathsf{M}(A)}\,.$$

Thus if $F_{\mathsf{TV}}(t) = t$ for some $t > 0$, then $(1-\eta(A))t \le (1-\eta(A))\frac{2a}{\mathsf{M}(A)}$ for all $A > 0$. Therefore we must have $\eta(A) = 1$ whenever $\mathsf{M}(A) > \frac{2a}{t}$. □

2.3 Bounds on F_{TV} via coupling

Theorem 4. *Define $\theta_{lb}(s) = \sup_{x:|x|\le s} \theta(x)$ and let $\theta_c : \mathbb{R}_+ \to [0,1]$ be the concave envelope (i.e., the smallest concave majorant) of θ_{lb} on $\mathbb{R}_+$, Then*

$$t\theta_{lb}\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right) \le F_{\mathsf{TV}}(t) \le t\theta_c\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right) \tag{41}$$

Remark 3. Note that for the upper bound (41) to be non-trivial, i.e., better than $F_{\mathsf{TV}}(t) \le t$, for all $t > 0$, it is necessary and sufficient to have $\theta_c(|x|) < 1$ for all x. This is consistent with Theorem 3.

Proof. Recall that $\mathsf{TV}(P_{Z+a}, P_{Z+b}) = \theta(a-b)$, by definition of the function θ in (38). Fix any $(P,Q) \in \mathcal{G}_a$. The map $(P,Q) \mapsto \mathsf{TV}(P,Q)$ is convex (as is any Wasserstein distance), thus for any coupling P_{AB} with $P_A = P$ and $P_B = Q$ we have

$$\mathsf{TV}(P * P_Z, Q * P_Z) \le \mathbb{E}[\theta(A-B)] \tag{42}$$

Furthermore, θ_c is necessarily continuous on $(0,\infty)$, strictly increasing on $\{x : \theta_c(x) < 1\}$ and concave. Thus,

$$\begin{aligned}
\mathbb{E}[\theta(|A-B|)] &= \mathbb{P}[A \ne B]\mathbb{E}[\theta(A-B) \,|\, A \ne B] && (43)\\
&\le \mathbb{P}[A \ne B]\mathbb{E}[\theta_c(|A-B|) \,|\, A \ne B] && (44)\\
&\le \mathbb{P}[A \ne B]\theta_c\left(\mathbb{E}[|A-B| \,|\, A \ne B]\right) && (45)
\end{aligned}$$

where (45) is by Jensen's inequality and the concavity of θ_c. Then

$$\begin{aligned}
\mathsf{M}\left(\frac{\mathbb{E}[|A-B||A \ne B]}{2}\right) &\le \mathbb{E}\left[\mathsf{M}\left(\frac{|A-B|}{2}\right)\Big| A \ne B\right] && (46)\\
&= \frac{1}{\mathbb{P}\left[A \ne B\right]}\mathbb{E}\left[\mathsf{M}\left(\frac{|A-B|}{2}\right)\right] && (47)\\
&\le \frac{1}{\mathbb{P}\left[A \ne B\right]}\mathbb{E}\left[\mathsf{M}\left(\frac{|A|+|B|}{2}\right)\right] && (48)\\
&\le \frac{\mathbb{E}[\mathsf{M}(|A|)] + \mathbb{E}[\mathsf{M}(|B|)]}{2\mathbb{P}\left[A \ne B\right]} && (49)\\
&\le \frac{a}{\mathbb{P}\left[A \ne B\right]} && (50)
\end{aligned}$$

where (46) and (49) are by Jensen's inequality and the convexity of M, (47) is by $\mathsf{M}(0) = 0$, (48) is by the monotonicity of M, and (50) is by the constraint $(P,Q) \in \mathcal{G}_a$. Applying M^{-1} to both sides of (50) and plugging into (45), we obtain

$$\mathbb{E}[\theta(|A-B|)] \le \mathbb{P}[A \ne B]\theta_c\left(2\mathsf{M}^{-1}\left(\frac{a}{\mathbb{P}\left[A \ne B\right]}\right)\right). \tag{51}$$

Note that both M^{-1} and θ_c are increasing concave functions. Thus their composition $\theta_c \circ 2\mathsf{M}^{-1}$ is concave and increasing too. Furthermore it is easy to show that

$$t \mapsto t\theta_c\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right) \tag{52}$$

is increasing. Hence the upper bound (51) is tightest for the coupling minimizing $\mathbb{P}[A \neq B]$. Recall that by Strassen's characterization [Str65] we have

$$\inf_{P_{AB}} \mathbb{P}[A \neq B] = \mathsf{TV}(P, Q), \tag{53}$$

where the infimum is over all couplings P_{AB} of P and Q such that $P_A = P$ and $P_B = Q$. Then (42) and (51) and the continuity of θ_c imply the upper bound in (41).

For the lower bound, we choose

$$P = (1-t)\delta_0 + t\delta_x \tag{54}$$
$$Q = (1-t)\delta_0 + t\delta_{-x} \tag{55}$$

with $|x| \leq \mathsf{M}^{-1}(a/t)$, which ensures that $(P, Q) \in \mathcal{G}_a$. It is straightforward to show that $\mathsf{TV}(P, Q) = t$ and $\mathsf{TV}(P * P_Z, Q * P_Z) = t\theta(x)$. Taking the supremum over x yields the left inequality of (41). □

Corollary 5. *If the dimension $d = 1$ and θ_{lb} is concave on $\mathbb{R}_+$ then*

$$F_{\mathsf{TV}}(t) = t\theta_{lb}\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right). \tag{56}$$

Remark 4. Examples of the noise distributions satisfying assumptions of Corollary 5 are given by (39) and (40). Note that from concavity of θ the map

$$u \mapsto \theta_{lb}(2\mathsf{M}^{-1}(u)) \tag{57}$$

is also concave. Therefore, the map

$$(a, t) \mapsto t\,\theta_{lb}\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right)$$

is the *perspective* of the concave function (57), and hence is concave on $\mathbb{R}_+^2$ [HUL96, p. 161]. Consequently, for fixed $a > 0$, F_{TV} is concave, which, as we mentioned, does not immediately follow from the definition of F_{TV}.

For the purpose of showing Theorem 1 we next point out the particularization of Corollary 5 to the AWGN channel. A representative plot of the F_{TV} for the AWGN channel and average power constraint (second-order moment) is given in Fig. 2, which turns out to be dimension-independent.

Corollary 6 (Vector Gaussian)**.** *Let $d \in \mathbb{N} \cup \{\infty\}$, $P_Z = \mathcal{N}(0, \mathbf{I}_d)$ and $|x| = (\sum_{i=1}^d x_i^2)^{1/2}$ be the Euclidean norm. Then*

$$F_{\mathsf{TV}}(t) = t\left(1 - 2\mathsf{Q}\left(\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right)\right). \tag{58}$$

Proof. From (26) we have that $\theta_{lb}(u) = \theta_c(u) = 1 - 2\mathsf{Q}(u/2)$ regardless of dimension and thus Theorem 4 yields (58). □

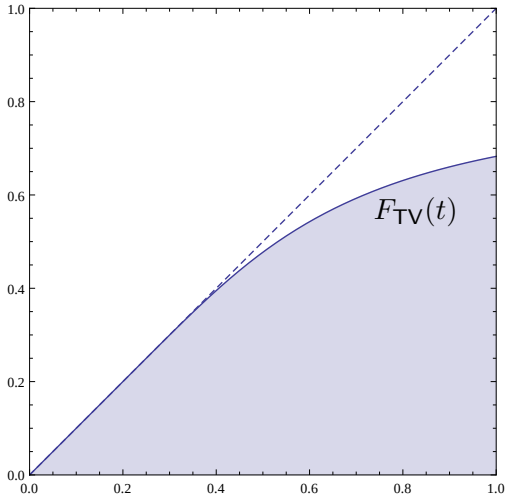


Figure 2: The region $\{(\mathsf{TV}(P,Q), \mathsf{TV}(P * P_Z, Q * P_Z)) : (P,Q) \in \mathcal{G}_1\}$ in the Gaussian case $Z \sim \mathcal{N}(0,1)$ with $\mathsf{M}(x) = |x|^2$.

2.4 By-product: CLT in smoothed total variation

Recall the following 1-Wasserstein distance between distributions with finite first moment:

$$W_1(P,Q) = \inf_{P_{AB}} \{\|A - B\|_1 : P_A = P, P_B = Q\}. \tag{59}$$

Then the same coupling method in the proof of Theorem 4 yields the following bound, which relates the total variation between convolutions to the W_1 distance.

Proposition 7. *If P_Z has a symmetric density which is non-increasing on $\mathbb{R}_+$. Then for any P and Q,*

$$\mathsf{TV}(P * P_Z, Q * P_Z) \le \mathbb{P}\left[|Z| \le \frac{W_1(P,Q)}{2}\right]. \tag{60}$$

Proof. By (40), the function $\theta(x) = \mathbb{P}[|Z| \le x/2]$ is concave and non-decreasing in x. Applying Jensen's inequality to (42) and optimizing over the coupling yields (60). □

Remark 5. It is worth mentioning that for Gaussian smoothing, using similar coupling and convexity arguments, the following counterpart of (60) for KL divergence has been proved in [Wu11], which provides a simple proof of Otto-Villani's HWI inequality [OV00] in the Gaussian case:

$$D(P * \mathcal{N}(0,\sigma^2) \| Q * \mathcal{N}(0,\sigma^2)) \le \frac{W_2^2(P,Q)}{2\sigma^2},$$

where the W_2 distance is analogously defined as (59) with L_2-norm replacing L_1-norm.

In particular, if P_Z has a bounded density near zero, then the right-hand side of (60) is $O(W_1(P,Q))$. As an application, we consider a central limit theorem setting and let

$$S_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j,$$

where X_j are iid, zero-mean and unit-variance. Choosing $P_Z = \mathcal{N}(0, \sigma^2)$ and applying Proposition 7 to P_{S_n} and $\mathcal{N}(0,1)$, we obtain

$$\mathsf{TV}(P_{S_n} * \mathcal{N}(0,\sigma^2), \mathcal{N}(0, 1+\sigma^2)) \le \frac{W_1(P_{S_n}, \mathcal{N}(0,1))}{\sqrt{2\pi\sigma^2}} \le \frac{3\,\mathbb{E}[|X_1|^3]}{\sqrt{2\pi\sigma^2 n}} \tag{61}$$

where the convergence rate in W_1 can be obtained from Stein's method and the dual representation of $W_1(P,Q) = \sup\{\int f \mathrm{d}P - \int f \mathrm{d}Q : f \text{ is 1-Lipschitz}\}$ (see, e.g., [BC05, Theorem 3.2]). In other words, smoothing the law of S_n by convolving with a Gaussian density (or any other bounded density that satisfies the conditions of (40)) results in a distribution that is closer in total variation to the Gaussian distribution. On the other hand, the law of S_n might never converge to Gaussian (e.g., for discrete X_1).

The non-asymptotic estimate (61) should be contrasted with the sharp asymptotics of total variation in CLT due to Sirazhdinov and Mamatov [SM62], which states that the left-hand side of (61) is equal to $\frac{(1+4\mathrm{e}^{-3/2})\mathbb{E}[X_1^3]}{6\sqrt{2\pi n(1+\sigma^2)^3}}(1+o(1))$ when $n \to \infty$ and σ is fixed.

3 From total variation to f-divergences

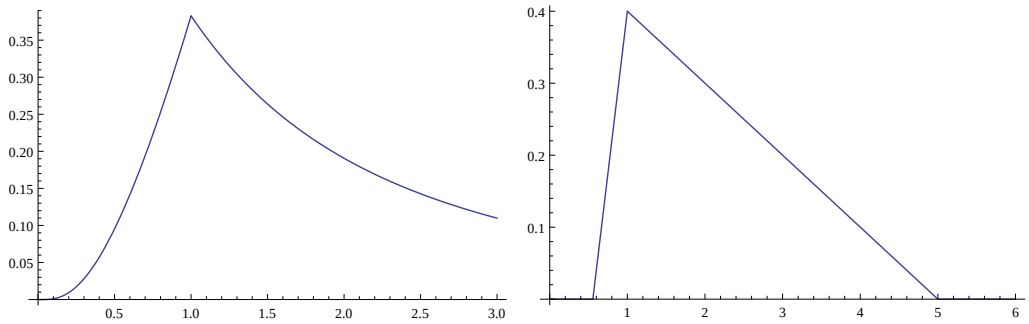
The main apparatus for obtaining the Dobrushin curve of total variation in Theorem 4 is the infimum-representation via couplings, thanks to the special role of the total variation as a Wasserstein distance. Unfortunately such representation is not known for other divergences such as the Hellinger distance or KL divergence. To extend the contraction property of total variation, our strategy is as follows: We first study a special family of f-divergences $\{\mathcal{E}_\gamma(P\|Q) : \gamma > 0\}$, which enjoys the same contraction property as the total variation for any channel. Then using an integral representation of general f-divergences [CKZ98] in terms of $\mathcal{E}_\gamma$, we extend the contraction results in Section 2.3 for additive-noise channels to f-divergences, in particular, Rényi divergences.

3.1 A parameterized family of f-divergences

For a pair of distributions P, Q, define the following family of f-divergences parameterized by $\gamma \ge 0$:

$$\mathcal{E}_\gamma(P\|Q) = \frac{1}{2}\int |\mathrm{d}P - \gamma \mathrm{d}Q| - \frac{1}{2}|1-\gamma|\,. \tag{62}$$

Typical plots of $\gamma \mapsto \mathcal{E}_\gamma(P\|Q)$ are given in Fig. 3 where P and Q are Gaussians or Bernoullis.



(a) $P = \mathcal{N}(1,1)$, $Q = \mathcal{N}(0,1)$. (b) $P = \mathrm{Bern}(0.5)$, $Q = \mathrm{Bern}(0.1)$.

Figure 3: Plot of $\gamma \mapsto \mathcal{E}_\gamma(P\|Q)$.

Some general properties of $\mathcal{E}_\gamma$ are as follows:

1. $\mathcal{E}_0(P\|Q) = \mathcal{E}_\infty(P\|Q) = 0$.
2. $\mathcal{E}_1(P\|Q) = \mathsf{TV}(P,Q)$.
3. $\gamma \mapsto \mathcal{E}_\gamma(P\|Q)$ is convex, positive, increasing on $[0,1]$, and decreasing on $[1,+\infty)$.
4. Reciprocity:
$$\mathcal{E}_\gamma(P\|Q) = \gamma \mathcal{E}_{\gamma^{-1}}(Q\|P).$$
5. Derivative of $\mathcal{E}$ recovers $Q[\frac{\mathrm{d}P}{\mathrm{d}Q} < \gamma]$:
$$\begin{aligned}\frac{d\mathcal{E}_\gamma}{d\gamma} &= \frac{1}{2}\operatorname{sign}(1-\gamma) + Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} < \gamma\Big] - \frac{1}{2} \\ &= \mathbf{1}_{\{\gamma<1\}} - Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma\Big] \end{aligned} \tag{63}$$
6. F-contraction property: If P', Q' are outputs of P, Q under some channel $P_{Y|X}$ with known F_{TV}, then
$$\mathcal{E}_\gamma(P'\|Q') \le F_{\mathsf{TV}}(\mathcal{E}_\gamma(P\|Q)). \tag{64}$$
This follows from the more general result below, which shows that the divergence $\mathcal{E}_\gamma$ for general γ enjoys the same (if not better) contraction property as the total variation, i.e., $\mathcal{E}_1$.

Proposition 8. *Assume that for each choice of $a > 0$ in (34) the corresponding F_{TV} curve is denoted by $t \mapsto F_{\mathsf{TV}}(t,a)$. Then for any channel $P_{Y|X}$ and any $(P,Q) \in \mathcal{G}_a$ we have*
$$\mathcal{E}_\gamma(P_{Y|X}\circ P\|P_{Y|X}\circ Q) \le F_{\mathsf{TV}}(\mathcal{E}_\gamma(P\|Q), a(\gamma\wedge 1)) \tag{65}$$
and, in particular, (64) holds.

Proof. First notice that if ν is any signed measure on $\mathcal{X}$ satisfying
$$\int \mathrm{d}\nu = 0, \quad \int \mathrm{d}|\nu| \le 2, \quad \int \mathsf{M}(|x|)\,\mathrm{d}|\nu| \le 2a'. \tag{66}$$
for some $a' > 0$, then we have[5]
$$\mathsf{TV}(P_{Y|X}\circ\nu, 0) \le F_{\mathsf{TV}}(\mathsf{TV}(\nu,0), a'). \tag{67}$$
Indeed, let $\nu = \nu^+ - \nu^-$ be the Jordan decomposition of ν. Then by the assumption (66) we have that $\nu^\pm$ are mutually singular sub-probability measures. Thus by introducing $P = \nu^+ + \epsilon\delta_0, Q = \nu^- + \epsilon\delta_0$ for some constant $\epsilon \ge 0$ chosen so that P and Q are probability measures, we get
$$\mathsf{TV}(P_{Y|X}\circ P, P_{Y|X}\circ Q) \le F_{\mathsf{TV}}(\mathsf{TV}(P,Q), a') \tag{68}$$
since $(P,Q) \in \mathcal{G}_{a'}$. In turn, (68) is equivalent to (67).

Now consider $\gamma < 1$ and a pair of probability measures $(P,Q) \in \mathcal{G}_a$. Write $\mathcal{E}_\gamma = \mathcal{E}_\gamma(P\|Q)$ and set
$$\nu = c(P-\gamma Q)^+ - (P-\gamma Q)^-, \qquad c = \frac{\mathcal{E}_\gamma}{\mathcal{E}_\gamma + 1 - \gamma} \le 1\,.$$

[5]The push-forward operation is extended to signed non-probability measures in the obvious way: $P_{Y|X}\circ\nu(\cdot) = \int P_{Y|X}(\cdot)\nu^+(\mathrm{d}x) - \int P_{Y|X}(\cdot)\nu^-(\mathrm{d}x)$.

Since $\mathcal{E}_\gamma \le \gamma \mathsf{TV}(P,Q) \le \gamma$, which follows from the convexity of $\gamma \mapsto \mathcal{E}_\gamma$, we have $c \le \gamma$. Then

$$\int \mathsf{M}(|x|)\,\mathrm{d}|\nu| \le \gamma \mathbb{E}_Q[\mathsf{M}(|X|)] + c\,\mathbb{E}_P[\mathsf{M}(|X|)] \le 2\gamma a\,.$$

Consequently ν satisfies condition (66) with $a' = \gamma a$. Furthermore, observe that for $\gamma \le 1$ we have

$$\int (\mathrm{d}P - \gamma \mathrm{d}Q)^- = \mathcal{E}_\gamma, \quad \int (\mathrm{d}P - \gamma \mathrm{d}Q)^+ = \mathcal{E}_\gamma + 1 - \gamma, \tag{69}$$

we have $\mathsf{TV}(\nu, 0) = \mathcal{E}_\gamma$. Thus from (67) we get

$$\mathsf{TV}(P_{Y|X} \circ \nu, 0) \le F_{\mathsf{TV}}(\mathcal{E}_\gamma, \gamma a).$$

Next from the representation

$$P - Q = \nu + (1-c)(P - \gamma Q)^+$$

and the triangle inequality we have

$$\mathsf{TV}(P_{Y|X} \circ P, P_{Y|X} \circ Q) \le \mathsf{TV}(P_{Y|X} \circ \nu, 0) + \frac{1-c}{2} \int (\mathrm{d}P - \gamma \mathrm{d}Q)^+\,. \tag{70}$$

In view of (69), it remains to notice that the last term in (70) equals $\frac{1-\gamma}{2}$, from which (65) follows via

$$\mathcal{E}_\gamma(P_{Y|X} \circ P \| P_{Y|X} \circ P) = \mathsf{TV}(P_{Y|X} \circ P, P_{Y|X} \circ Q) - \frac{1-\gamma}{2}\,.$$

For $\gamma > 1$ the proof is entirely analogous, except that we set

$$\nu = (P - \gamma Q)^+ - c(P - \gamma Q)^-\,, \qquad c = \frac{\mathcal{E}}{\mathcal{E} + \gamma - 1} \le 1$$

and the best bound we have on $\int \mathsf{M}(|x|)\mathrm{d}|\nu|$ is $2a$, which follows from the fact that $\mathcal{E}_\gamma \le 1$ and hence $c\gamma \le 1$. □

3.2 Integral representation and contraction of Rényi divergences

For an f-divergence, analogous to the Dobrushin curve (32) we define

$$F_f(t) \triangleq \sup\{D_f(P_{Y|X} \circ P, P_{Y|X} \circ Q) : D_f(P,Q) \le t, (P,Q) \in \mathcal{G}_a\}, \tag{71}$$

Note that the usual data processing inequality amounts to $F_f(t) \le t$. We say the channel $P_{Y|X}$ *contracts* the f-divergence D_f if $F_f(t) < t$ for all t in a neighborhood near zero. We have already shown that the total variation is always contracted by additive noise satisfying the necessary and sufficient condition in Theorem 3. In view of Proposition 8, the formulas in Corollaries 5 and 6 apply to $\mathcal{E}_\gamma$ as well. A natural question is in order: Do other divergences, such as the KL divergence, also contract in additive noise? To this end, we need the following integral representation of f-divergences in terms of the family of divergence $\mathcal{E}_\gamma$: If $f \in C^2(\mathbb{R}_+)$, then (see [CKZ98, Corollary 3.7, p. 99])

$$D_f(P\|Q) = \int_0^\infty \mathcal{E}_\gamma(P\|Q) f''(\gamma)\mathrm{d}\gamma. \tag{72}$$

For instance, the area under the curve $\gamma \mapsto \mathcal{E}_\gamma$ is half the χ^2-divergence $\chi^2(P\|Q) = \int \frac{\mathrm{d}P^2}{\mathrm{d}Q} - 1$.

For conciseness, below we focus on the scalar AWGN channel under the first moment constraint and the special case of Rényi divergence of order α, which is a monotonic transformation of the f_α-divergence with

$$f_\alpha(x) = \begin{cases} 1 - x^\alpha & \alpha \in (0,1) \\ x \log x & \alpha = 1 \\ x^\alpha - 1 & \alpha > 1 \end{cases}. \tag{73}$$

Note that the special case of $\alpha = 1, 2, \frac{1}{2}$ corresponds to the KL divergence $D(P\|Q)$, the χ^2-divergence $\chi^2(P\|Q)$, and half the squared Hellinger distance $H^2(P,Q) = \int(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$, respectively. The following result shows that the AWGN channel contracts Rényi divergence of order α if and only if $\alpha \in (0,1)$. Consequently, the Hellinger distance always contracts when passing through the AWGN channel, but χ^2 and KL divergences do not.

Theorem 9. *Consider the scalar AWGN channel $P_{Y|X} = \mathcal{N}(X,1)$. Let $\mathsf{M}(|x|) = |x|^2$ and $a > 0$. Then*

1. *For $\alpha \in (0,1)$, for any $\epsilon > 0$,*

$$F_{f_\alpha}(t) \le t\left(1 - 2\mathsf{Q}\left(\sqrt{a}t^{-\frac{1+\epsilon}{2\alpha}}\right)\right) + (1 + \alpha - \alpha^2)t^{1+\epsilon}, \quad 0 < t < 1. \tag{74}$$

2. *For $\alpha \ge 1$,*

$$F_{f_\alpha}(t) = t, \tag{75}$$

which holds for all $t > 0$ if $\alpha > 1$ and $t < \frac{a}{8}$ if $\alpha = 1$, respectively.

Proof. 1° Fix $\alpha \in (0,1)$ and $(P,Q) \in \mathcal{G}_a$ such that $D_{f_\alpha}(P\|Q) \in (0,1)$. Let $\mathcal{N}$ denote the standard normal distribution. Fix $\delta > 0$. Applying the integral representation (72) to $f_\alpha(x) = 1 - x^\alpha$, we have

$$\begin{aligned} D_{f_\alpha}(P * \mathcal{N}\|Q * \mathcal{N}) &= \alpha(1-\alpha)\int_0^\infty \mathcal{E}_\gamma(P * \mathcal{N}\|Q * \mathcal{N})\gamma^{\alpha-2}\mathrm{d}\gamma \\ &\le \alpha(1-\alpha)\int_0^\infty \mathcal{E}_\gamma\left(1 - 2\mathsf{Q}\left(\sqrt{\frac{a}{\mathcal{E}_\gamma}}\right)\right)\gamma^{\alpha-2}\mathrm{d}\gamma \qquad (76) \\ &\le \left(1 - 2\mathsf{Q}\left(\sqrt{\frac{a}{\delta}}\right)\right) D_{f_\alpha}(P\|Q) + \alpha(1-\alpha)\int_0^\infty \mathcal{E}_\gamma\gamma^{\alpha-2}\mathbf{1}_{\{\mathcal{E}_\gamma \le \delta\}}\mathrm{d}\gamma \qquad (77) \\ &\le \left(1 - 2\mathsf{Q}\left(\sqrt{\frac{a}{\delta}}\right)\right) D_{f_\alpha}(P\|Q) + \alpha(1-\alpha)\int_0^1 \mathcal{E}_\gamma\gamma^{\alpha-2}\mathbf{1}_{\{\mathcal{E}_\gamma \le \delta\}}\mathrm{d}\gamma + \alpha\delta, \qquad (78) \end{aligned}$$

where (76) follows from Corollary 6 with $\mathcal{E}_\gamma = \mathcal{E}_\gamma(P\|Q)$, and (77) follows from (72), and (78) is due to $\mathcal{E}_\gamma \le \mathsf{TV} \le 1$. Using (63), for all $\gamma \in (0,1)$, we have $\mathcal{E}'_\gamma = Q[\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma] \le 1$. By the convexity of $\gamma \mapsto \mathcal{E}_\gamma$ and $\mathcal{E}_0 = 0$, we have $\mathcal{E}_\gamma \le Q[\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma]\gamma \le \gamma$. Therefore

$$\begin{aligned} \int_0^1 \mathcal{E}_\gamma\gamma^{\alpha-2}\mathbf{1}_{\{\mathcal{E}_\gamma \le \delta\}}\mathrm{d}\gamma &= \int_0^\delta \mathcal{E}_\gamma\gamma^{\alpha-2}\mathbf{1}_{\{\mathcal{E}_\gamma \le \delta\}}\mathrm{d}\gamma + \int_\delta^1 \mathcal{E}_\gamma\gamma^{\alpha-2}\mathbf{1}_{\{\mathcal{E}_\gamma \le \delta\}}\mathrm{d}\gamma \\ &\le \int_0^\delta \gamma^{\alpha-1}\mathrm{d}\gamma + \delta^{\alpha-1}\int_\delta^1 \mathbf{1}_{\{\mathcal{E}_\gamma \le \delta\}}\mathcal{E}'_\gamma\mathrm{d}\gamma \\ &\le \frac{1+\alpha}{\alpha}\delta^\alpha. \end{aligned} \tag{79}$$

Plugging (79) into (78) and by the arbitrariness of $\delta > 0$, we obtain

$$D_{f_\alpha}(P * \mathcal{N} \| Q * \mathcal{N}) \le \inf_{0<\delta<1} \left\{ \left(1 - 2\mathsf{Q}\left(\sqrt{\frac{a}{\delta}}\right)\right) D_{f_\alpha}(P\|Q) + (1 + \alpha - \alpha^2)\delta^\alpha \right\},$$

which implies the desired (74) upon choosing $\delta = (D_{f_\alpha}(P\|Q))^{(1+\epsilon)/\alpha}$.

2° Turning to the case of $\alpha \ge 1$, we construct examples where D_{f_α} does not contract. Fix $t > 0$ and let $q > 0$ be sufficiently small. Let $P_q = (1-p)\delta_0 + p\delta_b, Q_q = (1-q)\delta_0 + q\delta_b$ with $b = \sqrt{\frac{a}{p}}$ and $p = \frac{t}{\log\frac{1}{q}}$ if $\alpha = 1$ and $p = q(\frac{t}{q})^{1/\alpha}$ if $\alpha > 1$. Then it is clear that $(P_q, Q_q) \in \mathcal{G}_a$ for all sufficiently small q. Furthermore,

$$D_{f_\alpha}(P_q\|Q_q) = d_\alpha(p\|q) = t + o(1), \qquad q \to 0\,,$$

where $d_\alpha(p\|q) \triangleq q^{1-\alpha}p^\alpha + (1-q)^{1-\alpha}(1-p)^\alpha$ if $\alpha > 1$ and $p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ if $\alpha = 1$.

Next, by applying the data-processing inequality to the transformation $y \mapsto \mathbf{1}_{\{y \ge b/2\}}$ we get

$$D_{f_\alpha}(P_q * \mathcal{N} \| Q_q * \mathcal{N}) \ge d_\alpha(p'\|q'),$$

where $p' = p + (1-2p)\mathsf{Q}(b/2) = p(1+o(1))$ and $q' = q + (1-2q)\mathsf{Q}(b/2) = q(1+o(1))$. This follows from the fact that $\mathsf{Q}(b/2) = o(q)$, which is obvious for $\alpha > 1$; for $\alpha = 1$, since we have assumed that $t < a/8$, we have $\mathsf{Q}(b/2) \le \exp(-b^2/8) = q^{a/8t} = o(q)$. Consequently, $D_{f_\alpha}(P_q * \mathcal{N}\|Q_q * \mathcal{N}) \ge d_\alpha(p'\|q') = t + o(1)$ as $q \to 0$, which completes the proof of (75). □

Remark 6. Theorem 9 extends in the following directions:

1. For general additive noise Z, (74) continues to hold with $1 - 2\mathsf{Q}(\cdot)$ replaced by the concave envelope $\theta_c(\cdot)$ in Theorem 4.
2. For the p^{th}-moment constraint with $\mathsf{M}(|x|) = |x|^p$ and $p > 2$, (75) holds for all $t, a > 0$ if $\alpha > 1$. For KL divergence ($\alpha = 1$), however, it remains unclear whether (75) holds in a neighborhood near zero since the above construction no longer applies.

4 Proof of Theorem 1

Theorem 1 follows from Propositions 11, 13 and 14 given in Sections 4.1, 4.3 and 4.4, respectively. The special case of finite-alphabet W is much simpler and is treated by Proposition 12 (Section 4.2). Finally, Section 4.5 shows that our converse bounds are optimal for total variation, mutual information and correlation in the scalar Gaussian case.

4.1 Convergence in total variation

The development in Section 2 deals with comparing a pair of distributions and studies by how much their total variation shrinks due to smoothing by the additive noise. Therefore these results are applicable to binary sources, i.e., transmitting one bit. What if the sources takes more than two, or rather, a continuum of, values? To this end, the data processing inequality for mutual information is relevant, which states that $W \to X \to Y \to Z$ implies that $I(W;Z) \le I(X;Y)$. In other words, dependency decreases on Markov chains. Our goal next is to find a quantitative data pre-processing and post-processing inequalities as a counterpart of Theorem 4. Since we know, in

view of Theorem 9, that KL divergence does not contract, it is natural to turn to total variation and define the following T-*information*:

$$T(X;Y) \triangleq \mathsf{TV}(P_{XY}, P_X P_Y), \tag{80}$$

which has been studied in, e.g., [Csi96, Pin05]. Similar to mutual information, it is easy to see that the T-information satisfies the following properties:

1. $T(X;Y) = \mathbb{E}[\mathsf{TV}(P_{Y|X}, P_Y)] = \mathbb{E}[\mathsf{TV}(P_{X|Y}, P_X)] = T(Y;X)$.
2. Data-processing inequality: $W \to X \to Y \to Z$ implies that $T(W;Z) \le T(X;Y)$.
3. If S is Bern($\frac{1}{2}$), then
$$T(S;X) = \frac{1}{2}\mathsf{TV}(P_{X|S=0}, P_{X|S=1}). \tag{81}$$
4. If S and $\hat{S}$ are both binary, then[6]
$$T(S;\hat{S}) \ge \min\{\mathbb{P}[S=0], \mathbb{P}[S=1]\} - \mathbb{P}[S \ne \hat{S}]. \tag{82}$$
5. Pinsker's inequality:
$$I(X;Y) \ge 2 \log \mathrm{e}\, T(X;Y)^2. \tag{83}$$

The next theorem gives a quantitative data processing theorem for the T-information with additive noise:

Theorem 10. *Let $W \to X \to Y$, where $Y = X + Z$ and $\mathbb{E}[\mathsf{M}(|X|)] \le a$. Let θ_c be as in Theorem 4. Then*

$$T(W;Y) \le f(T(W;X), a), \quad f(t,a) \triangleq t\theta_c\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right). \tag{84}$$

Remark 7. Exactly the same inequality holds for the following functional of real-valued random variables

$$T'(A;B) \triangleq \inf_{\mathbb{E}_{Q_B}[\mathsf{M}(|B|)] \le a} \mathsf{TV}(P_{AB}, P_A Q_B),$$

which is a natural extension of the K-information of Sibson [Sib69] and Csiszár [Csi95] and satisfies $T'(A;B) \le T(A;B)$. Optimizing over Q_B instead of taking $Q_B = P_B$ may lead to more powerful converse bounds, see [PV10] for details.

Proof. By the definition of $T(W;Y)$ and the Markov chain condition, we have

$$T(W;Y) = \int \mathsf{TV}(P_{X|W=w} * P_Z, P_X * P_Z)\, P_W(\mathrm{d}w).$$

Then Theorem 4 yields

$$\mathsf{TV}(P_{X|W=w} * P_Z, P_X * P_Z) \le f\left(\mathsf{TV}(P_{X|W=w}, P_X), \frac{1}{2}\mathbb{E}[\mathsf{M}(|X|)|W=w] + \frac{1}{2}\mathbb{E}[\mathsf{M}(|X|)]\right). \tag{85}$$

In view of Remark 4, the function f defined in (84) is jointly concave and non-decreasing in each argument. Thus taking expectation over $w \sim P_W$ on the right-hand side of (85) and applying Jensen's inequality, we complete the proof. □

[6] To see this, let $S \sim \mathrm{Bern}(p), \hat{S} \sim \mathrm{Bern}(q)$, $p_0 = \mathbb{P}[S=1|\hat{S}=0]$ and $p_1 = \mathbb{P}[S=0|\hat{S}=1]$. Then $T(S;\hat{S}) = \bar{q}|p_0 - p| + q|p_1 - \bar{p}| \ge p\bar{q} + \bar{p}q - (\bar{q}p_0 + qp_1) \ge \min\{p, \bar{p}\} - \mathbb{P}[\hat{S} \ne S]$.

As an application of Theorem 10, next we describe how the T-information decays on the Markov chain (1).

Proposition 11. *Assume the Markov chain (1), where Z_j are i.i.d. and $\mathbb{E}[\mathsf{M}(X_j)] \le a$ for all $j \in [n]$. Then for all $a > 0$ and $n \ge 2$,*

$$T(W;Y_n) \le \frac{a}{f^{-1}(n-1)}, \tag{86}$$

where $f(s) \triangleq \int_1^s \frac{1}{y\,(1-\theta_c(2\mathsf{M}^{-1}(y)))}\mathrm{d}y$.

In particular, if $Z_j \sim \mathcal{N}(0,1)$ are i.i.d., then

$$T(W;Y_n) \le Ca\exp(-g^{-1}(n)), \tag{87}$$

where $g(s) \triangleq \int_0^s \exp[\frac{1}{2}\mathsf{M}^{-1}(\exp(\tau))^2]\mathrm{d}\tau$, and C is a positive constant only depending on the cost function M.

Remark 8 (Gaussian noise)**.** Particularizing the result of Proposition 11 to the AWGN channel and the following cost functions we obtain the corresponding convergence rates

a) p^{th}-moment constraint: $\mathbb{E}|X_k|^p \le a$ for some $p \ge 1$. Then $T(W;Y_n) = O((\log n)^{-p/2})$. In particular, for power constraint $\mathsf{M}(x) = x^2$, (8) holds.

b) Sub-exponential: $\mathbb{E}\exp(\alpha|X_k|^2) \le a$ for some $\alpha > 0$ and $a > 1$. Then $T(W;Y_n) = O(\mathrm{e}^{-\sqrt{2\alpha\log n}})$.

c) Sub-Gaussian: $\mathbb{E}\exp(\alpha|X_k|^2) \le a$ for some $\alpha > 0$ and $a > 1$. Then $T(W;Y_n) = O(n^{-2\alpha})$.

Intuitively, the faster the cost function grows, the closer we are to amplitude-constrained scenarios, where we know that information contracts linearly thanks to the Dobrushin's coefficient being strictly less than one. Hence we expect the convergence rate to be faster and closer to, but always strictly slower than, exponential decay. In view of (81), Proposition 11 implies that transmitting one bit is impossible under any cost constraint, since the optimal Type-I+II error probability is given by $\frac{1}{2} - \mathsf{TV}(P_{Y_n|W=0}, P_{Y_n|W=1})$ (see [LR06, Theorem 13.1.1]) and the total-variation vanishes as $n \to \infty$.

The slow convergence rates obtained above for Gaussian noise can be explained as follows: In view of (88), the T-information obeys the iteration $T(W;Y_n) \le F_{\mathsf{TV}}(T(W;Y_{n-1}))$. For instance, consider the Dobrushin curve under unit power constraint is given by $F_{\mathsf{TV}}(t) = t(1 - 2\mathsf{Q}(1/\sqrt{t}))$, which satisfies $F_{\mathsf{TV}}{}'(0) = 1$ and all other derivatives vanish at zero. Therefore F_{TV} is smooth but not real analytic at zero, and the rate of convergence of the iteration $x_n = F_{\mathsf{TV}}(x_{n-1})$ to the fixed point zero is very slow. See Fig. 2 for an illustration.

Proof. By Theorem 10, we have

$$T(W;Y_n) \le F_{\mathsf{TV}}(T(W;X_n)) \le F_{\mathsf{TV}}(T(W;Y_{n-1})), \tag{88}$$

where the first inequality follows from Theorem 10, and the second inequality follows from the data processing theorem for T and the monotonicity of F_{TV}. Applying Theorem 4, we have

$$F_{\mathsf{TV}}(t) \le t\theta_c\left(2\mathsf{M}^{-1}\left(\frac{a}{t}\right)\right).$$

Repeating the above argument leads to

$$T(W;Y_n) \le at_n,$$

where the sequence $\{t_n\}$ is defined iteratively via

$$t_{n+1} = t_n - h(t_n) \tag{89}$$

with $h(t) = t(1-\theta_c\left(2\mathsf{M}^{-1}\left(\frac{1}{t}\right)\right))$ and $t_1 = 1$. By Theorem 4, θ_c is strictly increasing. Therefore h is an increasing function. Applying Lemma 20 in Appendix A, the convergence rate of the sequence (89) satisfies

$$t_n \le G^{-1}(n-1) = \frac{1}{f^{-1}(n-1)},$$

where $G(t) = \int_t^1 \frac{1}{2y(1-\theta_c(2\mathsf{M}^{-1}(\frac{1}{y})))}\mathrm{d}y$.

For the Gaussian noise, we have $\theta_c(x) = \theta(x) = 1 - 2\mathsf{Q}(x/2)$ (see Corollary 5). In view of the bound $\mathsf{Q}(u) \ge \frac{\varphi(u)u}{u^2+1}$ for $u > 0$, where ϕ denote the standard normal density, (87) follows from (86) upon changes of variables. □

4.2 Special case: finite-alphabet W

A consequence of the total variation estimates in Theorem 10 and Proposition 11 is that for finitely-valued message W they entail estimates on the mutual information and maximal correlation, as the next proposition shows.[7]

Proposition 12. *Assume W take values on a finite set $\mathcal{W}$ and let $p_{W,\min}$ denote the minimal non-zero mass of P_W. Then*

$$I(W;Y) \le \log(|\mathcal{W}|-1)T(W;Y) + h(T(W;Y)) \tag{90}$$

$$S^2(W;Y) \le \chi^2(P_{WY}\|P_W P_Y) \tag{91}$$

$$\le \frac{1}{p_{W,\min}} T(W;Y)\,, \tag{92}$$

where $S(W;Y)$ and χ^2 are defined in (15) and (12), respectively, and $h(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$ is the binary entropy function.

Proof. By coupling and Fano's inequality, for any P and Q on $\mathcal{W}$, we have

$$|H(P) - H(Q)| \le \mathsf{TV}(P,Q)\log(|\mathcal{W}|-1) + h(\mathsf{TV}(P,Q)).$$

Then

$$\begin{aligned} I(W;Y) &= H(W) - H(W|Y) \\ &\le \mathbb{E}_{y\sim P_Y}[\log(|\mathcal{W}|-1)\mathsf{TV}(P_W, P_{W|Y=y}) + h(\mathsf{TV}(P_W, P_{W|Y=y}))] \\ &\le \log(|\mathcal{W}|-1)\mathsf{TV}(P_W P_Y, P_{WY}) + h(\mathsf{TV}(P_W P_Y, P_{WY})), \end{aligned}$$

where the last step is due to the concavity of $h(\cdot)$.

The inequality (91) follows [Wit75] by noticing that $\chi^2(P_{WY}\|P_W P_Y)$ is the sum of squares of the singular values of $f(W) \mapsto \mathbb{E}[f(W)|Y]$ minus 1 (the largest one), while $S(W;Y)$ is the second

[7] The bound (90) is essentially [Csi96, Lemma 1]. The bound (92) was shown by F. P. Calmon <flavio@mit.edu> and included here with his permission.

largest singular value. Bound (92) follows from the chain:

$$\begin{aligned}\chi^2(P_{WY}\|P_W P_Y) &= \mathbb{E}_{P_{WY}}\left[\frac{P_{W|Y}(W|Y)}{P_W(W)}\right] - 1\\ &= \mathbb{E}_{P_{WY}}\left[\frac{P_{W|Y}(W|Y)}{P_W(W)}\right] - \mathbb{E}_{P_W P_Y}\left[\frac{P_{W|Y}(W|Y)}{P_W}\right]\\ &\le \operatorname*{esssup}_{w,y}\frac{P_{W|Y}(w|y)}{P_W(w)}\cdot \mathsf{TV}(P_{WY}, P_W P_Y)\\ &\le \frac{1}{p_{W,\min}} T(W;Y)\,,\end{aligned}$$

where first step is by (12) and the rest are self-evident. □

Combining Propositions 11 and 12, we conclude that both $S(W;Y_n)$ and $I(W;Y_n)$ vanish for finitely-valued W. In particular, for Gaussian noise, by Remark 8 (second moment constraint) we have $T(W;Y_n) = O(\frac{1}{\log n})$. Then the maximal correlation satisfies $S(W;Y_n) = O(\frac{1}{\sqrt{\log n}})$ and the mutual information vanishes according to

$$I(W;Y_n) = O\left(h\left(\frac{1}{\log n}\right)\right) = O\left(\frac{\log\log n}{\log n}\right). \tag{93}$$

4.3 Convergence of mutual information

In this subsection we focus on the AWGN channel and show that the convergence rate (93) continues to hold for *any* random variable W, which will be useful for applications in optimal stochastic control where W is Gaussian distributed.[8] To deal with non-discrete W, a natural idea to apply is *quantization*. By Propositions 11 and 12, for any quantizer $q : \mathbb{R}^d \to [m]$, we have

$$I(q(W);Y_n) \le \frac{C\log m}{\log n} + h\left(\frac{C}{\log n}\right) \tag{95}$$

for some universal constant C. A natural conjecture is the following implication: For any sequence of channels $P_{Y_n|X}$ we have:

$$\forall m \in \mathbb{N}, \forall q : \mathbb{R}^d \to [m] : I(q(W);Y_n) \to 0 \quad \Longrightarrow \quad I(W;Y_n) \to 0,$$

which would imply the desired conclusion that mutual information vanishes. Somewhat counter-intuitively, this conjecture is generally false, as the following counterexample shows: Consider $X \sim \text{Uniform}([0,1])$ and

$$Y_n = \begin{cases} 0, \frac{1}{n} \le X \le 1, \\ k, (k-1)\frac{2^{-n}}{n} \le X < k\frac{2^{-n}}{n}, \end{cases} \qquad k = 1, \dots, 2^n.$$

[8] *Added in print:* Another method of showing (93) is to directly use the strong data processing inequality for mutual information in Gaussian noise, cf. [PW15, CPW15]. Namely, it is possible to show the existence of certain non-linear function F_I such that $F_I(t) < t$ and

$$I(W; X+Z) \le F_I(I(W;X)) \tag{94}$$

for all $(W,X) \perp\!\!\!\perp Z$ and $\mathbb{E}[|X|^2] \le E$. Then (93) follows by applying (94) repeatedly and the behavior of F_I curve near zero: $F_I(t) = t - e^{-\frac{E}{t}\ln\frac{1}{t} + \Theta(\ln\frac{1}{t})}$.

On one hand it is clear that $I(X;Y_n) \to \infty$. On the other hand, among all m-point quantizers q, it is clear that the optimal one is to quantize to some levels corresponding to the partition that Y_n incurs (other quantizers are just equivalent to randomization). Thus

$$\sup_{q:[0,1]\to[m]} I(q(X);Y_n) = \sup_{q:[2^n+1]\to[m]} H(q(Y_n)).$$

But the RHS tends to zero as $n \to \infty$ for any fixed m because the dominating atom shoots up to 1. The same example also shows that

$$T(X;Y_n) \to 0 \quad \not\Longrightarrow \quad I(X;Y_n) \to 0\,. \tag{96}$$

Nevertheless, under additional constraints on kernels $P_{Y_n|W}$, we can prove that (96) indeed holds and obtain the convergence rate. The main idea is to show that the set of distributions $\{P_{Y_n|W=w}, w \in \mathbb{R}^d\}$ can be grouped into finitely many clusters, so that the diameter (in KL divergence) of each cluster is arbitrarily small. This can indeed be done in our setting since the channel $P_{Y_n|W}$ is a stochastically degraded version of an AWGN channel.

Proposition 13. *Let W, X_k, Y_k be as in Theorem 1. If $\mathbb{E}[\|X_k\|^2] \le dE$ for all $k \in [n]$, then*

$$I(W;Y_n) \le \frac{d}{2}\log\left(1+\frac{dE}{\log n}\right)+\frac{d^2E}{2\log n}\log\left(1+\frac{\log n}{d}\right)+\frac{Cd^2E}{\log n}\log\left(1+\frac{2\log n}{d\sqrt{E}}\right)+h\left(\frac{CdE}{\log n}\wedge 1\right), \tag{97}$$

where C is the absolute constant in (8). In particular, for fixed d and E,

$$I(W;Y_n) = O\left(\frac{\log\log n}{\log n}\right). \tag{98}$$

Remark 9. Note that the upper bound (97) deteriorates as the dimension d grows, which is to be expected. Indeed, for large d one can employ very reliable error-correcting codes for the AWGN channel with blocklength d, that can tolerate a large number of hops over the AWGN channels. If the blocklength $d = d_n$ grows with n such that $d_n = O(\log n)$ and the power per coordinate E is fixed, then (97) reduces to

$$I(W;Y_n) \le O\left(\frac{d_n^2}{\log n}\log\frac{\log n}{d_n}\right).$$

Using Fano's inequality, this implies that in order to reliably communicate over n hops at some positive rate, thereby $I(W;Y_n) = \Omega(d_n)$, it is necessary to have the blocklength d_n grow at least as fast as

$$d_n = \Omega(\log n). \tag{99}$$

This conclusion has been obtained in [Sub12] under the simplified assumption of almost sure power constraint of the codebook (see (11)). Here Proposition 13 extends it to power constraint in expectation.

Proof of Proposition 13. Fix $u, \epsilon > 0$ to be specified later. It is well-known that the ℓ_2-ball in $\mathbb{R}^d$ of radius u can be covered by at most $m = \lfloor(1+\frac{2u}{\epsilon})^d\rfloor$ ℓ_2-balls of radius ϵ, whose centers are denoted by $x_1, \ldots, x_m$. Define $q : \mathbb{R}^d \to [m+1]$ by

$$q(x) = \left(\operatorname*{argmin}_{i\in[m]} \|x_i - x\|\right)\mathbf{1}_{\{\|x\|\le u\}} + (m+1)\mathbf{1}_{\{\|x\|>u\}}.$$

Then $\mathbb{E}\left[\|X_1 - x_i\|_2^2 | q(X_1) = j\right] \le \epsilon^2$ for any $j \in [m]$. Hence

$$I(X_1; Y_n | q(X_1) = j) \le I(X_1; Y_1 | q(X_1) = j) \tag{100}$$

$$\le \frac{d}{2} \log \left(1 + \frac{\mathbb{E}[\|X_1 - x_j\|^2 | q(X_1) = j]}{d}\right) \tag{101}$$

$$\le \frac{d}{2} \log \left(1 + \frac{\epsilon^2}{d}\right), \tag{102}$$

where in (100) we used the Markov relation $q(X_1) \to X_1 \to Y_1 \to Y_n$, and (101) follows from the vector AWGN channel capacity:

$$\sup_{P_X : \mathbb{E}[\|X\|_2^2 \le P]} I(X; X + Z) = \frac{d}{2} \log \left(1 + \frac{P}{d}\right), \tag{103}$$

where $Z \sim \mathcal{N}(0, \mathbf{I}_d)$ is independent of X. Similarly,

$$\begin{aligned} I(X_1; Y_n | q(X_1) = m+1) &\le I(X_1; Y_1 | q(X_1) = m+1) \\ &\le \frac{d}{2} \log \left(1 + \frac{\mathbb{E}[\|X_1\|^2 | \|X_1\| > u]}{d}\right) \\ &\le \frac{d}{2} \log \left(1 + \frac{E}{\mathbb{P}\left[\|X_1\| > u\right]}\right), \end{aligned} \tag{104}$$

where (104) follows from the fact that $\mathbb{E}[\|X_1\|^2 | \|X_1\| > u] \mathbb{P}\left[\|X_1\| > u\right] \le \mathbb{E}[\|X_1\|^2]$.

Averaging (102) and (104) over $q(X_1) = j \in [m+1]$, we obtain

$$\begin{aligned} I(X_1; Y_n | q(X_1)) &\le \frac{d}{2} \log \left(1 + \frac{\epsilon^2}{d}\right) + \frac{d}{2} \mathbb{P}\left[\|X_1\| > u\right] \log \left(1 + \frac{E}{\mathbb{P}\left[\|X_1\| > u\right]}\right) \\ &\le \frac{d}{2} \log \left(1 + \frac{\epsilon^2}{d}\right) + \frac{d^2 E}{2u^2} \log \left(1 + \frac{u^2}{d}\right), \end{aligned} \tag{105}$$

where (105) follows from the fact that $x \mapsto x \ln(1 + \frac{1}{x})$ is increasing on $\mathbb{R}_+$[9] and the Chebyshev's inequality:

$$\mathbb{P}\left[\|X_1\| \ge u\right] \le \frac{\mathbb{E}[\|X_1\|^2]}{u^2} \le \frac{dE}{u^2}. \tag{106}$$

Applying Proposition 12, we have

$$I(q(X_1); Y_n) \le t_n d \log \left(1 + \frac{2u}{\epsilon}\right) + h(t_n), \tag{107}$$

where $t_n = T(q(X_1); Y_n) \le T(X_1; Y_n) \le \frac{CdE}{\log n}$ in view of (8)

Combining (104) and (107) yields

$$I(W; Y_n) \le I(X_1; Y_n) \tag{108}$$

$$= I(q(X_1); Y_n) + I(X_1; Y_n | q(X_1)) \tag{109}$$

$$\le \frac{d}{2} \log \left(1 + \frac{\epsilon^2}{d}\right) + \frac{d^2 E}{2u^2} \log \left(1 + \frac{u^2}{d}\right) + \frac{Cd^2 E}{\log n} \log \left(1 + \frac{2u}{\epsilon}\right) + h\left(\frac{CdE}{\log n} \wedge 1\right). \tag{110}$$

Choosing $u = \sqrt{\log n}$ and $\epsilon^2 = \frac{d^2 E}{\log n}$ yields the desired (97). □

[9] Indeed, $(x \ln(1 + \frac{1}{x}))' = -\ln(1 - \frac{1}{1+x}) - \frac{1}{1+x} \ge 0$.

4.4 Convergence of correlation coefficients

Given a pair of random variables X, Y, the conditional expectation of X given Y has the maximal correlation with X among all functions of Y, i.e.

$$\sup_{g \in L_2(P_Y)} \rho(X, g(Y)) = \rho(X, \mathbb{E}[X|Y]) = \frac{\|\mathbb{E}[X|Y] - \mathbb{E}[X]\|_2}{\sqrt{\mathsf{var}(X)}},$$

which is a simple consequence of the Cauchy-Schwartz inequality. As the next result shows, vanishing mutual information provides a convenient sufficient condition for establishing vanishing correlation coefficients.

Proposition 14. *Assume that $\mathbb{E}[W^2] < \infty$. For any sequence of $P_{Y_n|W}$,*

$$\lim_{n\to\infty} I(W; Y_n) = 0 \quad \Longrightarrow \quad \lim_{n\to\infty} \rho(W, \mathbb{E}[W|Y_n]) = 0. \tag{111}$$

Moreover, if W is Gaussian, then

$$\rho^2(W, \mathbb{E}[W|Y_n]) \le 1 - \exp(-2I(W; Y_n)) \le 2I(W; Y_n). \tag{112}$$

Proof. For the Gaussian case, (112) follows from the inequality

$$I(W; \hat{W}) \ge \frac{1}{2} \log \frac{1}{1 - \rho^2(W, \hat{W})}, \tag{113}$$

which is equivalent to the Gaussian rate-distortion formula. To see the implication (111), first notice the equivalence

$$\mathbb{E}[(W - \mathbb{E}[W|Y_n])^2] \to \mathsf{var}(W) \quad \Longleftrightarrow \quad \rho(W, \mathbb{E}[W|Y_n]) \to 0.$$

From here Proposition 14 follows from the next (probably well-known) lemma. □

Lemma 15. *Assume that $\mathbb{E}[X^2] < \infty$. Let $\mathsf{var}(X) = \sigma^2$. Denote the rate-distortion function of X with respect to the mean-square error by*

$$R(D) = \inf_{P_{\hat{X}|X}: \mathbb{E}(\hat{X} - X)^2 \le D} I(X; \hat{X}).$$

Then

$$D \to \sigma^2 \Leftrightarrow R(D) \to 0. \tag{114}$$

Proof. ($\Rightarrow$) The rate-distortion function is dominated by that of the Gaussian distribution [Ber71]:

$$R(D) \le \frac{1}{2} \log^+ \frac{\sigma^2}{D}, \tag{115}$$

where $\log^+ \triangleq \max\{\log, 0\}$.

($\Leftarrow$) Note that $D \mapsto R(D)$ is decreasing and concave on $[0, \sigma^2]$, hence continuous on the open interval $(0, \sigma^2)$. Suppose there exists $D_0 < \sigma^2$ such that $R(D_0) = 0$. Then by definition of the rate-distortion function, there exists a sequence of $P_{\hat{X}_n|X}$ such that $\mathbb{E}(\hat{X}_n - X)^2 \le D$. $I(X; \hat{X}_n) \to 0$. Note that $\mathbb{E}X_n^2 \le 2D + 2\mathbb{E}X^2$ for all n. Therefore the sequence $P_{\hat{X}_n, X}$ is tight. By Prokhorov's theorem, there exists a subsequence $P_{\hat{X}_{n_k}, X}$ which converges weakly to some $P_{\hat{X}, X}$. By the lower semicontinuity of the divergence and the second-order moment, $\mathbb{E}(\hat{X}_n - X)^2 \le \liminf \mathbb{E}(\hat{X}_{n_k} - X)^2 \le D$ and $I(\hat{X}; X) \le \liminf I(\hat{X}_{n_k}; X) = 0$. Hence $\hat{X} \perp\!\!\!\perp X$, contradicting $\mathbb{E}(\hat{X}_n - X)^2 \le D < \sigma^2$. □

Proposition 14 allows us to capitalize on the results on mutual information in Section 4.3 to obtain correlation estimates for the Markov chain (1). In particular, combining (111) with Proposition 13 yields (10). Additionally, if W is Gaussian, then (112) yields

$$\rho(W, \mathbb{E}[W|Y_n]) = O(\sqrt{I(W;Y_n)}) = O\left(\sqrt{\frac{\log\log n}{\log n}}\right). \tag{116}$$

These prove the correlation part of the main result Theorem 1.

However, the estimate (116) is not entirely satisfactory in the sense that it highly depends on the Gaussianity of W; if W is not Gaussian, the rate-distortion function of W is not explicitly known and it is unclear whether (112) still applies. How to obtain quantitative estimates on the correlation coefficient if we only have sub-Gaussianity or moment constraints on W? It turns out that one can circumvent mutual information completely and directly obtain correlation estimate from the T-information, whose convergence rate has been found in Section 4.1. The key connection between total variation and correlation is the following simple observation:

Proposition 16. *Assume W is zero-mean, unit-variance. For any $q \in (1, \infty]$ we have*

$$\rho^2(W, \mathbb{E}[W|Y]) \le 4T(W;Y)^{1-\frac{1}{q}} \|W\|_{2q}^2. \tag{117}$$

If W is sub-Gaussian and $T(W;Y) < e^{-2/e}$, we have

$$\rho^2(W, \mathbb{E}[W|Y]) \le \frac{8}{\log e} \|W\|_{\psi_2}^2 \, T(W;Y) \log \frac{1}{T(W;Y)}. \tag{118}$$

where $\|W\|_{\psi_2} \triangleq \inf\{c > 0 : \mathbb{E}[e^{W^2/c^2}] \le 2\}$ is an Orlicz norm.

Proposition 16 is reminiscent of Tao's inequality [Tao06, Ahl07] and [WV12, Theorem 10], which use mutual information to produce correlation estimates for bounded random variables: If $\mathsf{var}(W) = 1$, then

$$\rho^2(W, \mathbb{E}[W|Y]) \le \frac{2}{\log \mathrm{e}} \|W\|_\infty^2 I(W;Y).$$

In contrast, Proposition 16 uses T-information in lieu of mutual information and allows more general tail condition.

Remark 10. Combining Proposition 16 with the convergence rate of the T-information in Proposition 11, we obtain the corresponding convergence rate of correlation under various cost constraints on the relays and tail conditions on the original message W. For example, in view of Remark 8, if the cost function is $\mathsf{M}(x) = |x|^p$ and W is sub-Gaussian, then

$$\rho(W, \mathbb{E}[W|Y_n]) = O\left(\frac{\sqrt{\log\log n}}{(\log n)^{p/4}}\right). \tag{119}$$

In particular, for average power constraint ($p = 2$), the convergence rate (116) applies to all sub-Gaussian W. We will show in the next subsection that (119) is in fact optimal for all p when W is Gaussian.

Proof of Proposition 16. Since $T(W;Y) = \mathbb{E}[\mathsf{TV}(P_{W|Y}, P_W)]$ we may construct a probability space with three variables W, W', Y such that $W' \perp\!\!\!\perp Y$ and furthermore

$$\mathbb{P}[W \ne W'] = T(W;Y).$$

Then, consider an arbitrary zero-mean $g(Y)$ and write

$$\mathbb{E}[Wg(Y)] = \mathbb{E}[Wg(Y)] - \mathbb{E}[W'g(Y)] \le \mathbb{E}[|g(Y)| \cdot |W - W'|\mathbf{1}_{\{W \neq W'\}}] \tag{120}$$

$$\le \|g(Y)\|_2 \|W - W'\|_{2q} T(W;Y)^{\frac{1}{2q'}}\,, \tag{121}$$

where the last step is by Hölder's inequality since $\frac{1}{2} + \frac{1}{2q} + \frac{1}{2q'} = 1$ and $q' = \frac{q}{q-1}$. Since $\|W - W'\|_{2q} \le 2\|W\|_{2q}$, normalizing both sides of (121) by $\|g(Y)\|_2$ and $\|W\|_2 = 1$ yields the desired (117).

For the second part of the proposition, consider arbitrary non-negative, convex $\psi : \mathbb{R} \to \mathbb{R}$ with $\psi(0) < 1$ and define the following Orlicz norm

$$\|X\|_\psi \triangleq \inf\{c > 0 : \mathbb{E}[\psi(X/c)] \le 1\}\,.$$

If ψ^* is the Legendre dual of ψ then from Young's inequality we have for arbitrary X, Y:

$$XY \le \psi(X) + \psi^*(Y),$$

and, hence,

$$\mathbb{E}[XY] \le 2\|X\|_\psi \|Y\|_{\psi^*} \tag{122}$$

Consider

$$\psi_1(x) = \frac{1}{2}e^x, \qquad \psi_1^*(y) = y \ln \frac{2y}{e},$$
$$\psi_2(y) = \frac{1}{2}e^{x^2},$$

and notice an easy identity

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2. \tag{123}$$

Then, proceeding as above we only need to upper-bound $\mathbb{E}[|W - W'|^2\mathbf{1}_{\{W \neq W'\}}]$ in (120). From inequality (122) and (123) we get

$$\mathbb{E}[|W - W'|^2\mathbf{1}_{\{W \neq W'\}}] \le 2\|W - W'\|_{\psi_2}^2 \|\mathbf{1}_{\{W \neq W'\}}\|_{\psi_1^*}\,.$$

For the first term we apply triangle inequality. The ψ_1^*-norm of the indicator is found as a unique solution of

$$c = \mathbb{P}[W \neq W'] \ln \frac{2}{ec}\,,$$

with $c \in (0, \frac{2}{e})$. It is easy to show that if $\mathbb{P}[W \neq W'] < e^{-2/e}$ then

$$\|\mathbf{1}_{\{W \neq W'\}}\|_{\psi_1^*} \le \mathbb{P}[W \neq W'] \ln \frac{1}{\mathbb{P}[W \neq W']}\,,$$

from which the proposition follows. □

4.5 Achievable schemes

For the scalar case we construct a relay scheme under which the T-information, mutual information and the correlation between the initial message $W \sim \mathcal{N}(0,1)$ and the final output Y_n achieve the lower bounds (8) – (10) up to constants. This scheme is also useful for the optimal control problem in Section 5.1. For simplicity we only consider the p^{th} moment constraint $\mathbb{E}|X_k|^p \le a$ and assume $W \sim \mathcal{N}(0,1)$ and $a = 2$ for notational conciseness.

Binary-messaging scheme In view of the converse results in Sections 4.1 – 4.4, the majority of the information will be inevitably lost regardless of the relay design. Thus we only aim to transmit a small fraction of the original message, e.g., a highly skewed quantized version, reliably. To this end, let

$$\mu = 4\sqrt{\log n}, \quad a = \mathsf{Q}^{-1}(\mu^{-p}) = \sqrt{p \log\log n} + o(1). \tag{124}$$

Let $X_1 = \mu \mathbf{1}_{\{W \geq a\}}$, which satisfies $\mathbb{E}|X_1|^p = 1$. At each stage, the relay decodes the previous message by $X_{k+1} = \mu \mathbf{1}_{\{Y_k \geq \mu/2\}}$. Note that all X_k's take values in $\{0, \mu\}$. Then $\mathbb{P}[X_{k+1} \neq X_k] \leq \mathbb{P}[|Z_k| \geq \mu/2] = 2\mathsf{Q}(\mu/2)$. For any $k \in [n+1]$, applying the union bound and the fact that $\mathsf{Q}(a) \leq \varphi(a)/a$, we obtain

$$\mathbb{P}[X_k \neq X_1] \leq 2n\mathsf{Q}(\mu/2) \leq n^{-1}. \tag{125}$$

Moreover, the moment constraint is satisfied since

$$\mathbb{E}|X_k|^p = \mu^p \mathbb{P}[X_k \neq 0] \leq \mu^p (\mathbb{P}[X_1 \neq 0] + \mathbb{P}[X_k \neq X_1]) \leq 1 + \frac{1}{n}(16 \log n)^{p/2} \leq 2$$

for all sufficiently large n.

Total variation and Mutual information We show that

$$T(W; Y_n) = \Omega\left(\frac{1}{(\log n)^{p/2}}\right), \tag{126}$$

$$I(W; Y_n) = \Omega\left(\frac{\log\log n}{(\log n)^{p/2}}\right), \tag{127}$$

which matches the upper bound in Remark 8 and the upper bound (98) (for $p = 2$), respectively. Since X_1 and X_{n+1} are deterministic functions of W and Y_n, respectively, we have $X_1 \to W \to Y_n \to X_{n+1}$ and

$$T(W; Y_n) \geq T(X_1; X_{n+1}) \geq \mu^{-p} - \mathbb{P}[X_{n+1} \neq X_1] = \Omega((\log n)^{-p/2}),$$

where the first inequality follows from data processing, the second inequality follows from (82), and the last inequality is by (125). Similarly,

$$I(W; Y_n) \geq I(X_1; X_{n+1}) = H(X_1) - H(X_1|X_{n+1}) \geq h(\mu^{-p/2}) - h(1/n) = \Omega(\mu^{-p/2} \log \mu).$$

Correlation Denote $B = \mathbf{1}_{\{W \geq a\}} = X_1/\mu$ and $\hat{W} = \mathbb{E}[W|B] = g(B)$, where

$$g(0) = \mathbb{E}[W|W \leq a] = -\frac{\varphi(a)}{\Phi(a)}, \quad g(1) = \mathbb{E}[W|W > a] = \frac{\varphi(a)}{\mathsf{Q}(a)}. \tag{128}$$

Using the fact that $\mathsf{Q}(x) = \frac{\varphi(x)}{x}(1 + o(1))$ as $x \to \infty$, we have

$$\mathbb{E}[W\hat{W}] = \mathbb{E}[\hat{W}^2] = \frac{\varphi^2(a)}{\Phi(a)\mathsf{Q}(a)} = \mathsf{Q}(a)a^2(1 + o(1)) = \Theta\left(\frac{\log\log n}{(\log n)^{p/2}}\right), \tag{129}$$

where the last inequality follows from the choice of a in (124).

Set $B_n = \mathbf{1}_{\{Y_n \geq \mu/2\}} = X_{n+1}/\mu$ and $W_n = g(B_n)$. By (125), we have $\mathbb{P}[B \neq B_n] \leq \frac{1}{n}$. Therefore

$$\begin{aligned}
\mathbb{E}[WW_n] &= \mathbb{E}[W\hat{W}] + \mathbb{E}[W(W_n - \hat{W})\mathbf{1}_{\{B\neq B_n\}}] \\
&\geq \mathbb{E}[W\hat{W}] - \max\{|g(0)|, |g(1)|\}\mathbb{E}[|W|\mathbf{1}_{\{B\neq B_n\}}] \\
&\geq \mathbb{E}[W\hat{W}] - g(1)\sqrt{\mathbb{P}[B \neq B_n]} \qquad (130) \\
&= \Theta\left(\frac{\log\log n}{(\log n)^{p/2}}\right), \qquad (131)
\end{aligned}$$

where (130) is by Cauchy-Schwartz, (131) is by (129) and $g(1) = a(1 + o(1)) = \Theta(\sqrt{\log\log n})$. Similarly,

$$|\mathbb{E}[W_n^2] - \mathbb{E}[\hat{W}^2]| = |\mathbb{E}[(W_n^2 - \hat{W}^2)\mathbf{1}_{\{B\neq B_n\}}]| \leq g(1)^2\mathbb{P}[B \neq B_n] = O(\log\log n/n). \qquad (132)$$

Therefore $\|W_n\|_2 = \|\hat{W}\|_2(1 + o(1))$. Consequently, the correlation satisfies

$$\rho(W, \mathbb{E}[W|Y_n]) = \sup_{g\in L_2(P_{Y_n})} \rho(W, g(Y_n)) \geq \rho(W, W_n) = \frac{\mathbb{E}[WW_n]}{\|W_n\|_2} = \Omega\left(\frac{\sqrt{\log\log n}}{(\log n)^{p/4}}\right), \qquad (133)$$

which meets the upper bound (119).

5 Applications

5.1 Optimal memoryless control in Gaussian noise

The problem of optimal memoryless control in Gaussian noise was investigated in [LM11]. Consider the n-stage stochastic control problem in Fig. 1 in one dimension ($d = 1$) where the input $W = X_0 + Z_0$ with $X_0 \sim \mathcal{N}(0, \sigma_0^2)$ independent of $Z_0 \sim \mathcal{N}(0, 1)$. The additive noise $Z_1, \ldots, Z_n$ are i.i.d. standard Gaussian, and the relay function f_j plays the role of a memoryless controller mapping the noisy observation Y_{i-1} into a control signal X_i. Let $X_{n+1} = f_{n+1}(Y_n)$ denote the final estimate. Then we have the following Markov chain which has two more stages than (1):

$$X_0 \to W \to X_1 \to Y_1 \to X_2 \to Y_2 \to \cdots \to X_n \to Y_n \to X_{n+1}.$$

The major difference is that, instead of requiring that each controller satisfies the same power constraint as in (3), here only a *total* power budget is imposed:

$$\sum_{j=1}^{n} \mathbb{E}[X_j^2] \leq nE\,. \qquad (134)$$

The objective is to maximize the correlation between X_0 and X_{n+1}.

The main results of [LM11] show that although linear controllers are optimal for two stages ($n = 1$) [LM11, Proposition 7], for multiple stages they can be strictly sub-optimal. Specifically, subject to the constraint (134), the optimal squared correlation $\rho^2(X_0, X_{n+1})$ achieved by linear controllers is [LM11, Lemma 6]

$$\frac{\sigma_0^2}{1+\sigma_0^2}\left(\frac{E}{1+E}\right)^n, \qquad (135)$$

which vanishes exponentially as $n \to \infty$. [LM11, Theorem 15] shows that (135) can be improved by using binary quantizers in certain regimes, although the correlation still vanishes exponentially

fast albeit with a better exponent. The optimal performance of non-linear controllers is left open in [LM11].

Capitalizing on the results developed in Section 4, next we show that the squared correlation achieved by the best non-linear controllers is $\Theta(\frac{\log\log n}{\log n})$, which is significantly better than the exponentially small correlation (135) achieved by the best linear controllers.

- For any sequence $\{f_j\}$ satisfying the total power constraint (134), the correlation necessarily satisfies
$$\rho^2(X_0, X_{n+1}) = O\left(\frac{\log\log n}{\log n}\right). \tag{136}$$
To see this, applying the data processing inequality Theorem 10 and the F_{TV} curve in Corollary 6 with $\mathsf{M}(|x|) = |x|^2$, we have
$$T(X_0; X_{n+1}) \le T(W; Y_n) \le F_1 \circ \cdots \circ F_n(1),$$
where $F_i(t) = t(1-2\mathsf{Q}(\sqrt{a_i/t}))$ and $a_i = \mathbb{E}[X_i^2]$. Since $\sum_{i=1}^n a_i \le nE$, we have $\sum_{i=1}^n \mathbf{1}_{\{a_i \ge 2E\}} \le n/2$. Consequently, Proposition 11 applies with n replaced by $n/2$ and, by Remark 8, we have $T(W; Y_n) \le C/\log n$ for some constant C only depending on E. Since X_0 is Gaussian, applying Proposition 16 yields the upper bound (136).

- Conversely, the binary-quantizer scheme described in Section 4.5 (with $p = 2$) achieves
$$\rho^2(X_0, X_{n+1}) = \Omega\left(\frac{\log\log n}{\log n}\right).$$
Set $X_{n+1} = W_n = g(\mathbf{1}_{\{Y_n \ge \mu/2\}})$, where g is defined in (128). Since $W = X_0 + Z_0$ and $V \triangleq X_0 - \sigma_0^2 Z_0$ are independent, we have $\mathbb{E}[X_0 X_{n+1}] = \frac{\sigma_0^2}{1+\sigma_0^2}\mathbb{E}[W W_n]$ and the rest follows from (133).

The fact that linear control only achieves exponentially decaying correlation can also be understood from the perspective of contraction coefficient of KL divergence. Note that if all controllers are linear, then all input X_i's to the AWGN channel are Gaussian. Recall the distribution-dependent contraction coefficient $\eta_{\mathrm{KL}}(Q)$ defined in (13). For AWGN channel with noise variance σ^2 and Gaussian input with variance P, Erkip and Cover showed in [EC98, Theorem 7] that $\eta_{\mathrm{KL}}(\mathcal{N}(\mu, P)) = \frac{P}{P+\sigma^2}$, which is strictly less than one. This results in exponentially small mutual information:

$$\begin{aligned} I(W; \hat{W}) &\le I(W; Y_1) \prod_{i=2}^n \eta_{\mathrm{KL}}(\mathcal{N}(\mathbb{E}[X_i], \mathsf{var}(X_i))) \\ &\le \frac{\log(1+\sigma_0^2)}{2} \prod_{i=2}^n \frac{\mathsf{var}(X_i)}{1+\mathsf{var}(X_i)} \le \frac{\log(1+\sigma_0^2)}{2} \left(\frac{E}{1+E}\right)^{n-1}, \end{aligned}$$

where the last step follows from (134) and the concavity and monotonicity of $x \mapsto \log\frac{x}{x+1}$. Together with the Gaussian rate-distortion function (113), this implies $\rho(W, \hat{W})$ must vanish as $(\frac{E}{1+E})^n$ which agrees with (135). Therefore from a control-theoretic perspective, it is advantageous to design the controller to steer the output away from Gaussian, which requires, of course, non-linear control.

5.2 Uniqueness of Gibbs measures

In this section we rely on the notations and results from the theory of infinite-volume Gibbs measures; in particular we assume familiarity with [Geo11, Chapter 2]. Consider a $\mathbb{R}$-valued Markov random field $\{X_n : n \in \mathbb{Z}\}$ specified by pairwise potentials $\Phi_j(x_j, x_{j+1})$. We assume that for every $k \in \mathbb{Z}$ and every $L \geq 1$ we have

$$\int \exp\left\{-\sum_{j=k}^{k+L} \Phi_j(x_j, x_{j+1})\right\} dx_k \cdots dx_{k+L} < \infty\,.$$

This specification translates into requiring the conditional probabilities to be of the following form:

$$P_{X_{k+1}^{k+L}|X_{-\infty}^{k}, X_{k+L+1}^{\infty}} \propto \exp\left\{-\sum_{j=k}^{k+L} \Phi_j(x_j, x_{j+1})\right\} dx_k \cdots dx_{k+L}\,, \tag{137}$$

and in particular X_n form a doubly-infinite Markov chain:

$$\cdots - X_{-1} - X_0 - X_1 - \cdots \tag{138}$$

One of the principal questions in Gibbs theory is: Do there exist none, one or many joint distributions satisfying conditional probabilities (137)? Such a joint distribution is called a Gibbs measure consistent with the specification (137). It is believed that the existence of multiple Gibbs measures corresponds to the existence of second-order phase transitions in physics (such as the Curie temperature in ferromagnets).

A typical method for proving non-existence of multiple phases is the application of Dobrushin contraction, cf. [Dob70]. Next we extend this technique to cases where Dobrushin contraction is not available ($\eta_{\mathrm{TV}} = 1$) by relying on the knowledge of the Dobrushin curve F_{TV}. Here is an illustration.

Theorem 17. *Suppose that potentials Φ_j are such that each conditional distribution (137) factors through the Gaussian channel, i.e. for each k, L there exists a representation*

$$P_{X_{k+1}^{k+L}|X_{-\infty}^{k}, X_{k+L+1}^{\infty}} = P_{X_{k+1}^{k+L}|Y} \circ P_{Y|X_k, X_{k+L+1}}\,, \tag{139}$$

with $P_{Y|X_k, X_{k+L+1}}$ a two-dimensional Gaussian channel (2). Then there may exist at most one joint distribution of $X_{-\infty}^{\infty}$ satisfying

$$\sup_{j \in \mathbb{Z}} \mathbb{E}[|X_j|^2] < \infty. \tag{140}$$

Remark 11. Assumptions of Theorem 17 guarantee that "strengths" of all links in (138) are uniformly upper-bounded. Thus we can see that on $\mathbb{Z}$ the only possibilities for a phase transition are: 1) when the links become asymptotically noiseless, or 2) when the (non shift-invariant) solutions are allowed to grow unbounded. This is in accord with known examples of systems with non-unique Gibbs measures: e.g., the asymptotically noiseless example in [Geo11, Chapter 6], or the non shift-invariant examples of Spitzer-Cox and Kalikow in [Geo11, Chapter 11].

Proof. We recall the following idea due to Dobrushin [Dob70, Lemma 5]:

Proposition 18. *Let π be any coupling of P_{AB} to Q_{AB} (i.e. $\pi_{ABA'B'}$ is P_{AB} or Q_{AB} when restricted to first pair or second pair). Assume also that for every a and a' we have*[10]

$$W_\rho(P_{B|A=a}, Q_{B|A=a'}) \le r(a,a')\,.$$

Then there exists a coupling $\tilde\pi$ between P_{AB} and Q_{AB} such that $\tilde\pi_{A,A'} = \pi_{A,A'}$ and

$$\mathbb{E}_{\tilde\pi}[\rho(B,B')] \le \mathbb{E}_\pi[r(A,A')].$$

When $r(a,a') = c\rho(a,a')$ and $c<1$ (Dobrushin contraction), we can progressively refine the coupling at various points between two distributions P and Q and show that they must coincide. This is a brilliant idea of Dobrushin [Dob70]. We apply the same recursion here, except without relying on $c<1$.

Suppose that there exist two distributions P and Q of $X_{-\infty}^{\infty}$ satisfying (139) and (140). Let $E>0$ denote the left-hand side of (140), i.e., the common upper bound on the second moment of X_j. Given a coupling π between P and Q, that is

$$\pi_{X_k^n} = P_{X_k^n}, \quad \pi_{\tilde X_k^n} = Q_{X_k^n}, \quad k \le n$$

denote

$$\epsilon_N = \pi[X_{-N}^N \ne \tilde X_{-N}^N] \le 1,$$

where $N \ge 1$ is large integer.

Denote $x_{\pm N} = (x_N, x_{-N})$ and $|x_{\pm N}| = \sqrt{x_N^2 + x_{-N}^2}$ its Euclidean norm. Using the factorization condition (139) and the data processing inequality for total variation, we have

$$\begin{aligned}
&\mathsf{TV}\left(P_{X_{-N+1}^{N-1}|X_{\pm N}=a_{\pm N}}, Q_{X_{-N+1}^{N-1}|X_{\pm N}=b_{\pm N}}\right)\\
&\le \mathsf{TV}(P_{Y|X_{\pm N}=a_{\pm N}}, Q_{Y|X_{\pm N}=b_{\pm N}}) = \mathsf{TV}(\mathcal{N}(a_{\pm N}, \mathbf{I}_2), \mathcal{N}(b_{\pm N}, \mathbf{I}_2))\\
&= \theta_c\left(|a_{\pm N} - b_{\pm N}|\right)\,,
\end{aligned}$$

where $\theta_c(u) = 1 - 2\mathsf{Q}(u/2)$, cf. Corollary 6. Applying Proposition 18 with $\rho(a,a') = \mathbf{1}_{\{a\ne a'\}}$ and $r(a,a') = \theta_c(|a-a'|)$, we can produce a new coupling π' so that $\pi'_{X_{\pm N}, \tilde X_{\pm N}} = \pi_{X_{\pm N}, \tilde X_{\pm N}}$ and

$$\pi'[X_{-N+1}^{N-1} \ne \tilde X_{-N+1}^{N-1}] \le \mathbb{E}_\pi[\theta_c(|X_{\pm N} - \tilde X_{\pm N}|)].$$

In view of the moment constraint (140), we have

$$\mathbb{E}_\pi[|X_{\pm N}|^2 + |\tilde X_{\pm N}|^2] = \mathbb{E}_P[|X_{-N}|^2 + |X_N|^2] + \mathbb{E}_Q[|X_{-N}|^2 + |X_N|^2] \le 4E. \tag{141}$$

Thus, as we noticed in the proof of Theorem 4, the constraint (141) leads to

$$\mathbb{E}_\pi[\theta_c(|X_{\pm N} - \tilde X_{\pm N}|)] \le f(\pi[X_{\pm N} \ne \tilde X_{\pm N}]) \le f(\epsilon_N),$$

where the concave non-decreasing function f is

$$f(t) = t\theta_c\left(\sqrt{\frac{8E}{t}}\right)\,.$$

[10] Here W_ρ is a Wasserstein distance with respect to the metric ρ, analogously defined as in (59) with the L_1 distance replaced by ρ.

Therefore, starting from any coupling π which achieves ϵ_N we produced a new coupling π' which achieves

$$\epsilon_{N-1} \le f(\epsilon_N)\,.$$

As we have seen in the proof of Proposition 11, Lemma 20 shows that such iterations lead to ϵ_N decreasing to zero. Hence for any n, starting with sufficiently large $N \gg n$, we have shown that $\mathsf{TV}(P_{X^n_{-n}}, Q_{X^n_{-n}})$ is arbitrarily small, hence zero. In other words, distributions P and Q have the same finite-dimensional marginals, and must therefore coincide. □

As one can see our proof crucially relies on the fact that boundary of the interval $[-N, N]$ on the chain graph (138) always consists of two points $X_{\pm N}$ (see (141)). This is why a similar argument is not applicable to Markov random fields on $\mathbb{Z}^2$, where the number of variables in the boundary of $[-N, N]^2$ grows with N. But in that case it is well-known that even for binary-valued X there can exist multiple Gibbs measures (the two-dimensional Ising model example).

5.3 Circuits of noisy gates

A circuit is a directed acyclic graph emanating from n inputs $X_1, \dots X_n$, going through multiple intermediate nodes ("gates") and terminating at a final node W. Each gate i with inputs $S_i = (S_{i,1}, \dots, S_{i,k})$ performs a simple operation $f_i(S_i)$ and produces an output, which is then subjected to additive Gaussian noise, so that the output value O_i of the i^{th} gate is given by

$$O_i = f_i(S_i) + Z_i\,, \qquad Z_i \sim \mathcal{N}(0,1)\,. \tag{142}$$

The outputs of the i^{th} gate are connected to the inputs S_j of subsequent gates according to the graph. The value of W is the output of the last gate.

We say that the circuit computes the Boolean function $F : \{0,1\}^n \to \{0,1\}$ with probability of error ϵ if

$$\mathbb{P}[F(x_1, \dots, x_n) = g(W)] \ge 1 - \epsilon\,,$$

for some $g : \mathbb{R} \to \{0,1\}$ and all binary vectors x^n. We assume that all gates have at most k inputs. We say that the function F depends essentially on input x_i if there exist $x, x' \in \mathbb{R}^n$ differing in the i^{th} coordinate *only*, such that

$$F(x) \ne F(x')\,.$$

We show below that it is not possible to have small ϵ, complicated F, large n and small power consumed by outputs of each gate:

$$\mathbb{E}[|f_i(S_i)|^2] \le P\,. \tag{143}$$

This is a natural extension of the well-studied model of binary symmetric noise (bit flips) [vN56, Pip88]. We note that even for the settings of binary symmetric channels (BSC), quite a few open questions remain. For example, it is known that for each k there exists a threshold of maximum tolerable noise beyond which arbitrarily complex circuits are not possible [HW91, ES03]. However, this threshold is generally unknown and is sensitive to whether BSCs have crossover probability exactly δ or $\le \delta$, cf. [Ung10], and whether the output of one gate is allowed to be used at one or multiple consequent gates, cf. [ES03].

Proposition 19. *For any signal-to-noise ratio $P > 0$, any Boolean function F essentially depending on n inputs, and any circuits of noisy k-input gates computing F, the probability of error satisfies*

$$\epsilon \ge \frac{1 - t_k^*}{2} + o(1)\,, \qquad n \to \infty\,, \tag{144}$$

where

$$t_k^* = \sup\{t : F_{\mathsf{TV}}(kt \wedge 1) \geq t, 0 \leq t \leq 1\}$$

and $F_{\mathsf{TV}}(t)$ *is given by* (58) *with* $a = P$.

For three-input gates, the lower bound (144) is evaluated in Fig. 4 as a function of P.

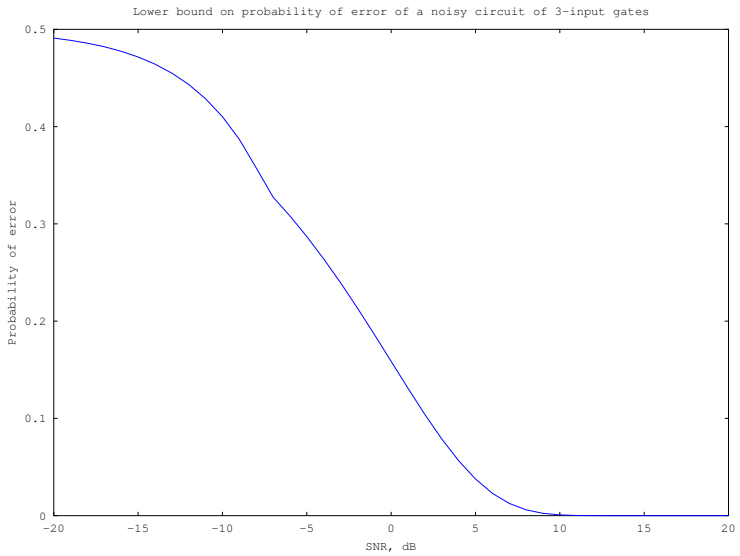


Figure 4: Lower bound of probability of erroneous computation (144) versus signal-to-noise ratio for $k = 3$.

Proof. We recall a combinatorial fact shown in the proof of [ES99, Theorem 2]: For every Boolean function F essentially depending on n inputs, and for every circuit that computes F with probability of error strictly less than $\frac{1}{2}$, there must exist at least one input, say X_1, such that *every* path from X_1 to W has length at least

$$\ell \geq \frac{\log n}{\log k}\,. \tag{145}$$

Since F essentially depends on X_1, we can assume, without loss of generality, that

$$F(0, 0, \ldots, 0) \neq F(1, 0, \ldots, 0).$$

Note that the random variables in the circuit consist of the inputs $X = (X_i)$, inputs $S = (S_i)$ and outputs $O = (O_i)$ of the gates, and the final output W, which is equal to some O_i. To simplify notation, let $O_0 = X_1$. Denote the neighbors of the gate i by

$$N_i = \{j \geq 0 : O_j \in S_i\},$$

whose outputs serve as inputs to gate i. Then $|N_i| \leq k$ by assumption. Without loss of generality, we assume that all gates are numbered so that i^{th} gate's inputs all come from gates with indices strictly less than i. Then $N_i \subset \{0, \ldots, i-1\}$ by construction.

Consider now probability distributions P and Q of all random variables in the circuit, such that under P we have $X_1 = 0$ and under Q we have $X_1 = 1$, while $X_2 = \cdots = X_n = 0$ under both. The idea is to progressively build coupling between P and Q to show that

$$\mathsf{TV}(P_W, Q_W) \leq t_k^* + o(1)\,, \tag{146}$$

from which the desired lower bound (144) follows.

To prove (146), suppose that there is a joint distribution π such that

$$(X, O, S, W) \sim P, \quad (X', O', S', W') \sim Q,$$

i.e. π is a coupling of P to Q. Consider an arbitrary gate i with input S_i and output O_i. In view of the noise model (142), the proof of Theorem 17 shows that the moment constraint (143) enables us to use Proposition 18 to build another coupling $\tilde{\pi}$, such that a) $(X, X', O_{<i}, O'_{<i}, S_{\le i}, S'_{\le i})$ have identical joint distribution under either π or $\tilde{\pi}$, and b) at the i^{th} gate we have

$$\tilde{\pi}[O_i \neq O'_i] \le F_{\mathsf{TV}}(\pi[S_i \neq S'_i])\,. \tag{147}$$

Recall that $X_2 = X'_2 = \ldots = X_n = X'_n = 0$ under π. Then S_i is determined by the outputs of the neighboring gates and possibly $O_0 = X_1$, collectively denoted by $\{O_j : j \in N_i\}$. By the union bound, we have

$$\pi[S_i \neq S'_i] \le k \max_{j \in N_i} \pi[O_j \neq O'_j]\,.$$

So if we introduce the function

$$F_k(t) \triangleq F_{\mathsf{TV}}(kt \wedge 1),$$

then we can relax (147) to

$$\tilde{\pi}[O_i \neq O'_i] \le F_k\Big(\max_{j \in N_i} \pi[O_j \neq O'_j]\Big). \tag{148}$$

Now, let π_0 be the trivial (independent) coupling. Since $X_1 = 1$ and $X'_1 = 0$ under π_0, we have $\pi_0[O_0 \neq O'_0] = 1 \triangleq t_0$. Consider the first gate, whose inputs can be either X_1 or constants. Applying the previous construction yields a coupling π_1 such that

$$t_1 \triangleq \pi_1[O_1 \neq O'_1] \le F_k(1)\,.$$

Here t_1 measures the quality of coupling at the output of the first gate. Next, suppose that all gates $< i$ are similarly coupled by π_{i-1} with respective $t_1, \ldots, t_{i-1}$. We refine the coupling at gate i to get π_i, so that a) the joint distribution of $(O_{<i}, O'_{<i})$ and hence $t_1, \ldots, t_{i-1}$ are unchanged, and b)

$$t_i \triangleq \pi_i[O_i \neq O'_i] \le F_k\Big(\max_{j \in N_i} t_j\Big)\,, \tag{149}$$

which follows from (148). Continuing similarly, we arrive at the last gate which outputs W. Now let us construct a path from W back to $X_1 = O_0$ as follows: starting from W go back from gate i to the neighboring gate $j < i$ that achieves $\max_{j \in N_i} t_j$. Let m be the length of this path and let the indices (in increasing order) be

$$i_0 = 0 < i_1 = 1 < i_2 < \ldots < i_m.$$

By (145) we must have $m = \Omega(\log n)$. By construction of the path, we have $t_0 = 1, t_1 \le F_k(1)$, $t_{i_2} \le F_k(t_1)$, etc. So finally

$$\pi[W \neq W'] \le t_{i_m} \le F_k\Big(\max_{j \in N_i} t_j\Big) = F_k(t_{i_{m-1}}) \le \ldots \le \underbrace{F_k \circ F_k \cdots \circ F_k}_{m \text{ times}}(1).$$

Hence as $n \to \infty$ this repeated composition of F_k's must converge to a fixed point t_k^*, thus proving (146). □

5.4 Broadcasting on trees

Consider the setting studied in [EKPS00]: the original bit $W = \pm 1$ is to be broadcasted along the binary tree of noisy channels:

$$W \to X_{1,1} \begin{cases} \nearrow Y_{2,2} \to X_{2,2} \begin{cases} \nearrow \\ \searrow \end{cases} \cdots \\ \searrow Y_{2,1} \to X_{2,1} \begin{cases} \nearrow \\ \searrow \end{cases} \cdots \end{cases}$$

where arrows $X \to Y$ represent independent noisy channels and $Y \to X$ are relays. The goal is to design the relay functions so that for some $\epsilon > 0$ one can reconstruct W with probability of error at most $\frac{1}{2} - \epsilon$ based on the values at the n^{th} layer $\{X_{n,1}, \ldots, X_{n,2^{n-1}}\}$ for all sufficiently large n; to wit, the total variation of the distributions conditioned on $W = 1$ or -1 is strictly bounded away from zero. One of the main results of [EKPS00] is that when all channels are BSC with flip probability δ such broadcasting is possible if and only if $2(1-2\delta)^2 > 1$, thus establishing a certain "phase transition" in this problem.

In fact, the impossibility part of the BSC result follows from a result of Evans and Schulman [ES99]: for a binary tree of discrete channels the probability of error tends to $\frac{1}{2}$ as the depth tends to infinity whenever $2\eta_{\text{KL}} < 1$. For Gaussian channels we know that $\eta_{\text{KL}} = 1$ which suggests that such transition *does not occur* for a tree of Gaussian channels. Indeed, in this section we demonstrate that it is possible to broadcast some information to arbitrarily deep layers regardless of how small the SNR is.

Specifically, consider channels

$$Y_{k,j} = X_{k-1,j} + Z_{k,j}\,, \qquad Z_{k,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

with cost constraint

$$\mathbb{E}[|X_{k,j}|^2] \le E\,. \qquad \forall k, j \tag{150}$$

Choose the initial (randomized) encoder as follows:

$$X_{1,1} = \mu B W, \qquad \mathbb{P}[B=1] = 1 - \mathbb{P}[B=0] = 2p, B \perp\!\!\!\perp W,$$

with parameters p, μ to be specified later. Similar to the scheme in Section 4.5, choose relays as follows:

$$X_{k,j} = \begin{cases} +\mu, & Y_{k,j} \ge t\mu\,, \\ 0, & |Y_{k,j}| < t\mu\,, \\ -\mu, & Y_{k,j} \le -t\mu\,, \end{cases}$$

where $t \in (1/2, 1)$ can be set arbitrarily. Notice that if μ is selected so that

$$p = \frac{Q(t\mu)}{Q((1-t)\mu) + 2Q(t\mu) - Q((1+t)\mu)} \tag{151}$$

then a simple computation shows that for all k, j we have

$$\mathbb{P}[X_{k,j} = +\mu] = \mathbb{P}[X_{k,j} = -\mu] = p\,. \tag{152}$$

But from (151) and the fact that $t > 1/2$ for large μ we get

$$p = e^{-\mu^2(t-1/2)+O(1)}\,, \qquad \mu \to \infty.$$

In particular, regardless of how small E in (150) is and for any t, there exists a sufficiently large μ such that the cost constraint is satisfied. Another important parameter turns out to be

$$\theta = 1 - Q((1-t)\mu) - Q((1+t)\mu).$$

Again, taking μ large we may ensure

$$2\theta^2 > 1. \tag{153}$$

Thus we assume from now on that p, μ and t are selected in such a way that both (150) and (153) are satisfied.

Similarly to [EKPS00] we will employ the idea of T. Kamae, see [Hig77, Remark on p. 342], and consider the behavior of "spin sums":

$$S_k = \sum_{j=1}^{2^{k-1}} \sigma_{k,j}\,,$$

where $\sigma_{k,j} \triangleq \mathrm{sign}(X_{k,j})$ with $\mathrm{sign}(0) = 0$, or equivalently, $\sigma_{k,j} = X_{k,j}/\mu$. To show that it is possible to test $W = \pm 1$ based on the statistic S_n, we show that

$$\liminf_{n\to\infty} \mathsf{TV}(P_{S_n|W=+1} \| P_{S_n|W=-1}) \geq 2p\left(1 - \frac{1}{4\theta^2}\right), \tag{154}$$

which is strictly positive. According to [EKPS00, Lemma 4.2 (i) and (iii)] we have:

$$\mathsf{TV}(P_{S_n|W=+1} \| P_{S_n|W=-1}) \geq \frac{(\mathbb{E}[S_n|W=+1] - \mathbb{E}[S_n|W=-1])^2}{4\mathbb{E}[S_n^2]}.$$

So the estimate (154) follows from two results:

$$\mathbb{E}[S_n|W=\pm 1] = \pm 2p(2\theta)^{n-1}, \tag{155}$$

$$\mathbb{E}[S_n^2] \leq 2^n p + 2p\frac{(2\theta)^{2n}}{(2\theta)^2 - 1}. \tag{156}$$

Both of these are verified below: Consider two arbitrary nodes (k,j) and (k,j') at the k^{th} level and let (u,i) be their common ancestor in the tree. Denote the parent node of (k,j) by $(k-1,j'')$. Then

$$\mathbb{E}[\sigma_{k,j}|\sigma_{u,j'}] = \mathbb{E}[\mathbb{E}[\sigma_{k,j}|\sigma_{k-1,j''}]|\sigma_{u,j'}] = \theta\,\mathbb{E}[\sigma_{k-1,j''}|\sigma_{u,j'}] = \ldots = \theta^{k-u}\sigma_{u,i}.$$

Furthermore, $\sigma_{k,j}$ and $\sigma_{k,j'}$ are independent conditioned on $\sigma_{u,i}$. Note that $\mathbb{E}[\sigma_{1,1}|W=\pm 1] = \pm\mathbb{P}[B=1] = \pm 2p$, which yields (155). Next, note that $\mathbb{E}[S_n^2] = \sum_{j=1}^{2^{k-1}} \mathbb{E}[\sigma_{n,j}^2] + 2\sum_{j'<j} \mathbb{E}[\sigma_{n,j}\sigma_{n,j'}]$, where the first term is $2^{n-1} \times 2p$ since $\sigma_{k,j}^2 \sim \mathrm{Bern}(2p)$ in view of (152). To estimate the cross term, denote the depth of the common ancestor of (n,j) and (n,j') by $u(j,j') \in \{1,\ldots,n-1\}$. Then

$$\sum_{j'<j} \mathbb{E}[\sigma_{n,j}\sigma_{n,j'}] = \sum_{u=1}^{n-1} \sum_{u(j',j)=u} \mathbb{E}[\sigma_{n,j}\sigma_{n,j'}] = \sum_{u=1}^{n-1} \sum_{u(j',j)=u} \theta^{2(n-u)} 2p \tag{157}$$

$$= 2p\sum_{u=1}^{n-1} \theta^{2(n-u)} \binom{2^{n-u}}{2} \leq p\frac{(2\theta)^{2n}}{(2\theta)^2 - 1}, \tag{158}$$

which yields (156).

Acknowledgment

It is a pleasure to thank Max Raginsky (UIUC) for many helpful discussions and Flavio du Pin Calmon (MIT) for Proposition 12.

Appendix A Convergence rate analysis

Consider the following iteration

$$t_{n+1} = t_n - h(t_n), \quad t_1 = 1$$

where $h : [0,1] \to [0,1]$ satisfies $h(0) = 0$ and $0 < h(t) \le t$ for all $0 < t \le 1$. Then $\{t_n\} \subset [0,1]$ a monotonically decreasing sequence converging to the unique fixed point zero as $n \to \infty$. Under the monotonicity assumption of the function h, the following result gives a non-asymptotic upper estimate of this sequence.

Lemma 20. *Define $G : [0,1] \to \mathbb{R}_+$ by $G(t) = \int_t^1 \frac{1}{h(\tau)} \mathrm{d}\tau$. If h is increasing, then for any $n \in \mathbb{N}$,*

$$t_n \le G^{-1}(n-1). \tag{159}$$

Proof. By the positivity and monotonicity of h, G is a strictly decreasing and concave function. Hence $G^{-1} : \mathbb{R}_+ \to [0,1]$ is well-defined. Put $b_n = G(t_n)$. Then

$$b_n - b_{n-1} = \int_{t_n}^{t_{n-1}} \frac{1}{h(\tau)} \mathrm{d}\tau \ge \frac{t_{n-1} - t_n}{h(t_{n-1})} = 1. \tag{160}$$

Hence $b_n \ge n - 1$ since $b_1 = G(1) = 0$. □

Appendix B Contraction coefficient for mutual information: General case

We shall assume that P_X is not a point mass, namely, there exists a measurable set E such that $P_X(E) \in (0,1)$. Define

$$\eta_{\mathrm{KL}}(P_X) = \sup_{Q_X} \frac{D(Q_Y \| P_Y)}{D(Q_X \| P_X)}$$

where the supremum is over all Q_X such that $0 < D(Q_X \| P_X) < \infty$. It is clear that such Q_X always exists (e.g., $Q_X = P_{X|X \in E}$ and $D(Q_X \| P_X) = \log \frac{1}{P_X(E)} \in (0, \infty)$). Let

$$\eta_I(P_X) = \sup \frac{I(U;Y)}{I(U;X)}$$

where the supremum is over all Markov chains $U \to X \to Y$ with fixed P_{XY} such that $0 < I(U;X) < \infty$. Such Markov chains always exist, e.g., $U = \mathbf{1}_{\{X \in E\}}$ and then $I(U;X) = h(P_X(E)) \in (0, \log 2)$. The goal of this appendix is to prove (22), namely

$$\eta_{\mathrm{KL}}(P_X) = \eta_I(P_X)\,.$$

The inequality $\eta_I(P_X) \le \eta_{\mathrm{KL}}(P_X)$ follows trivially:

$$I(U;Y) = D(P_{Y|U}\|P_Y|P_U) \le \eta_{\mathrm{KL}}(P_X) D(P_{X|U}\|P_X|P_U) = \eta_{\mathrm{KL}}(P_X) I(X;U)\,.$$

For the other direction, fix Q_X such that $0 < D(Q_X\|P_X) < \infty$. First, consider the case where $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ is bounded, namely, $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X} \le a$ for some $a > 0$ Q_X-a.s. For any $\epsilon \le \frac{1}{2a}$, let $U \sim \mathrm{Bern}(\epsilon)$ and define the probability measure $\tilde{P}_X = \frac{P_X - \epsilon Q_X}{1-\epsilon}$. Let $P_{X|U=0} = \tilde{P}_X$ and $P_{X|U=1} = Q_X$, which defines a Markov chain $U \to X \to Y$ such that X, Y is distributed as the desired P_{XY}. Note that

$$\frac{I(U;Y)}{I(U;X)} = \frac{\bar{\epsilon} D(\tilde{P}_Y\|P_Y) + \epsilon D(Q_Y\|P_Y)}{\bar{\epsilon} D(\tilde{P}_X\|P_X) + \epsilon D(Q_X\|P_X)}$$

where $\tilde{P}_Y = P_{Y|X} \circ \tilde{P}_X$. We claim that

$$D(\tilde{P}_X\|P_X) = o(\epsilon), \tag{161}$$

which, in view of the data processing inequality $D(\tilde{P}_X\|P_X) \le D(\tilde{P}_Y\|P_Y)$, implies $\frac{I(U;Y)}{I(U;X)} \xrightarrow{\epsilon\downarrow 0} \frac{D(Q_Y\|P_Y)}{D(Q_X\|P_X)}$ as desired. To establish (161), define the function

$$f(x,\epsilon) \triangleq \begin{cases} \frac{1-\epsilon x}{\epsilon(1-\epsilon)} \log \frac{1-\epsilon x}{1-\epsilon}\,, & \epsilon > 0 \\ (x-1)\log e, & \epsilon = 0\,. \end{cases}$$

One easily notices that f is continuous on $[0,a] \times [0, \frac{1}{2a}]$ and thus bounded. So we get, by bounded convergence theorem,

$$\frac{1}{\epsilon} D(\tilde{P}_X\|P_X) = \mathbb{E}_{P_X}\left[f\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}, \epsilon \right) \right] \to \mathbb{E}_{P_X}\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X} - 1 \right] \log e = 0\,.$$

To drop the boundedness assumption on $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ we simply consider the conditional distribution $Q'_X \triangleq Q_{X|X\in A}$ where $A = \{x : \frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x) < a\}$ and $a > 0$ is sufficiently large so that $Q_X(A) > 0$. Clearly, as $a \to \infty$, we have $Q'_X \to Q_X$ and $Q'_Y \to Q_Y$ pointwise, where $Q'_Y \triangleq P_{Y|X} \circ Q'_X$. Hence the lower-semicontinuity of divergence yields

$$\liminf_{a\to\infty} D(Q'_Y\|P_Y) \ge D(Q_Y\|P_Y)\,.$$

Furthermore, since $\frac{\mathrm{d}Q'_X}{\mathrm{d}P_X} = \frac{1}{Q_X(A)} \frac{\mathrm{d}Q_X}{\mathrm{d}P_X} \mathbf{1}_A$, we have

$$D(Q'_X\|P_X) = \log \frac{1}{Q_X(A)} + \frac{1}{Q_X(A)} \mathbb{E}_Q\left[\log \frac{\mathrm{d}Q_X}{\mathrm{d}P_X} \mathbf{1}\left\{ \frac{\mathrm{d}Q_X}{\mathrm{d}P_X} \le a \right\} \right]. \tag{162}$$

Since $Q_X(A) \to 1$, by dominated convergence (note: $\mathbb{E}_Q[|\log \frac{\mathrm{d}Q_X}{\mathrm{d}P_X}|] < \infty$) we have $D(Q'_X\|P_X) \to D(Q_X\|P_X)$. Therefore,

$$\liminf_{a\to\infty} \frac{D(Q'_Y\|P_Y)}{D(Q'_X\|P_X)} \ge \frac{D(Q_Y\|P_Y)}{D(Q_X\|P_X)}\,,$$

completing the proof.

References

[AG76] R. Ahlswede and P. Gács. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.*, pages 925–939, 1976.

[AGKN13] Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *arXiv preprint arXiv:1304.6133*, 2013.

[Ahl07] R. Ahlswede. The final form of Tao's inequality relating conditional expectation and conditional mutual information. *Advances in Mathematics of Communications (AMC)*, 1(2):239–242, 2007.

[BC05] A.D. Barbour and L.H.Y. Chen. *An introduction to Stein's method*. World Scientific, Singapore, 2005.

[Ber71] Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[CIR+93] J.E. Cohen, Yoh Iwasa, Gh. Rautu, M.B. Ruskai, E. Seneta, and Gh. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear algebra and its applications*, 179:211–235, 1993.

[CKZ98] J. E. Cohen, J. H. B. Kempermann, and Gh. Zbăganu. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*. Springer, 1998.

[CPW15] F. Calmon, Y. Polyanskiy, and Y. Wu. Strong data processing inequalities in power-constrained Gaussian channels. In *Proc. 2015 IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, CN, June 2015.

[CRS94] M. Choi, M.B. Ruskai, and E. Seneta. Equivalence of certain entropy contraction coefficients. *Linear algebra and its applications*, 208:29–36, 1994.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:229–318, 1967.

[Csi95] I. Csiszár. Generalized cutoff rates and Renyi's information measures. *IEEE Trans. Inf. Theory*, 41(1):26 –34, January 1995.

[Csi96] Imre Csiszár. Almost independence and secrecy capacity. *Prob. Peredachi Inform.*, 32(1):48–57, 1996.

[DMLM03] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probab. Theory Relat. Fields*, 126:395–420, 2003.

[Dob56] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory Probab. Appl.*, 1(1):65–80, 1956.

[Dob70] R. L. Dobrushin. Definition of random variables by conditional distributions. *Theor. Probability Appl.*, 15(3):469–497, 1970.

[EC98] Elza Erkip and Thomas M. Cover. The efficiency of investment information. *IEEE Trans. Inf. Theory*, 44(3):1026–1040, 1998.

[EKPS00] William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, pages 410–433, 2000.

[ES99] William S Evans and Leonard J Schulman. Signal propagation and noisy circuits. *IEEE Trans. Inf. Theory*, 45(7):2367–2373, 1999.

[ES03] William S Evans and Leonard J Schulman. On the maximum tolerable noise of k-input gates for reliable computation by formulas. *IEEE Trans. Inf. Theory*, 49(11):3094–3098, 2003.

[Geo11] Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9. Walter de Gruyter, 2011.

[Hig77] Yasunari Higuchi. Remarks on the limiting Gibbs states on a $(d+1)$-tree. *Publ. RIMS Kyoto Univ.*, 13(2):335–348, 1977.

[HUL96] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I: Fundamentals*. Springer, Berlin, Germany, 1996.

[HW91] Bruce Hajek and Timothy Weller. On the maximum tolerable noise for reliable computation by formulas. *IEEE Trans. Inf. Theory*, 37(2):388–391, Mar. 1991.

[Led99] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. *Seminaire de probabilites XXXIII*, pages 120–216, 1999.

[LM11] Gabriel M. Lipsa and Nuno C. Martins. Optimal memoryless control in Gaussian noise: A simple counterexample. *Automatica*, 47(3):552–558, 2011.

[LR06] Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[OV00] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[Pin05] Mark S. Pinsker. On estimation of information via variation. *Prob. Peredachi Inform.*, 41(2):71–75, 2005.

[Pip88] Nicholas Pippenger. Reliable computation by formulas in the presence of noise. *IEEE Trans. Inf. Theory*, 34(2):194–197, 1988.

[Pro04] V. Prokaj. A Characterization of Singular Measures. *Real Analysis Exchange*, 29:805–812, 2004.

[PV10] Y. Polyanskiy and S. Verdú. Arimoto channel coding converse and Rényi divergence. In *Proc. 2010 48th Allerton Conference*. Allerton Retreat Center, Monticello, IL, USA, September 2010.

[PW15] Yury Polyanskiy and Yihong Wu. Strong data-processing of mutual information: beyond Ahlswede and Gács. In *Proc. Information Theory and Applications Workshop*, San Diego, CA, February 2015.

[Rag13] Maxim Raginsky. Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels. In *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 419–423, 2013.

[Rag14] Maxim Raginsky. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *arXiv preprint arXiv:1411.3575*, November 2014.

[Sar58] O. V. Sarmanov. A maximal correlation coefficient. *Dokl. Akad. Nauk SSSR*, 121(1), 1958.

[Sib69] R. Sibson. Information radius. *Z. Wahrscheinlichkeitstheorie und Verw. Geb.*, 14:149–161, 1969.

[SM62] S. Kh. Sirazhdinov and M. Mamatov. On convergence in the mean for densities. *Theory of Probability & Its Applications*, 7(4):424–428, 1962.

[Str65] V. Strassen. The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[Sub12] R. Subramanian. The relation between block length and reliability for a cascade of AWGN links. In *Proc. 2012 Int. Zurich Seminar on Communications (IZS)*, pages 71–74, Feb. 2012.

[SVL13] R. Subramanian, B.N. Vellambi, and I. Land. An improved bound on information loss due to finite block length in a Gaussian line network. In *Proceedings of 2013 IEEE International Symposium on Information Theory*, Istanbul, Turkey, Jul. 2013.

[Tao06] T. Tao. Szemerédi's regularity lemma revisited. *Contributions to Discrete Mathematics*, 1(1):8–28, 2006.

[Ung10] Falk Unger. Better gates can make fault-tolerant computation impossible. In *Elect. Colloq. Comp. Complexity (ECCC)*, volume 17, page 164, 2010.

[vN56] John von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, 34:43–98, 1956.

[Wit75] H.S. Witsenhausen. On sequences of pairs of dependent random variables. *SIAM J. Appl. Math.*, 28:100–113, 1975.

[Wu11] Yihong Wu. A simple proof of the Gaussian HWI inequality and extensions. *Preprint*, Sep. 2011.

[WV12] Yihong Wu and Sergio Verdú. Functional properties of MMSE and mutual information. *IEEE Trans. Inf. Theory*, 58(3):1289 – 1301, Mar. 2012.