Finite Blocklength Analysis of MISO Block Fading Channels

by

Austin Daniel Collins

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

Masters of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2014

Author .....
Department of Electrical Engineering and Computer Science
August 29, 2014

Certified by .....
Yury Polyanskiy
Robert J. Shillman Assistant Professor
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses

Finite Blocklength Analysis of MISO Block Fading Channels

by

Austin Daniel Collins

Submitted to the Department of Electrical Engineering and Computer Science
on August 29, 2014, in partial fulfillment of the
requirements for the degree of
Masters of Science in Computer Science and Engineering

Abstract

Write Me

Thesis Supervisor: Yury Polyanskiy
Title: Robert J. Shillman Assistant Professor

Acknowledgments

This is the acknowledgements section. You should replace this with your own acknowledgements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Wireless Communication and Fading Channels

A common statistical model for a wireless communication medium is the fading channel, which generalizes the AWGN by introducing multiplicative fading coefficients to model multipath interference effects. The model is flexible, allowing for different fading distributions (E.g. Rayleigh (richly scattered) or Rician (line of sight)), types of fading processes (block fading, quasi-static, fast fading), types of channel state information (available at the transmitter, receiver, both, or none), and for the inclusion of multiple antennas at the transmitter or receiver (the fading process accounts for independently faded signal paths between the transmitter and receiver). This model is well studied in information theory literature, its capacity is known in all the above scenarios (for an overview see [2]).

Amongst the most notable successes of this channel model is the discovery that using multiple antennas at both the transmitter and receiver boosts the capacity of the channel linearly, proportional to the minimum number of transmit and receive antennas. This discovery opened up huge new potential gains in communication systems, since power and bandwidth are often so strictly constrained. This also led to the investigation of space time codes, which are communication schemes that take advantage of having multiple antennas. More recently, as new hardware is developed, new MIMO systems are becoming possible. For example Mega-MIMO [13] which attempts to make many independent transmitters act as if they were a single MIMO system (through synchronization) in order to gain rate benefits, or Large Scale Antenna Systems (LSAS) [7], which places a very large number of antennas at a base station in order to improve downlink communication.

We are interested in characterizing the fundamental communication limits of this class of channels. The classical fundamental limits of fading channels (i.e. their capacity) are well known in most cases, and these limits tell code designers the maximum possible near-error-free data rate achievable through these channels. However, these classical fundamental limits make the assumption that the communication system is able to use arbitrarily long coding blocks. The decoder must wait until it has received the entire block in order to decode. In many applications, namely when there are

delay constraints (e.g. voice or video), the system cannot wait very long before decoding each block. So the channel capacity gives us a maximum possible transmission rate, but if we can only use coding blocks of length n, then the capacity may be quite an inaccurate estimate of the actual achievable rate.

Although the capacity of these fading channels is well known, finding achievable rates at finite blocklength is a new area of study in information theory (at least, it has been revitalized recently, but has been addressed off and on before), and finite blocklength results are only known for a small subset of the possible fading channel models. Knowledge of these limits will guide code designers in the search for the optimal codes, as well as show how these limits scale with power, bandwidth, and number of antennas. Recent powerful techniques [9] have been developed in order to analyze the maximum cardinality of the codebook that achieves block length n and error probability ϵ, denoted by $\log M^*(n,\epsilon)$. Knowledge of $\log M^*(n,\epsilon)$ completely characterized the limits of communication over a channel, but it is infeasible to commute directly. Instead, the quantity must be bounded from above and below.

As a resolution of this computational difficulty [9] proposed a closed-form normal approximation, based on the asymptotic expansion:

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)\,, \tag{1.1}$$

where capacity C and dispersion V are two intrinsic characterstics of the channel and $Q^{-1}(\epsilon)$ is the inverse of the Q-function[1]. One immediate consequence of the normal approximation is an estimate for the minimal blocklength (delay) required to achieve a given fraction η of channel capacity:

$$n \gtrsim \left(\frac{Q^{-1}(\epsilon)}{1-\eta}\right)^2 \frac{V}{C^2}\,. \tag{1.2}$$

Asymptotic expansions such as (1.1) are rooted in the central-limit theorem and have been known classically for discrete memoryless channels [4, 15] and later extended in a wide variety of directions; see [10] for a survey.

We are interested in extending this dispersion analysis to fading channels. Specifically, we analyze the dispersion of the MISO Coherent block fading channel with channel state information available to the receiver and isotropic fading fading (the channel model is described in detail in Chapter **??**). This model may represent the downlink channel from a cell tower to a mobile receiver, where the tower has the space to have many antennas, while the mobile device only can only have one antennas. The isotropic noise assumption means that the communication is in a "rich scattering" environment where the signal has no line of sight path to the receiver. A special case of this assumption is the common Rayleigh Fading process, where all fading coefficients have i.i.d. Gaussian distribution.

We notice that this MISO block fading channel has *non-unique* capacity achieving input distributions. This non-uniqueness hasn't appeared in any other finite

[1] As usual, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt\,.$

blocklength analysis done thus far, so handling this problem is new, and turns out to be non-trivial. Specifically, orthogonal designs (for example, Alamouti's scheme) achieve capacity in this channel for dimensions where they exist. It turns out that $\mathrm{Var}[i(X;Y,H)|X]$ is an achievable dispersion for any capacity achieving input distribution P_X. For the AWGN channel, there was only one input distribution that achieved capacity, and the conditional variance with this input gave the true dispersion. In this case there are non unique input distributions that achieve capacity, and it will be shown that orthogonal designs minimize the conditional variance over all distributions that achieve capacity. Orthogonal designs were introduced into the field of communications by Tarokh et al in [16], where it was showed that, in a MIMO channel, orthogonal designs achieve the maximum diversity order of the channel while having a simple liner decoder. This thesis provides a theoretical justification for the optimally of orthogonal designs from finite blocklength analysis, in the given channel.

This thesis is organized as follows: Chapter 2 gives the channel model and notation used throughout the thesis, Chatper 3 describes the multiple non-unique capacity achieving input distribution for this channel, as well as computes its conditional variance, Chapter 4 gives the proof of achievable dispersion and compares the result to error exponents, and Chapter 5 discusses Orthogonal Designs and their relation to the achievable dispersion.

Chapter 2

Channel Model and Basic Definitions

The main object of study in this thesis is the coherent MISO (Multiple-Input Single-Output) block fading channel. This channel models a wireless communication channel where the transmitter has multiple antennas and the receiver only has one. In this chapter, we will give the definitions and notation necessary to define a communication channel (Section **??**), a brief description of the many statistical models of wireless channels (Section **??**), and then describe the coherent MISO block fading channel model (Section **??**).

2.1 The Coherent MISO Block Fading Channel

In this work, we examine the block fading with channel state information available to the receiver (CSIR) where the system is MISO (Multiple-Input-Single-Output, i.e. multiple antennas at the transmitter and a single antenna at the receiver). Motivated by a recent surge of orthogonal frequency division (OFDM) technology, this thesis focuses on the frequency-nonselective coherent real block fading discrete-time channel.

The situation where this channel is most practical is in the downlink of a communication system where there is no line of sight between the transmitter and receiver. Here, a base station is able to have many antennas whereas smaller receivers (such as cell phones) only have the space for one antenna. In any urban environment, there will likely be no direct line of sight between the base station and the mobile device, so the signal will reflect off various buildings and such before reaching the receiver.

Formally, let $n_t \geq 1$ be the number of transmit antennas and $T \geq 1$ be the coherence time of the channel. The input-output relation at block j (spanning time instants $(j-1)T+1$ to jT) with $j = 1, \ldots, n$ is given by

$$Y_j = H_j X_j + Z_j \,, \tag{2.1}$$

where $\{H_j, j = 1, \ldots\}$ is a $1 \times n_t$ vector-valued random fading process, X_j is a $n_t \times T$ matrix channel input, Z_j is a $1 \times T$ Gaussian random vector with independent entries of variance 1, and Y_j is the $1 \times T$ vector-valued channel output. The process H_j is

assumed to be i.i.d. with isotropic distribution P_H (that is, $H \sim HU$ for any $n_t \times n_t$ orthogonal matrix U) satisfying

$$\mathbb{E}[\|H\|^2] = 1\,. \tag{2.2}$$

Note that because of merging channel inputs at time instants $1, \ldots, T$ into one matrix-input, the block-fading channel becomes memoryless. We assume coherent demodulation so that the channel state information H_j is fully known to the receiver (CSIR). Now we give the definition of a code for this channel.

Definition 1. An $(nT, M, \epsilon, P)_{CSIR}$ code of blocklength nT, probability of error ϵ and power-constraint P is a pair of maps: the encoder $f : \{1, \ldots, M\} \to (\mathbb{R}^{n_t \times T})^n$ and the decoder $g : (\mathbb{R}^{1\times T})^n \times (\mathbb{R}^{1\times n_t})^n \to \{1, \ldots, M\}$ such that the probability of decoding incorrectly satisfies

$$\mathbb{P}[W \neq \hat{W}] \leq \epsilon\,. \tag{2.3}$$

and the codewords must satisfy the power constraint

$$\sum_{j=1}^{n} \|X_j\|_F^2 \leq nTP \qquad \mathbb{P}\text{-a.s.}\,,$$

($\|A\|_F^2 = \sum_{ij} |a_{ij}|^2$ is the Frobenius norm of the matrix) on the probability space

$$W \to X^n \to (Y^n, H^n) \to \hat{W}\,,$$

Where W is uniform on $\{1, \ldots, M\}$, $X^n = f(W)$, $X^n \to (Y^n, H^n)$ is as described in (2.1), and $\hat{W} = g(Y^n, H^n)$.

We will focus on the real where all quantities are real valued, the analysis is very similar in the complex case.

2.2 Capacity

Under the isotropy assumption on P_H, the capacity C appearing in (1.1) of this channel is given by [17]

$$C(P) = \mathbb{E}\left[C_{AWGN}\left(\frac{P}{n_t}\|H\|^2\right)\right]\,, \tag{2.4}$$

where $C_{AWGN}(P) = \frac{1}{2}\log(1+P)$ is the capacity of the additive white Gaussian noise (AWGN) channel with SNR P.

With this, we define a *capacity achieving input distribution* (CAID), which will play a large role in this work.

Definition 2. A distribution P_X satisfying the power constraint $\mathbb{E}[\|X\|_F^2] \leq TP$ is a

capacity achieving input distribution if

$$I(P_X, P_{YH|X}) = C(P) \tag{2.5}$$

Where $I(P_X, P_{Y|X})$ is the mutual information when the channel input has distribution P_X.

We will show in Chapter 3 that many input distributions achieve capacity, including orthogonal designs. Then we will analyze which input gives the best finite blocklength performance over this set of distributions.

2.3 Dispersion

A channel code maps a message space $\{1, \ldots, M\}$ into a sequence of n symbols which are sent over the channel. The decoder waits to receive all n symbols, then decodes to the most likely message. Classically, the capacity of a channel tells us the maximum data rate we can send over a channel with arbitrarily small error probability, given that the blocklength n can be arbitrarily large. In practice, this means the decoder must wait a very long time before decoding a block, causing long *delay*. If we insist that the code can only use blocklength n, then for most cases, we cannot achieve data rates arbitrarily close to capacity. The *channel dispersion* quantifies the "rate penalty" incurred for transmitting at a fixed blocklength n and error probability ϵ.

The dispersion V was formally defined in [9] as

$$V = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \left(\frac{nC - \log M^*(n, \epsilon)}{Q^{-1}(\epsilon)} \right)^2 \tag{2.6}$$

For DMC's, it was shown in [9] that the dispersion is given by the variance of the information density

$$\mathrm{Var}(i(X, Y)) \tag{2.7}$$

Where P_X is a capacity achieving input distribution, and the information density is defines as

$$i(x; y) \triangleq \log \frac{P_{Y|X}(y|x)}{P_Y(y)} \tag{2.8}$$

Where $P_Y = \int P_{Y|X}(y|x) dP_X$ is the output distribution induced through the channel by input P_X. For the AWGN channel and the SISO coherent fading channel, the dispersion was shown to instead be the *conditional variance* of the information density

$$\mathbb{E}_X \mathrm{Var}(i(X; Y)|X) \tag{2.9}$$

This will be denoted by simply $\mathrm{Var}[i(X; Y)|X]$. Note that for a DMC, quantities

(2.7) and (2.9) are the same, since we can expand (2.7) as

$$\mathrm{Var}(i(X;Y)) = \mathbb{E}\mathrm{Var}(i(X;Y)|X) + \mathrm{Var}\left(\mathbb{E}[i(X;Y)|X]\right) \tag{2.10}$$

In the second term above, for all inputs x, $\mathbb{E}[i(x,Y)] = C$ for a DMC, where C is the capacity of the channel. Hence this term is the variance of a constant, which is zero. We'll show that for the MISO coherent block fading channel, $\mathrm{Var}[i(X;Y,H)|X]$ is achievable, and then we'll look at minimum over the set of input distributions that achieve capacity.

Chapter 3

Characterizing the Channel

The coherent MISO block fading channel has an interesting property: the capacity achieving input distribution (CAID) P_X^* is not unique. All the channels with a known dispersion, e.g. BSC, BEC, AWGN, SISO coherent fading channel, have a unique capacity achieving input distribution. In these channels, the dispersion turns out to be either $\mathrm{Var}[i(X;Y)]$ (discrete case) or $\mathrm{Var}[i(X;Y)|X]$ (cost constrained case), where in both the distribution of X was simply chosen as he unique CAID. However, when the are multiple CAIDs, the question becomes: do some capacity achieving input distributions give better dispersion than others? If so, which ones?

In this chapter, we'll give necessary and sufficient conditions for the channel input X to achieve capacity for the general MISO fading channel with n_t transmit antennas and coherence time T. We will compute the conditional variance of the information density of this channel, which will be shown to be achievable in Chapter 4.

3.1 Motivation Example: Alamouti's Scheme vs Independent Gaussians

As a motivating example, we first consider a special case of $n_t = T = 2$, meaning there are 2 antennas at the transmitter and the coherence time is 2. As argued by Telatar [17], the following input achieves capacity

$$X = \sqrt{\frac{P}{2}} \begin{bmatrix} \xi_1 & \xi_3 \\ \xi_2 & \xi_4 \end{bmatrix}, \tag{3.1}$$

where here and below ξ_j are i.i.d. standard normal random variables. Reflecting upon ingenious scheme of Alamouti [1] we observe that the following input is also capacity achieving:

$$X = \sqrt{\frac{P}{2}} \begin{bmatrix} \xi_1 & -\xi_2 \\ \xi_2 & \xi_1 \end{bmatrix} \tag{3.2}$$

The Alamouti scheme sends two data symbol in two timeslots (i.e. achieves a

multiplexing gain of 2 and a *diversity gain* of 2). The iid Gaussian scheme sends four symbols in two time slots (multiplexing gain of 4 and diversity gain of 1). The channel only has one degree of freedom per timeslot, so both schemes use all available degrees of freedom.

The intuition why both of these distributions achieve capacity is as follows (see Figure 3-1). The independent Gaussian scheme sends two vectors (one $\mathbb{R}^2$ vector in each time slot), each independently isotropically distributed. The channel acts by projecting each input vector onto the vector of fading coefficients H, effectively killing all information sent in the direction orthogonal to H. With this, sometimes we get lucky and both transmitted vectors are nearly parallel with H, and sometimes we're unlucky and both vectors are nearly orthogonal. The Alamouti scheme sacrifices multiplexing gain (only sends 2 symbols instead of 4), but is more robust to fading. Alamouti's scheme sends two orthogonal data vectors, so that if one happens to be nearly orthogonal to H, the other will be nearly parallel. With this, the magnitude of the projection onto H is the same regardless of the angle of the data vectors. This intuition will be useful later when we discuss the dispersion achieved by each of these inputs.

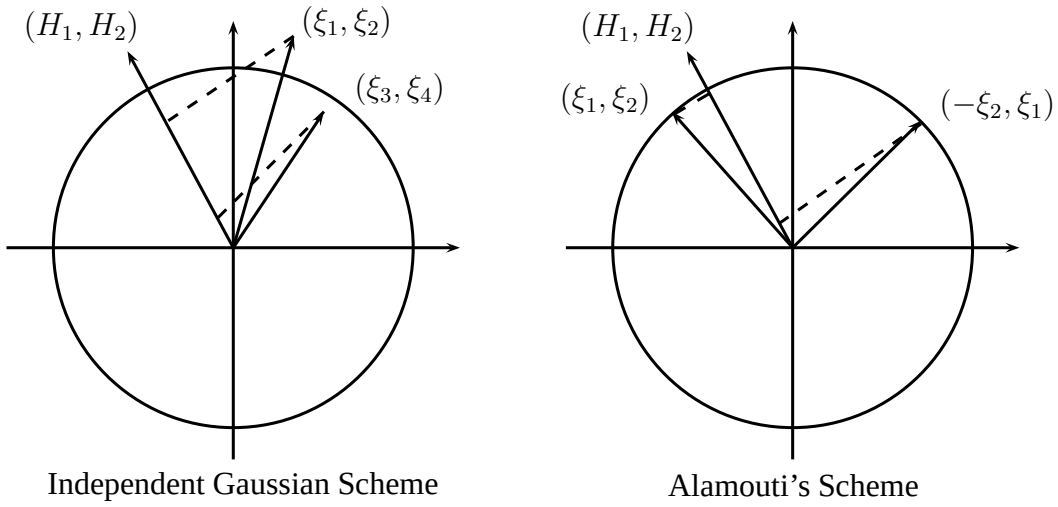

Figure 3-1: (Left) The independent Gaussian scheme transmits two vectors, each isotropically distributed independently, while (Right) Alamouti's scheme transmits two perpendicular vectors.

It will be shown that Alamouti's scheme achieves the smaller value of $\mathrm{Var}[i(X;Y,H)|X]$ than Teletar's i.i.d. Gaussian input. In Chapter 4, we'll show that $\mathrm{Var}[i(X;Y,H)|X]$ is achievable, and thus the Alamouti's scheme gives a strictly better dispersion. We are then lead to the question: is there a CAID for $n_t = T = 2$ that gives a better dispersion than Alamouti's scheme? Fefore we analyze the dispersion of these inputs, we first characterize all capacity achieving input distributions for the channel.

3.2 Capacity Achieving Input Distributions

We saw that both the input X with i.i.d. Gaussian entries and the input based on Alamouti's scheme achieve capacity in the MISO block fading channel. The following proposition gives necessary and sufficient conditions for an input distribution to achieve capacity. An immediate consequence of this proposition is that if any $n_t \times T$ orthogonal design exists in indeterminates $x_1, \ldots, x_T$ (for $T \geq n_t$), then the input distribution with this structure achieves capacity.

Proposition 1. *P_X is a capacity achieving input distribution iff all rows and columns of X are jointly Gaussian and either of the following holds*

1. *Let R_i denote the i-th row of X, then:*

$$\mathbb{E}[R_i^T R_i] = \frac{P}{n_t} I_T, \qquad i = 1, \ldots, n_t \tag{3.3}$$

$$\mathbb{E}[R_i^T R_j] = -\mathbb{E}[R_j^T R_i], \qquad i \neq j \tag{3.4}$$

2. *Let C_i be the i-th column of X, then:*

$$\mathbb{E}[C_i C_i^T] = \frac{P}{n_t} I_T, \qquad i = 1, \ldots, T \tag{3.5}$$

$$\mathbb{E}[C_i C_j^T] = -\mathbb{E}[C_j C_i^T], \qquad i \neq j \tag{3.6}$$

Example: In the $n_t = T = 2$ case, the set of jointly Gaussian CAIDs is given by

$$\left\{ \sqrt{\frac{P}{2}} \begin{bmatrix} \xi_1 & -\rho\xi_2 + \sqrt{1-\rho^2}\xi_3 \\ \xi_2 & \rho\xi_1 + \sqrt{1-\rho^2}\xi_4 \end{bmatrix} \; : \; -1 \leq \rho \leq 1 \right\} \tag{3.7}$$

Where $\xi_1, \xi_2, \xi_3, \xi_4 \sim \mathcal{N}(0,1)$ iid. Note that there are CAIDs that are not joinly Gaussian (although all rows and columns are jointly Gaussian), for example if we take a mixture of two CAIDS from the collection above.

Remark 1. These conditions imply that if X is caid, then X^T and any submatrix of X are caids too (for different n_t and T). Elementwise, these conditions require that all elements in a row are pairwise independent, all elements in a column are pairwise independent, each 2×2 minor has equal and opposite correlation across diagonal entries, and each of the entries have the same distribution $X_{ij} \sim \mathcal{N}(0, \frac{P}{n_t})$.

Proof. Observe that distribution P_X achieves capacity iff for P_H-almost every h_0 it induces the optimal output distribution

$$P^*_{Y|H=h_0} = \mathcal{N}\left(0, \left(1 + \frac{P}{n_t}\|h_0\|^2\right) I_T\right)$$

Indeed, note that

$$I(X; Y, H) = h(Y|H) - h(Z)\,, \tag{3.8}$$

where $h(\cdot)$ is a differential entropy [**?**, Chapter 9]. Since this expression only depends on $P_{Y|H}$, then P_X is optimal if it induces $P^*_{Y|H}$. Conversely, maximizing (3.8) is equivalent to maximizing $h(Y|H)$ given a second moment constraint. We know that for any P_X and any fixed $H = h_0$

$$h(Y|h_0) \le h\left(\mathcal{N}\left(0, (1+\frac{P}{n_t}\|h_0\|^2)I_t\right)\right) \tag{3.9}$$

i.e. iid Gaussian maximizes entropy subject to a second moment constraint, with equality iff $P_Y \sim \mathcal{N}\left(0, \left(1+\frac{P}{n_t}\|h_0\|^2\right) I_T\right)$. Hence, in order to attain capacity (2.4), $P_{Y|H}$ must coincide with $P^*_{Y|H}$ for P_H-almost every h_0.

Next, fix a capacity achieving distribution P_X and let E_0 be an almost sure set of those h_0 for which $P_{Y|H=h_0} = P^*_{Y|H=h_0}$. Let $\{U_k, k = 1, \ldots\}$ be a dense subset of all orthogonal $n_t \times n_t$ matrices. By isotropy of P_H we have $P_H[U_k(E_0)] = 1$ and therefore

$$E \stackrel{\triangle}{=} E_0 \cap \bigcap_{k=1}^{\infty} U_k(E_0)$$

is also almost sure: $P_H[E] = 1$. By assumption (2.2) E must contain a non-zero element, and hence (by density of $\{U_k\}$), must also be dense in some sphere in $\mathbb{R}^{1\times n_t}$, which without loss of generality we assume to have radius 1. Now take an $h_0 \in \mathbb{S}^{n_t-1} \cap E$. Then since $h_0 X + Z$ is Gaussian, a theorem of Cramer [3, Theorem 1] implies that $h_0 X$ must itself be jointly Gaussian. Equivalently, by uniqueness of the characteristic function, for arbitrary $\theta \in \mathbb{R}^{1\times n_t}$ we have

$$\mathbb{E}\left[e^{ih_0 X\theta^T}\right] = e^{-\frac{P}{2n_t}\|\theta\|^2\|h_0\|^2} \tag{3.10}$$

Since the characteristic function is continuous in h_0, identity (3.10) must hold for all $h_0 \in \mathbb{R}^{1\times n_t}$ (by density of E and scaling $h_0 \mapsto \lambda \cdot h_0$). Consequently, for every h_0 we have

$$h_0 X \sim \mathcal{N}\left(0, \frac{P}{n_t}\|h_0\|^2 I_T\right) \tag{3.11}$$

In particular, we have (denoting components of h_0 by $h_i, i = 1, \ldots n_t$)

$$\mathbb{E}[(h_0X)^T(h_0X)] = \mathbb{E}\left[\sum_{i,j=1}^{n_t} h_i h_j R_i^T R_j\right] = \frac{P}{n_t}\left(\sum_{i=1}^{n_t} h_i^2\right) I_T\,.$$

Intepreting the last equation as equality of bilinear forms yields the set of conditions 1), 2) follows similarly. Since (3.10) holds for all θ and h_0, choosing these appropriately shows that the rows and columns of X must be jointly Gaussian. □

Remark 2. The characteristic function $\mathbb{E}[e^{i\,\mathrm{tr}(s^T X)}]$ where $s \in \mathbb{R}^{n_t\times T}$, isn't fully specified. (3.10) only gives its value for those s such that $s = h_0^T \otimes \theta$ for row vectors $h_0 \in \mathbb{R}^{n_t}$, $\theta \in \mathbb{R}^T$, and $\otimes$ denotes the tensor product. This set is equivalent to all

$n_t \times T$ rank 1 matrices, which is clearly not equal to all of $\mathbb{R}^{n_t \times T}$. So some non-jointly Gaussian relations between entries of X can occur when s has rank greater than 1.

We will explore orthogonal designs in depth in Chapter 5, but to show the use of this proposition, we will quickly show that any orthogonal design input distribution achieve capacity. Any real orthogonal design can be represented as the sum $\sum_{i=1}^{k} x_i V_i$ where $\{x_1, \ldots, x_k\}$ are indeterminates, and $\{V_1, \ldots, V_k\}$ is a collection of $n \times n$ real matrices satisfying Hurwitz-Radon conditions:

$$V_i^T V_i = I_n \tag{3.12}$$

$$V_i^T V_j + V_j^T V_i = 0 \tag{3.13}$$

Suppose $n_t \leq T$. Given such a collection of $T \times T$ V_i's, form the orthogonal design input distribution by

$$X = \begin{bmatrix} V_1^T \xi^T & \cdots V_k^T \xi^T \end{bmatrix}^T \tag{3.14}$$

Where the row vector $\xi \sim \mathcal{N}(0, \frac{P}{n_t} I_T$. Then each row and column is jointly Gaussian, and applying the CAID conditions (3.3) and (3.4), we see immediately

$$\mathbb{E}[R_i^T R_i] = V_i^T \mathbb{E}[\xi^T \xi] V_i = \frac{P}{n_t} V_i^T V_i = \frac{P}{n_t} I_T \tag{3.15}$$

$$\mathbb{E}[R_i^T R_j] = V_i^T \mathbb{E}[\xi^T \xi] V_j = V_i^T V_j = -V_j^T V_i = -\mathbb{E}[R_j^T R_i] \tag{3.16}$$

So that the orthogonal design input distribution X satisfies the CAID condition, and hence achieves capacity. Note the striking resemblance between the CAID conditions (3.3) and (3.4) and the Hurwitz-Radon conditions (3.12) and (3.13).

3.3 Information Density and Conditional Variance

A key tool for finite blocklength results is the *information density*, from which the mutual information and conditional variance can be derived. It is defined as follows:

Definition 3. For a joint distribution P_{XY}, the information density is given by

$$i(x; y) = \log \frac{dP_{Y|X=x}(y)}{dP_Y(y)} \tag{3.17}$$

and $i(x, y) = \infty$ for all x where $P_{Y|X=x}$ is not absolutely continuous with respect to P_Y.

It is instructive to interpret the information density as a log likelihood ratio between the channel output distribution $P_{Y|X=x}$ induced when x is sent over the channel, and the "average noise" P_Y. Note too that $\mathbb{E}[i(X; Y)] = I(X; Y)$. Later we will be interested in the second and even third moments of the information density. We first calculate the information density for the channel under study.

For any channel, all CAIDs induce the same output distribution, since CAOD (capacity achieving output distribution – the distribution induced by P_X^* through the channel) is always unique. For the coherent MISO fading channel, the CAOD induced by one coherent block has distribution $P^*_{Y|H=h} \sim \mathcal{N}\left(0, \left(\frac{P}{n_t}\|h\|^2 + 1\right) I_T\right)$ where I_T is the $T \times T$ identity matrix. Hence over n blocks, the CAOD is

$$P^*_{Y^n|H^n=h^n} \sim \mathcal{N}\left(0, \frac{P}{n_t}A + I_{nT}\right) \tag{3.18}$$

Where A is a diagonal matrix $nT \times nT$ matrix with each of the $T \times T$ submatrices along A's diagonal are $\|h_i\|^2 I_T$ for $i = 1, \ldots, n$.

Letting h_j represent the $1 \times n_t$ real vector of fading coefficients for the j-th block, and x_i bet the $n_t \times T$ real input matrix for the j-th block, then the information density over n fading blocks is

$$i(x; y, h) = \frac{T}{2}\sum_{j=1}^{n} \log\left(1 + \frac{P}{n_t}\|h_j\|^2\right) + \frac{1}{2}\frac{\|h_j x_j\|^2 + 2h_j x_j Z_j^T - \frac{P}{n_t}\|h_j\|^2\|Z_j\|^2}{1 + \frac{P}{n_t}\|h_j\|^2} \tag{3.19}$$

where $Z = hx - y$ is a $1 \times T$ real vector. This can be computed simply by taking the log of the ratio of densities $\mathcal{N}([h_1 x_1, \cdots, h_n x_n], I_{nT})$ and $P^*_{Y|H=h}$ from (3.18). Note that indeed taking $\frac{1}{T}\mathbb{E}[i(X;Y)]$ gives the channel capacity. For finite blocklength analysis, we're interested in the conditional variance of the information density, $\mathrm{Var}[i(X;Y,H)|X]$. This will be shown to be achievable in Chapter 4. The following two propositions compute the more general $\mathrm{Var}(i(x;Y,H))$, and then the conditional variance, both of which will be useful later. Although the computation is mainly algebraic manipulation, it's an essential piece of many arguments in this thesis, so we give the full derivation.

Proposition 2. *For the information density given in (3.19),* $(Y^n, H^n) \sim P_{H^n} P^*_{Y^n|H^n}$,

$$\begin{aligned}\frac{1}{nT}\mathrm{Var}(i(x^n; Y^n, H^n)) = T\mathrm{Var}\left(C_{AWGN}\left(\frac{P}{n_t}\|H\|^2\right)\right) + \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\|H\|_2\right)\right] \\ + \frac{\chi_0}{n_t nT}\left(\|x^n\|_F^2 - nTP\right) + \frac{1}{4nT}\sum_{j=1}^{n}\mathrm{Var}\left(\frac{\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)\|H_j\|^2}{1 + \frac{P}{n_t}\|H_j\|^2}\right)\end{aligned} \tag{3.20}$$

And the last term reduces to

$$\frac{1}{4nT}\sum_{j=1}^{n}\mathrm{Var}\left(\frac{\left(\|\tilde{H}_j x_j\|^2-\frac{TP}{n_t}\right)\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right) \tag{3.21}$$

$$=\frac{1}{nT}\sum_{j=1}^{n}\chi_1\mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2-\frac{TP}{n_t}\right)^2\right]-\chi_2\left(\frac{\|x_j\|_F^2}{n_t}-\frac{TP}{n_t}\right)^2 \tag{3.22}$$

Where $\tilde{H}_j \triangleq \frac{H_j}{\|H_j\|}$ is the normalized fading vector for the j-th block, and

$$\chi_0 \triangleq 4\mathbb{E}\left[\left(\lambda\left(\frac{P}{n_t}\|H\|^2\right)\|H\|\right)^2\right] \tag{3.23}$$

$$\chi_1 \triangleq \mathbb{E}\left[\left(\lambda\left(\frac{P}{n_t}\|H\|^2\right)\|H\|^2\right)^2\right] \tag{3.24}$$

$$\chi_2 \triangleq \mathbb{E}^2\left[\lambda\left(\frac{P}{n_t}\|H\|^2\right)\|H\|^2\right] \tag{3.25}$$

$$C_{AWGN}(P) \triangleq \frac{1}{2}\log(1+P) \tag{3.26}$$

$$V_{AWGN}(P) \triangleq \frac{1}{2}\left(1-\left(\frac{1}{1+P}\right)^2\right) \tag{3.27}$$

$$\lambda(P) \triangleq \frac{1}{2(1+P)} = \frac{dC_{AWGN}(P)}{dP} \tag{3.28}$$

The last three quantities are the AWGN capacity, dispersion, and optimal Lagrange multiplier $\lambda(P)$ in the optimization

$$\sup_{P_X:\mathbb{E}[\|X\|^2]\le P} I(X;Y) \tag{3.29}$$

Proof. Note that since the channel is memoryless, each fading block is independent of the others, so we only need to consider a single fading block,

$$\mathrm{Var}(i(x^n;Y^n,H^n)) = \sum_{j=1}^{n}\mathrm{Var}(i(x_j;Y_j,H_j)) \tag{3.30}$$

Now, the variance distributes over the sum in the information density since the first term is independent of Z (hence zero covariance):

$$\mathrm{Var}(i(x_j;Y_j,H_j)) = \mathrm{Var}\left(\frac{T}{2}\log\left(1+\frac{P}{n_t}\|H_j\|^2\right)\right) + \mathrm{Var}\left(\frac{1}{2}\frac{\|H_j x_j\|^2+2H_j x_j Z_j^T-\frac{P}{n_t}\|H_j\|^2\|Z_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right) \tag{3.31}$$

The first term in (3.31) is $T^2\text{Var}(C_{AWGN}(\frac{P}{n_t}\|H\|^2))$. For the second term, we use the "iterated variance" identity. Let $f(x, H, Z)$ represent the second term in (3.31), then

$$\text{Var}(f(x_j, H_j, Z_j)) = \underbrace{\mathbb{E}_{H_j}\text{Var}_{Z_j}(f(x_j, H_j, Z_j)|H_j)}_{1} + \underbrace{\text{Var}_{H_j}\mathbb{E}_{Z_j}[f(x_j, H_j, Z_j)|H_j]}_{2} \tag{3.32}$$

We deal with each piece separately. Recall that $Z_j \sim \mathcal{N}(0, I_T)$ and H_j is isotropically distributed, and both H_j and Z_j are independent from all other variables. First, term 1 is

$$\mathbb{E}_{H_j}\text{Var}_{Z_j}(f(x_j, H_j, Z_j)|H_j) = \frac{1}{4}\mathbb{E}_{H_j}\left[\frac{4\|H_j x_j\|^2 + 2T\left(\frac{P}{n_t}\right)^2\|H_j\|^4}{\left(1+\frac{P}{n_t}\|H_j\|^2\right)^2}\right] \tag{3.33}$$

Massaging this into a more useful form give

$$= \frac{1}{2}\mathbb{E}_{H_j}\left[\frac{\left(2\|H_j x_j\|^2 - 2T\frac{P}{n_t}\|H_j\|^2\right) + \left(2T\frac{P}{n_t}\|H_j\|^2 + T\left(\frac{P}{n_t}\right)^2\|H_j\|^4\right)}{\left(1+\frac{P}{n_t}\|H_j\|^2\right)^2}\right] \tag{3.34}$$

$$= \mathbb{E}_{H_j}\left[\frac{\|H_j x_j\|^2 - T\frac{P}{n_t}\|H_j\|^2}{\left(1+\frac{P}{n_t}\|H_j\|^2\right)^2}\right] + \frac{1}{2}\mathbb{E}_{H_j}\left[\frac{2T\frac{P}{n_t}\|H_j\|^2 + T\left(\frac{P}{n_t}\right)^2\|H_j\|^4}{\left(1+\frac{P}{n_t}\|H_j\|^2\right)^2}\right] \tag{3.35}$$

$$= \mathbb{E}_{H_j}\left[\frac{(\|\tilde{H}_j x_j\|^2 - T\frac{P}{n_t})\|H_j\|^2}{\left(1+\frac{P}{n_t}\|H_j\|^2\right)^2}\right] + \frac{T}{2}\mathbb{E}_{H_j}\left[1 - \left(\frac{1}{1+\frac{P}{n_t}\|H_j\|^2}\right)^2\right] \tag{3.36}$$

Note that the second term here is now $T\mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\|H_j\|^2\right)\right]$. Since H_j is isotropically distributed, $\tilde{H}_j = \frac{H_j}{\|H_j\|}$ is independent from $\|H_j\|^2$. To see this, for scalar $a \in \mathbb{R}$, unit vector $b \in \mathbb{R}^{n_t}$, and orthogonal matrix U:

$$\begin{aligned}\mathbb{P}\left[\|H\|^2 = a \middle| \frac{H}{\|H\|} = b\right] &= \mathbb{P}\left[\|UH\|^2 = a \middle| \frac{UH}{\|UH\|} = b\right] \\ &= \mathbb{P}\left[\|H\|^2 = a \middle| \frac{H}{\|H\|} = U^T b\right]\end{aligned} \tag{3.37}$$

Since U^T acts transitively on the support of $\frac{H}{\|H\|}$ (i.e. unit vectors in $\mathbb{R}^{n_t}$), this quantity is the same for all b's, hence $\|H\|^2$ is independent from $\frac{H}{\|H\|}$. With this, 1 in

(3.32) becomes

$$\mathbb{E}\left[\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right]\mathbb{E}\left[\frac{\|H_j\|^2}{\left(1+\frac{P}{n_t}\|H_j\|^2\right)^2}\right] + T\mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\|H_j\|^2\right)\right] \tag{3.38}$$

We can take the expectation over H_j in $\mathbb{E}\left[\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right]$. Let H_j^l, $l = 1, \ldots, n_t$ be the elements of the vector H_j, and x_j^{lm} be the elements of the matrix x_j, for $l = 1, \ldots, n_t$ and $m = 1, \ldots, T$. Then

$$\mathbb{E}[\|\tilde{H}_j x_j\|^2] = \mathbb{E}\left[\sum_{m=1}^{T}\left(\sum_{l=1}^{n_t}\tilde{H}_j^l x_j^{lm}\right)^2\right] = \sum_{l=1}^{n_t}\sum_{k=1}^{n_t}\mathbb{E}[\tilde{H}_j^l\tilde{H}_j^k]\sum_{m=1}^{T}x_j^{lm}x_j^{km} \tag{3.39}$$

Now, since $\tilde{H}$ is isotropically distributed (in fact, since $\tilde{H}$ is normalized, it is uniform on the sphere S^{n_t-1}), $(\tilde{H}_j^l, H_j^k) \sim (-\tilde{H}_j^k, \tilde{H}_j^l)$, since this is simply a 90 degree rotation. So if $l \neq k$, then $E[\tilde{H}_j^l\tilde{H}_j^k] = -\mathbb{E}[\tilde{H}_j^l\tilde{H}_j^k] = 0$. And since $E[\|\tilde{H}_j\|^2] = 1$, by symmetry $\mathbb{E}\left[\left(\tilde{H_j^l}\right)^2\right] = \frac{1}{n_t}$. Summarizing:

$$\mathbb{E}[\tilde{H}_j^l\tilde{H}_j^k] = \begin{cases}\frac{1}{n_t} & \text{if } k = l\\ 0 & \text{if } k \neq l\end{cases} \tag{3.40}$$

With this, the sum in (3.39) becomes

$$= \frac{1}{n_t}\sum_{l=1}^{n_t}\sum_{m=1}^{T}\left(x_j^{lm}\right)^2 = \frac{\|x_j\|_F^2}{n_t} \tag{3.41}$$

Applying this to (3.38), we recover the term in the proposition

$$\frac{\chi_0}{n_t}\left(\|x_j\|_F^2 - TP\right) + T\mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\|H_j\|^2\right)\right] \tag{3.42}$$

Now we move the 2 in (3.32):

$$\begin{aligned}\mathrm{Var}_{H_j}\mathbb{E}_{Z_j}[f(x_j, H_j, Z_j)|H_j] &= \frac{1}{4}\mathrm{Var}\left(\frac{\|H_j x_j\|^2 - \frac{TP}{n_t}\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right)\\ &= \frac{1}{4}\mathrm{Var}\left(\frac{(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t})\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right)\end{aligned} \tag{3.43}$$

Substituting (3.43) and (3.38) into (3.31) gives the first statement of the proposition.

For the second statement, we decompose (3.43) using the iterated variance identity

once more, this time in terms of $\tilde{H}_j$ and $\|H_j\|^2$. With this, (3.43) becomes

$$= \frac{1}{4}\mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)^2\right] \mathrm{Var}\left(\frac{\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right) + \frac{1}{4}\mathrm{Var}\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)\mathbb{E}^2\left[\frac{\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right] \tag{3.44}$$

Expanding the second variance using $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X]$,

$$= \frac{1}{4}\mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)^2\right] \mathrm{Var}\left(\frac{\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right) \tag{3.45}$$

$$+ \frac{1}{4}\left(\mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)^2\right] - \mathbb{E}^2\left[\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right]\right)\mathbb{E}^2\left[\frac{\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right] \tag{3.46}$$

Combining like terms gives

$$= \frac{1}{4}\mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)^2\right]\mathbb{E}\left[\left(\frac{\|H_j\|^2}{1+\frac{P}{n_t}}\right)^2\right] - \frac{1}{4}\mathbb{E}^2\left[\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right]\mathbb{E}^2\left[\frac{\|H_j\|^2}{1+\frac{P}{n_t}\|H_j\|^2}\right] \tag{3.47}$$

In the notation of the proposition, the first term in (3.47) is

$$\chi_1 \mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)^2\right] \tag{3.48}$$

For the second term in (3.47), we can evaluate the first expectation. Using the same argument as (3.39) - (3.41), we see that

$$\mathbb{E}^2\left[\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right] = \left(\frac{\|x_j\|_F^2}{n_t} - \frac{TP}{n_t}\right)^2 \tag{3.49}$$

And hence (3.47) reduces to

$$\chi_1 \mathbb{E}\left[\left(\|\tilde{H}_j x_j\|^2 - \frac{TP}{n_t}\right)^2\right] - \chi_2\left(\frac{\|x_j\|_F^2}{n_t} - \frac{TP}{n_t}\right)^2 \tag{3.50}$$

As claimed in the Proposition. □

Next, we want to know the *conditional expectation* $Var[i(X;Y,H)|X]$ where X has the distribution of a capacity achieving input distribution. This is simply the expectation over X of the expression from Proposition 2. This quantity will be essential in later chapters and is computed in the follow proposition.

Proposition 3. *For the MISO $n_t \times T$ block fading CSIR channel $P_{YH|X}$, and capacity achieving input X, the conditional variance is given by*

$$\frac{1}{T}\mathrm{Var}[i(X;Y,H)|X] = V_1(P) - \frac{\chi_2}{n_t^2 T}\mathrm{Var}(\|X\|_F^2) \tag{3.51}$$

where $V_1(P)$ is independent of the capacity achieving input distribution X, and

$$V_1(P) \triangleq T\mathrm{Var}\left(C_{AWGN}\left(\frac{P}{n_t}\|H\|^2\right)\right) + \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\|H\|^2\right)\right] + 2\chi_1\left(\frac{P}{n_t}\right)^2 \tag{3.52}$$

And the rest of the notation is from Proposition 2.

Proof. We take the expectation over X in the expression given in Proposition 2. The first two terms are immediate since they do not dependent on X:

$$T\mathrm{Var}\left(C_{AWGN}\left(\frac{P}{n_t}\|H\|^2\right)\right) + \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\|H\|^2\right)\right] \tag{3.53}$$

The third term vanishes since for all CAIDs, $\mathbb{E}[\|X\|_F^2] = TP$, so

$$\frac{\chi_0}{n_t nT}\mathbb{E}[\|X\|_F^2 - nTP] = 0 \tag{3.54}$$

The third term, which is

$$\chi_1\mathbb{E}\left[\left(\|\tilde{H}X\|^2 - \frac{TP}{n_t}\right)^2\right] \tag{3.55}$$

To compute this expectation, we notice $P^*_{Y|H=h} \sim \mathcal{N}(0, (1+\frac{P}{n_t}\|h\|^2)I_T)$ for any CAID. To see this, take the characteristic function of Y to find the distribution of hX for a fixed h:

$$\mathbb{E}[e^{itY}] = \mathbb{E}[e^{itHX}]\mathbb{E}[e^{itZ}] \tag{3.56}$$

Since Z is i.i.d standard normal, it's characteristic function has no zeroes, hence we can solve as

$$\mathbb{E}[e^{itHX}] = \frac{\mathbb{E}[e^{itY}]}{\mathbb{E}[e^{itZ}]} \tag{3.57}$$

The characteristic function of an $\mathcal{N}(0,\sigma^2)$ random variables is $e^{-\frac{1}{2}\sigma^2 t^2}$, hence

$$\mathbb{E}[e^{itHX}] = \frac{e^{\frac{1}{2}(\frac{P}{n_t}\|h\|^2+1)t^2}}{e^{\frac{t^2}{2}}} = e^{\frac{1}{2}\frac{P}{n_t}\|h\|^2 t^2} \tag{3.58}$$

Hence we conclude by the inversion theorem for characteristic functions that $hX \sim$

$\mathcal{N}(0, \frac{P}{n_t}\|h\|^2 I_T)$ for all CAIDs X. This allows us to evaluate (3.55). First, this implies for fixed unit vector $\tilde{h}$,

$$\mathbb{E}[\|\tilde{h}X\|^2] = \frac{TP}{n_t}\|\tilde{h}\|^2 = \frac{TP}{n_t} \tag{3.59}$$

So that (3.55) reduces to

$$\chi_1 \mathbb{E}\left[\left(\|\tilde{H}X\|^2 - \frac{TP}{n_t}\right)^2\right] = \chi_1 \mathbb{E}_{H|X}\mathbb{E}_X\left[\left(\|\tilde{H}X\|^2 - \frac{TP}{n_t}\right)^2 \middle| \tilde{H}\right] \tag{3.60}$$

$$= \chi_1 \mathrm{Var}(\|\tilde{H}X\|^2|H) \tag{3.61}$$

$$= 2\chi_1 T\left(\frac{P}{n_t}\right)^2 \mathbb{E}[\|\tilde{H}\|^4] \tag{3.62}$$

$$= 2\chi_1 T\left(\frac{P}{n_t}\right)^2 \tag{3.63}$$

Where $\tilde{H}^4 = 1$ deterministically. This along with (3.53) gives $V_1(P)$ in the proposition. For the final term, from Propsition 2,

$$-\chi_2 \mathbb{E}\left[\left(\frac{\|X\|_F^2}{n_t} - \frac{TP}{n_t}\right)^2\right] \tag{3.64}$$

And since $\mathbb{E}[\|X\|_F^2] = TP$ for any CAID, this becomes

$$-\frac{\chi_2}{n_t^2}\mathrm{Var}(\|X\|_F^2) \tag{3.65}$$

Which completes the proof of the Proposition. □

So we've derived the expression for $\mathrm{Var}[i(X;Y,H)|X]$. We will show that this dispersion is achievable. Proposition (3) shows that this dispersion depends on the CAID through $-\mathrm{Var}(\|X\|^2)$. Hence to minimize the dispersion, we want the CAID that *maximizes* $\mathrm{Var}(\|X\|^2)$. In Chapter 4 we will show achievability, and in Chapter 5 we will minimize the dispersion over all CAIDs.

Chapter 4

Achievable Dispersion

The goal of this section is to prove that a class of dispersions are achievable for the coherent MISO fading channel. The proof relies on the $\kappa\beta$ bound [9, Theorem 25]. To understand this bound, first we must give some definitions, notation, and properties for binary and composite hypothesis tests (Section 4.1). Then in Section 4.2 we state and prove the main achievability theorem.

4.1 Binary and Composite Hypothesis Testing

Many finite blocklength results are derived by considering an optimal hypothesis between appropriate distributions. A *binary hypothesis test* $P_{Z|W} : \mathcal{W} \to \{0, 1\}$ is a test that, given a sample w from a space $\mathcal{W}$, chooses (perhaps non-deterministically) one of two distributions P or Q that could have generated w. $Z = 1$ indicates that the test choose P to be the true distribution, while $Z = 0$ indicates the test chooses Q. This is sometimes written as

$$H_0 : W \sim Q \tag{4.1}$$
$$H_1 : W \sim P \tag{4.2}$$

Two types of errors can be made in a binary hypothesis test: we can mistakenly choose P when Q is the actual distribution, or we can choose Q when P is the true distribution. These errors depends on the choice of test $P_{Z|W}$, and in general are asymmetric. Here we will use the convention that we always consider the error when the test chooses P when the actual distribution is Q.

We define $\beta_\alpha(P, Q)$ to be the minimum error probability of all statistical tests $P_{Z|W}$ between distributions P and Q, given that the test chooses P when P is correct with at least probability α. Formally:

$$\beta_\alpha(P, Q) = \inf_{P_{Z|W} : \int_{\mathcal{W}} P_{Z|W}(1|w) dP(w) \geq \alpha} \int_{\mathcal{W}} P_{Z|W}(1|w) dQ(w) \tag{4.3}$$

The *Neyman Pearson Lemma* tells us that for a given α, an optimal test $P^*_{Z|W}$

achieving error β_α exists, and has the form of a ratio test, i.e.

$$\beta_\alpha(P, Q) = Q\left[\frac{dP}{dQ} > \gamma\right] \tag{4.4}$$

Where γ is chosen to satisfy

$$\alpha = P\left[\frac{dP}{dQ} > \gamma\right] \tag{4.5}$$

In a *composite hypothesis test*, P and Q are now parametric families of distributions, $\{P_{\theta_1}\}_{\theta_1 \in \Theta_1}$ and $\{Q_{\theta_2}\}_{\theta_2 \in \Theta_2}$, i.e.

$$H_0 : W \sim Q_{\theta_2} \text{ s.t. } \theta_2 \in \Theta_2 \tag{4.6}$$
$$H_1 : W \sim P_{\theta_1} \text{ s.t. } \theta_1 \in \Theta_1 \tag{4.7}$$

In words: the test sees a sample w and must decide whether the distribution generating that sample was from the P_{θ_1} family or the Q_{θ_2} family. Similar to the binary hypothesis testing case, we denote the minimum error probability of a test $P_{Z|W}$ given that the test chooses H_1 when H_1 is true *for any* $\theta_1 \in \Theta_1$ with at least probability τ. Formally:

$$\kappa_\tau(\Theta_1, \Theta_2) = \inf_{P_{Z|W} : \inf_{\theta_1 \in \Theta_1} \{\int_{\mathcal{W}} P_{Z|W}(1|w) dP_{\theta_1}(w) \geq \tau\}} \sup_{\theta_2 \in \Theta_2} \int_{\mathcal{W}} P_{Z|W}(1|w) dQ_{\theta_2} \tag{4.8}$$

Our main case of interest will be between the set of distributions $\{P_{Y|X=x}\}_{x \in F}$ and a single distribution Q_Y. We will denote the minimum error probability in the composite hypothesis test in this case as $\kappa_\tau(F, Q_Y)$.

Now that we have the basic definintions, we'll need a few bounds that will be used in the next section. First, we can lower bound the minimum error in a composite hypothesis test by the minimum error of a binary hypothesis test. This is useful because often it is difficult to evaluate κ_τ, but for β_α the Neyman-Pearson lemma gives us the form of the optimal test.

Lemma 4. *For a composite hypothesis test between $\{P_{Y|X=x}\}_{x \in F}$ and Q_Y, and any distribution $P_{\tilde{X}}$ such that $F \subset supp(\tilde{X}) = A$,*

$$\kappa_\tau(F, Q_Y) \geq \beta_{\tau P_{\tilde{X}}[F]}(P_{\tilde{X}} \circ P_{Y|X}, Q_Y) \tag{4.9}$$

Here, $P_X \circ P_{Y|X} \triangleq \int P_{Y|X=x} dP_X(x)$.

Proof. Let $P_{Z|Y}$ be any test for the composite hypothesis test between $\{P_{Y|X=x}\}_{x \in F}$ and Q_Y satisfying

$$\inf_{x \in F} \sum_{y \in B} P_{Y|X}(y|x) P_{Z|Y}(1|y) \geq \tau \tag{4.10}$$

Where $Z = 1$ indicates the test choses $\{P_{Y|X=x}\}_{x \in F}$. Then we use this test $P_{Z|Y}$ for

testing P_Y vs Q_Y, where now $Z = 1$ indicates the test chooses P_Y. The corresponding probability of choosing P_Y when P_Y is correct is (note $P_Y = P_X \circ P_{Y|X}$ by assumption)

$$\sum_{y\in B} P_Y(y)P_{Z|Y}(1|y) = \sum_{y\in B}\left(\sum_{x\in A} P_X(x)P_{Y|X}(y|x)\right)P_{Z|Y}(1|y) \tag{4.11}$$

$$\geq \sum_{y\in B}\left(\sum_{x\in F} P_X(x)P_{Y|X}(y|x)\right)P_{Z|Y}(1|y) \tag{4.12}$$

$$\geq \sum_{x\in F} P_X(x)\left(\inf_{x\in F}\sum_{y\in B} P_{Y|X}(y|x)P_{Z|Y}(1|y)\right) \geq P_X[F]\tau \tag{4.13}$$

Since this hold for all tests $P_{Z|Y}$ for the composite HT, it holds for the test achieving κ_τ. Since $\beta_{\tau P_X[F]}$ lower bounds the $\pi_{1|0}$ error of all test for P_Y vs Q_Y, it lower bounds κ_τ. □

Furthermore, we can lower bound β_α from a binary hypothesis test in terms of the divergence between $D(P||Q)$ using the data processing inequality:

Lemma 5. *For all distributions P, Q s.t. $P \ll Q$, and all $\alpha \in [0, 1]$,*

$$\beta_\alpha(P, Q) \geq \exp\left(-\frac{D(P||Q) + h_B(\alpha)}{\alpha}\right) \tag{4.14}$$

Proof. Use the data processing inequality with the kernel $P_{Z|W}$ from our hypothesis test:

$$D(P||Q) \geq d(\alpha||\beta_\alpha) = -h(\alpha) + \alpha\log\frac{1}{\beta} + (1-\alpha)\log\frac{1}{1-\beta} \geq -h(\alpha) + \alpha\log\frac{1}{\beta} \tag{4.15}$$

Where $d(p||q)$ is the divergence between a Bernoulli(p) and Bernoulli(q) distribution. The lemma follows from solving for β_α. □

Finally, we are interested the case when P and Q are product distribution $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$. When this is the case, with a few regularity conditions we can expand β_α in terms of it's dependence on n by the following lemma from [11, Lemma 14], which we give here

Lemma 6. *Let $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$ with $P_i \ll Q_i$ be two measures on a measurable space $\mathcal{A}^n$ such that the third moment of $\log\frac{dP}{dQ}$ is bounded, then*

$$\log\beta_\alpha(P, Q) = -nD_n - \sqrt{nV_n}Q^{-1}(\alpha) + o(\sqrt{n}) \tag{4.16}$$

Where

$$D_n = \frac{1}{n}\sum_{i=1}^{n} D(P_i||Q_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\log \frac{dP_i}{dQ_i}\right] \tag{4.17}$$

$$V_n = \frac{1}{n}\sum_{i=1}^{n} V(P_i||Q_i) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\left(\log \frac{dP_i}{dQ_i}\right) \tag{4.18}$$

The proof is an application of the Berry-Esseen Theorem, which quantifies the error in approximating the CDF of a sum of independent random variables by a Gaussian distribution as in the Central Limit Theorem.

4.2 Achievability Theorem

In this section, we prove the achievability coding theorem for the coherent MISO channel. To prove achievability for channels with input constraints, the best known method is to use the $\kappa\beta$-bound from [9], restated here:

Theorem 7 ($\kappa\beta$-bound)**.** *For any distribution Q_Y on $\mathcal{B}$, and ϵ, τ such that $0 < \tau < \epsilon < 1$, there exists an (M, ϵ) code with codewords in $F \subset A$ (A is the input space of the channel) satisfying*

$$M \geq \frac{\kappa_\tau(F, Q_Y)}{\sup_{x\in F} \beta_{1-\epsilon+\tau}(P_{Y|X=x}, Q_Y)} \tag{4.19}$$

The "art" of applying this theorem is in choosing F and Q_Y appropriately. The intuition in choosing these is as follows: although we know the distributions in the collection $\{P_{Y|X=x}\}_{x\in F}$, we don't know which x is actually was sent, so if Q_Y is in the "center" of the collection, then the two hypotheses can be difficult to distinguish, making the numerator large. However, for a given x, $P_{Y|X=x}$ vs Q_Y may still be easily to distinguish, making the denominator small. The main principle for applying the $\kappa\beta$-bound is thus: *Choose F and Q_Y such that $P_{Y|X=x}$ vs Q_Y is easy to distinguish for any given x, yet $\{P_{Y|X=x}\}_{x\in F}$ vs Q_Y is hard to distinguish.*

The main theorem of this section gives achievable rates for the coherent MISO fading channel:

Theorem 8 (Achievability)**.** *For the coherent MISO fading channel*

$$\log M^*(nT, \epsilon, P) \geq nTC(P) - \sqrt{nTV(P)}Q^{-1}(\epsilon) + o(\sqrt{n}) \tag{4.20}$$

Where for any $n_t \times T$ capacity achieving input distribution X,

$$C(P) = \frac{1}{2}\mathbb{E}\left[\log\left(1 + \frac{P}{n_t}\|H\|^2\right)\right] \tag{4.21}$$

$$V(P) = \frac{1}{T}\mathrm{Var}[i(X; Y, H)|X] \tag{4.22}$$

We first give the main structure of the proof, then the more technical steps follow as lemmas.

Proof. We apply the $\kappa\beta$ bound (4.19). First we bound the numerator $\kappa_\tau(F_n, Q_{Y^n})$. From Lemmas 4 and 5, we can lower bound $\kappa_\tau(F_n, Q_{Y^n})$ for any F_n, $P_{\tilde{X}^n}$ and Q_{Y^n} satisfying the above lemmas by

$$\kappa_\tau(F_n, Q_{Y^n}) \geq \exp\left(-\frac{D(P_{\tilde{X}^n} \circ P_{Y^n H^n|X^n} || Q_{Y^n}) + \log 2}{\tau P_{\tilde{X}^n}[F_n]}\right) \tag{4.23}$$

This allows us to lower bound κ_τ, by upper bounding $D(P_{\tilde{X}^n} \circ P_{Y^n H^n|X^n} || Q_{Y^n})$. The next lemma shows that for a certain choice of $P_{\tilde{X}^n}$ and Q_{Y^n}, the divergence in (4.23) converges to a constant as $n \to \infty$, hence κ_τ is bounded away from zero for all n. Note that choosing $P_{\tilde{X}^n}$ to be i.i.d. will never work if Q_{Y^n} is i.i.d., since the divergence will go to infinity. Instead, we choose $P_{\tilde{X}^n}$ to be on a "shell" around Q_Y.

Lemma 9. *Let X be any CAID for the channel. For*

$$P_{\tilde{X}^n} \sim \frac{X^n}{\|X^n\|_F}\sqrt{nTP} \tag{4.24}$$

and Q_Y the (unique) CAOD ($Q_Y = P_X \circ P_{Y|X}$), we have as $n \to \infty$

$$D(P_{\tilde{X}^n} \circ P_{Y^n H^n|X^n} || Q_Y^n) \to O(1) \tag{4.25}$$

For the lower bound (4.23) to be meaningful, we must have $P_{\tilde{X}^n}[F_n]$ bounded away from zero. To choose the set F_n appropriately, we look to the denominator in the $\kappa\beta$ bound. Using Lemma 6, we can expand the denominator in terms of n as follows:

$$\begin{aligned} -\sup_{x \in F_n} \log \beta_\alpha(P_{Y^n H^n|X^n=x^n}, Q_{Y^n H^n}) &= -\sup_{x \in F_n}\left(-nD_n(x^n) + \sqrt{nV_n(x^n)}Q^{-1}(\alpha) + o(\sqrt{n})\right) \\ &= \inf_{x \in F_n}\left(nD_n(x^n) - \sqrt{nV_n(x^n)}Q^{-1}(\alpha) + o(\sqrt{n})\right) \end{aligned} \tag{4.26}$$

Where now D_n and V_n depends on x^n through $P_{Y^n H^n|X^n=x^n}$. We use (4.26) to help choose the set F_n. From the cost constraint on the channel, $F_n \subset \{x^n : \|x^n\|_F^2 \leq nTP\}$. Considering the first term

$$D_n(x^n) = \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\log \frac{P_{Y_j|H_j,X_j=x_j}(Y_j|H_j)}{P^*_{Y_j|H_j}(Y_j|H_j)}\right] = \frac{1}{n}\sum_{j=1}^n \mathbb{E}[i(x_j, Y_j, H_j)] \tag{4.27}$$

The expression for the information density is given in (3.19). As shown in (3.39)-(3.41), $\mathbb{E}[\|\tilde{H}_j x_j\|^2] = \frac{\|x_j\|_F^2}{n_t}$, so we discover the dependence on the input x^n in $D_n(x^n)$

$$= T\mathbb{E}\left[C_{AWGN}\left(\frac{P}{n_t}\|H_j\|^2\right)\right] + \frac{c_0}{2n}\left(\frac{\|x^n\|_F^2}{n_t} - \frac{nTP}{n_t}\right) \tag{4.28}$$

Where $c_0 > 0$ is a constant that doesn't depend on x^n. In order to maximize the $O(n)$ term in our expansion, we require $F_n \subset \{x^n : \|x^n\|_F^2 = nTP\}$ (i.e. x^n lies on the surface of a "sphere") on which $D_n(x^n)$ is $TC(P)$. Now, let V represent our "target" dispersion, and choose F_n so that (4.26) is forced to give this target dispersion. To do this, for arbitrary $\delta > 0$, take F_n to be

$$F_n = \{x^n : \|x^n\|_F^2 = nTP\} \cap \{x^n : \frac{1}{T} V_n(x^n) \leq V + \delta\} \tag{4.29}$$

Where

$$V_n(x^n) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{Var}(i(x_j; Y_j, H_j)) \tag{4.30}$$

Then (4.26) becomes

$$-\sup_{x \in F} \log \beta_\alpha(P_{Y^n H^n | X^n = x^n}, Q_{Y^n H^n}) \leq nTC(P) - \sqrt{nT(V+\delta)} Q^{-1}(\epsilon - \tau) + o(\sqrt{n}) \tag{4.31}$$

We require that F_n from (4.29) satisfies $P_{\tilde{X}}[F_n] > c > 0$, which is stated as the following lemma

Lemma 10. *For X CAID, $\tilde{X}^n$ distributed as $\frac{X^n}{\|X^n\|_F}\sqrt{nTP}$,*

$$\mathbb{P}\left[\left|\frac{1}{T} V_n(\tilde{X}^n) - V(P)\right| \geq \delta\right] \to 0 \quad as \quad n \to \infty \tag{4.32}$$

Where $V(P) = \frac{1}{T}\mathrm{Var}(i(X; Y, H)|X)$.

Applying this lemma shows that $P_{\tilde{X}^n}[F_n]$ tends to 1 asymptotically

$$P_{\tilde{X}^n}[F_n] = \mathbb{P}\left[\frac{1}{T} V(\tilde{X}^n) \leq V(P) + \delta\right] \to 1 \tag{4.33}$$

Thus for large enough n, there exists a constant $k_1 > 0$ such that

$$\kappa_\tau \geq \exp\left(-\frac{k_1}{\tau}\right) \tag{4.34}$$

All together, the $\kappa\beta$ bound gives

$$\log M^*(nT, \epsilon, P) \geq nTC(P) - \sqrt{nT(V(P)+\delta)} Q^{-1}(\epsilon - \tau) + o(\sqrt{n}) - \frac{k_1}{\tau} \tag{4.35}$$

Since δ and τ are arbitrary, we recover the statement in the Theorem. □

Now we prove the two lemmas used in the Theorem.

Proof of Lemma 9. We want to show, for X CAID and $\tilde{X} = \frac{X}{\|X\|_F}\sqrt{nTP}$, and channel law $P_{Y^n H^n|X^n}$, that the divergence between the output distribution induced by $P_{\tilde{X}}$ and Q_Y is asymptotically upper bounded by some constant:

$$D(P_{\tilde{X}^n} \circ P_{Y^n H^n|X^n} || P_{X^n} \circ P_{Y^n H^n|X^n}) \to O(1) \tag{4.36}$$

Here, $P_{\tilde{X}^n} \circ P_{Y^n H^n|X^n}$ describes the output distribution of the "modified" channel

$$Y_j = H_j \frac{X_j}{\|X_j\|_F}\sqrt{nTP} + Z_j, \qquad j = 1, \ldots, n \tag{4.37}$$

We can equivalently describe this distribution on (Y^n, H^n) through this modified channel in terms of the CAID X by

$$P_{\tilde{X}^n} \circ P_{Y^n H^n|X^n} = P_{X^n} \circ \tilde{P}_{Y^n H^n|X^n} \tag{4.38}$$

Where now

$$\tilde{P}_{Y^n H^n|X^n = x^n}(y^n, h^n) \sim P_{H^n}(h^n) \mathcal{N}\left(\frac{\sqrt{nTP}}{\|x^n\|_F}\left[h_1 x_1, \ldots, h_n x_n\right], I_{nT}\right) \tag{4.39}$$

With this representation, the divergence in (4.36) takes a simpler form

$$D(P_{X^n} \circ \tilde{P}_{Y^n H^n|X^n} || P_{X^n} \circ P_{Y^n H^n|X^n}) = D(\tilde{P}_{Y^n H^n|X^n} || P_{Y^n H^n|X^n} | P_{X^n}) \tag{4.40}$$

This is the average over the divergence between two standard normal distributions with different means. Recall that for any means $a, b \in \mathbb{R}^n$,

$$D\left(\mathcal{N}(a, I_n) || \mathcal{N}(b, I_n)\right) = \frac{1}{2}\|a - b\|^2 \tag{4.41}$$

Where the norm is the standard Euclidean norm. Applying this to (4.40), i.e. to

$$D\left(\mathcal{N}\left(\frac{\sqrt{nTP}}{\|x^n\|_F}\left[h_1 x_1, \ldots, h_n x_n\right], I_{nT}\right) \middle|\middle| \mathcal{N}\left(\left[h_1 x_n, \ldots, h_n x_n\right], I_{nT}\right) \middle| P_{X^n}\right) \tag{4.42}$$

Yields

$$\frac{1}{2}\mathbb{E}\left[\left(\|X^n\|_F - \sqrt{nTP}\right)^2 \frac{\|[H_1 X_1, \ldots, H_n X_n]\|^2}{\|X^n\|_F^2}\right] \tag{4.43}$$

We argue that (4.43) converges to a constant as $n \to \infty$. Taking the expectation over H^n while holding X^n fixed gives

$$\frac{1}{2}\mathbb{E}\left[\frac{\left(\|X^n\|_F - \sqrt{nTP}\right)^2}{\|X^n\|_F^2}\mathbb{E}\left[\|[H_1 X_1, \ldots, H_n X_n]\|^2 \middle| X^n\right]\right] \tag{4.44}$$

As computed in (3.39)-(3.41), the inner expectation simplifies to

$$\mathbb{E}\left[\|[H_1X_1,\ldots,H_nX_n]\|^2\middle|X^n=x^n\right]=\frac{\|x^n\|_F^2}{n_t} \tag{4.45}$$

Hence (4.43) becomes

$$\frac{1}{2n_t}\mathbb{E}\left[\left(\|X^n\|_F-\sqrt{nTP}\right)^2\right] \tag{4.46}$$

We show that this converges to a constant for any CAID X. The basic idea is that this behaves like the variance of a χ_n random variable, which converges to a constant as $n\to\infty$. First, we expand as

$$=\frac{1}{2}\left(\mathrm{Var}(\|X^n\|_F)+\left(\sqrt{nTP}-\mathbb{E}[\|X^n\|_F]\right)^2\right) \tag{4.47}$$

For the first term here, we use the concavity of the square root to get an upper bound in terms of a χ_n random variable. Expanding the first term gives

$$\mathrm{Var}(\|X^n\|_F)=\mathbb{E}[\|X^n\|_F^2]-\mathbb{E}[\|X^n\|_F]^2=nTP-\mathbb{E}[\|X^n\|_F]^2 \tag{4.48}$$

Now, for any CAID X, let X_i^{jk} be the (j,k)-th entry in the i-th block of the matrix X^n, then

$$\mathbb{E}[\|X^n\|_F]^2=\mathbb{E}\left[\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n_t}\sum_{k=1}^{T}(X_i^{jk})^2}\right]^2 \tag{4.49}$$

$$=n_tT\mathbb{E}\left[\sqrt{\sum_{j=1}^{n_t}\sum_{k=1}^{T}\frac{1}{n_tT}\sum_{i=1}^{n}(X_i^{jk})^2}\right]^2 \tag{4.50}$$

$$\geq n_tT\left(\sum_{j=1}^{n_t}\sum_{k=1}^{T}\frac{1}{n_tT}\mathbb{E}\left[\sqrt{\sum_{i=1}^{n}(X_i^{jk})^2}\right]\right)^2 \tag{4.51}$$

Now since X^n is block i.i.d. (each block with distribution X), and each entry of X^n has distribution $\mathcal{N}(0,\frac{P}{n_t})$, the sequence $(X_1^{jk},\ldots,X_n^{jk})$ is i.i.d. $\mathcal{N}(0,\frac{P}{n_t})$. Letting μ_n be the mean of χ_n distribution (i.e. $\mu_n=\sqrt{2}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$), we find

$$\mathbb{E}[\|X^n\|_F]^2\geq n_tT\frac{P}{n_t}\mu_n^2=\mu_n^2TP \tag{4.52}$$

Applying this bound to (4.48) gives

$$\mathrm{Var}(\|X^n\|_F)\leq nTP-\mu_n^2TP\to\frac{TP}{2} \tag{4.53}$$

(This is simply the variance of a χ_n random variable). Now, the second term in (4.47) goes to zero since

$$\mathrm{Var}(\|X^n\|_F) = (\sqrt{nTP} - \mathbb{E}[\|X^n\|_F])(\sqrt{nTP} + \mathbb{E}[\|X^n\|_F]) \to \frac{TP}{2} \tag{4.54}$$

$$\implies \sqrt{nTP} - \mathbb{E}[\|X^n\|_F] \to 0 \tag{4.55}$$

Hence we've shown that (4.46) convergence to a constant, which completes the proof. □

Proof of Lemma 10. We want to show, for any CAID X, and where $\tilde{X}^n \sim \frac{X^n}{\|X\|_F}\sqrt{nTP}$,

$$\mathbb{P}\left[V_n(\tilde{X}^n) - V(P) \geq \delta\right] \to 0 \quad \text{as} \quad n \to \infty \tag{4.56}$$

Where

$$V_n(\tilde{X}^n) = \frac{1}{nT}\mathrm{Var}\left(i(\tilde{X}^n; Y^n, H^n)|\tilde{X}^n\right) \tag{4.57}$$

$$V(P) = \frac{1}{T}\mathrm{Var}\left((i(X; Y, H)|X\right) = \mathbb{E}[V_1(X)] \tag{4.58}$$

The key is that as $n \to \infty$, $\|X^n\|_F \to \sqrt{nTP}$ (by LLN), so that $\frac{\tilde{X}_j}{\|X\_F}\sqrt{nTP} \to X_j$ □

4.3 Error Exponents

In this section, we show that Gallager's Error Exponent [5, Section 7.3] analysis is consistent with our result that orthogonal design input distributions preform better from a finite blocklength perspective. We will introduce error exponents for input constrained channel, compute these exponents for both the i.i.d. Gaussian input and orthogonal design input, then compare the two to see that the exponent for orthogonal designs is always larger.

The method of error exponents give a non-asymptotic upper bound (achievability bound) on the probability of error when using $M = e^{nR}$ codewords and blocklength n. Gallager shows us that for an arbitrary discrete-time memoryless channel with input constraint $\sum_{i=1}^n f(x_i) \leq nP$, exists an (n, e^{nR}, ϵ) code satisfying

$$\epsilon \leq \left(\frac{e^{r\delta}}{\mu}\right)^{1+\rho} \exp(-n(E_0(\rho, P_X, r) - \rho R)) \tag{4.59}$$

$E_0(\rho, P_X, r)$ is called the *random coding exponent* defined below, and the prefactor $(e^{r\delta}/\mu)^{1+\rho}$ grows with n as $n^{(1+\rho)/2}$ for fixed r and δ, so it does not effect the exponential dependence of the above bound. The tightest bound on error probability in (4.59) is given by maximizing over $\rho \in (0, 1)$, $r \geq 0$, and P_X satisfying the input

constraint. The bound exponential dependence on the input distribution P_X is only though $E_0(\rho, P_X, r)$, given as follows:

Definition 4. For $\rho \in [0,1]$, $r \geq 0$, input P_X and channel $P_{Y|X}$ with power constraint $\sum_{i=1}^{n} f(x_i) \leq nP$, the random coding exponent is

$$E_0(\rho, P_X, r) = -\log \int \left(\int P_X(x) e^{r(f(x)-P)} P_{Y|X}(y|x)^{\frac{1}{1+\rho}} dx \right)^{1+\rho} dy \tag{4.60}$$

The following argument will show that in the MISO block fading channel, taking input P_X as an orthogonal design input distribution gives a strictly better random coding exponent than the independent Gaussian input, for all values of ρ and r. This agrees with our result in Chapter 5 that orthogonal designs preform better from a finite blocklength standpoint. Hence this result could have been discovered with the classical method of error exponents rather than the more modern method of dispersion.

4.3.1 Computation of i.i.d. Gaussian and Orthogonal Design Exponents

Here we derive the expressions for the error exponent when the MISO channel input is i.i.d. Gaussian and when it is an orthogonal design. The full computation is given for completeness. The following proposition gives the expression for the random coding exponent for both the i.i.d. Gaussian input and the orthogonal design input.

Proposition 11. *The error exponents for the i.i.d. Gaussian input and the Orthogonal Design input are given by, respectively*

$$E_0(\rho, P_{X_G}, r) = rTP(1+\rho) + \frac{T(n_t + \rho n_t - \rho)}{2} \log\left(1 - 2r\frac{P}{n_t}\right) + \frac{\rho T}{2} \mathbb{E} \log\left(1 - 2r\frac{P}{n_t} + \frac{\frac{P}{n_t}\|H\|^2}{1+\rho}\right) \tag{4.61}$$

and

$$E_0(\rho, P_{X_{OD}}, r) = rT\frac{P}{n_t}(1+\rho) + \frac{T}{2} \log\left(1 - 2r\frac{P}{n_t}\right) + \frac{\rho T}{2} \mathbb{E} \log\left(1 - 2r\frac{P}{n_t} + \frac{\frac{P}{n_t}\|H\|^2}{1+\rho}\right) \tag{4.62}$$

Proof. The objective is to compute, for the respective input distributions,

$$E_0(\rho, P_X, r) = -\log \int \left(\int P_X(x) e^{\|x\|_F^2 - TP)} P_{YH|X}(y, h|x)^{\frac{1}{1+\rho}} dx \right)^{1+\rho} dy dh \tag{4.63}$$

Note that we can decompose the channel transition kernel into $P_H P_{Y|HX}$ and pull out

P_H to get to a simpler form

$$-\mathbb{E}_H \log \int \left(\int P_X(x) e^{r(\|x\|_F^2 - TP)} P_{Y|HX}(y|h,x)^{\frac{1}{1+\rho}} dx \right)^{1+\rho} dy \tag{4.64}$$

i.i.d. Gaussian input: Throughout, we will compute the integrals by viewing them as finding the output distribution through a "modified" channel. The "tilted" channel, $P_{Y|HX}^{\frac{1}{1+\rho}}$, has conditional density

$$P_{Y|HX}^{\frac{1}{1+\rho}} = \frac{1}{\sqrt{2\pi}^{T(1+\rho)}} e^{\frac{-\|y-hx\|^2}{2(1+\rho)}} = \sqrt{1+\rho}^{T} \sqrt{2\pi}^{\frac{\rho T}{1+\rho}} \mathcal{N}\left(hx, (1+\rho) I_T\right) \tag{4.65}$$

We can think of this as instead the channel

$$Y = HX + \sqrt{1+\rho} Z \tag{4.66}$$

The input P_{X_G} is also tilted:

$$P_{X_G} e^{r(\|x\|_F^2 - TP)} = e^{r(\|x\|_F^2 - TP)} \frac{1}{\sqrt{2\pi \frac{P}{n_t}}^{n_t T}} e^{-\frac{\|x\|_F^2}{2\frac{P}{n_t}}} = \frac{e^{-rTP}}{\sqrt{1 - 2r\frac{P}{n_t}}^{n_t T}} \mathcal{N}\left(0, \frac{\frac{P}{n_t}}{1 - 2r\frac{P}{n_t}} I_T \otimes I_{n_t}\right) \tag{4.67}$$

Here, $\otimes$ denotes the tensor product. The inner integral of (4.64) is the output distribution through channel (4.66) induced by the i.i.d Gaussian input, each entry having variance $\frac{P}{n_t} \frac{1}{1-2r\frac{P}{n_t}}$, so

$$\int P_X(x) e^{r(\|x\|_F^2 - TP)} P_{Y|HX}(y|h,x)^{\frac{1}{1+\rho}} dx = c_0 \mathcal{N}\left(0, \left(1 + \rho + \frac{\frac{P}{n_t}\|h\|^2}{1 - 2r\frac{P}{n_t}}\right) I_T\right) \tag{4.68}$$

Where the constant c_0 is

$$c_0 = \frac{e^{-rTP}}{\sqrt{1 - 2r\frac{P}{n_t}}^{n_t T}} \sqrt{1+\rho}^{T} \sqrt{2\pi}^{\frac{\rho T}{1+\rho}} \tag{4.69}$$

For the outer integral from (4.64), again we extract the appropriate constant to obtain the integral of a Gaussian density. Let P_Y be the normal distribution from (4.68), then the outer integral is

$$c_0^{1+\rho} \int P_Y^{1+\rho} dy = c_0^{1+\rho} c_1 \tag{4.70}$$

With constant

$$c_1 = \left(\sqrt{2\pi v}^{\rho T}\sqrt{1+\rho}^{T}\right)^{-1} \tag{4.71}$$

And v is the variance of P_Y, i.e.

$$v = 1 + \rho + \frac{\frac{P}{n_t}\|h\|^2}{1 - 2r\frac{P}{n_t}} \tag{4.72}$$

Putting the constants together, the error exponent for the Gaussian input is

$$E_0(\rho, P_{X_G}, r) = -\mathbb{E}\log\left(c_0^{1+\rho}c_1\right) \tag{4.73}$$

$$= \mathbb{E}\log\left(\frac{\sqrt{1+\rho}^{T}\sqrt{2\pi v}^{\rho T}}{\sqrt{1+\rho}^{\rho T+T}\sqrt{2\pi}^{\rho T}}\sqrt{1-2r\frac{P}{n_t}}^{n_t T(1+\rho)} e^{rTP(1+\rho)}\right) \tag{4.74}$$

$$= rTP(1+\rho) + \frac{n_t T(1+\rho)}{2}\log\left(1-2r\frac{P}{n_t}\right) + \mathbb{E}\frac{\rho T}{2}\log\left(1+\frac{\frac{P}{n_t}\|H\|^2}{(1-2r\frac{P}{n_t})(1+\rho)}\right) \tag{4.75}$$

A trivial rearrangement of this gives (4.61) from the proposition. Note this expression is only valid when $0 \le r \le \frac{1}{2\frac{P}{n_t}}$.

Orthogonal Design input: The orthogonal design error exponent even easier. When the input is an orthogonal design, the channel is equivalent to a SISO block fading channel. More concretely, if an orthogonal design of dimension $n_t \times T$ is used for the MISO block fading channel with power constraint TP and fading vector H, the equivalent SISO channel has the same coherence time T, but now with power constraint $\sum_{j=1}^{n}\|X_j\|^2 \le nT\frac{P}{n_t}$ and scalar fading coefficient $\|H\|$ for a length T block of symbols, i.e. one block of length T is given by

$$Y_j = \|H\|X_j + Z_j \tag{4.76}$$

Where $X_j \sim \mathcal{N}(0, \frac{P}{n_t}I_T)$. Hence we can calculate the error exponent of this SISO channel to give the error exponent for orthogonal design inputs in the MISO channel.

The steps are very similar to the i.i.d. Gaussian case above. The tilted channel is

$$P_{Y|HX}^{\frac{1}{1+\rho}} = \sqrt{1+\rho}^{T}\sqrt{2\pi}^{\frac{\rho T}{1+\rho}}\mathcal{N}(\|h\|x, (1+\rho)I_T) \tag{4.77}$$

The tilted input distribution is

$$P_X e^{r(\|x\|^2 - T\frac{P}{n_t})} = \frac{e^{-rT\frac{P}{n_t}}}{\sqrt{1-2r\frac{P}{n_t}}^{T}}\mathcal{N}\left(0, \frac{\frac{P}{n_t}}{1-2r\frac{P}{n_t}}I_T\right) \tag{4.78}$$

With these, the inner integral of (4.64) is

$$\int P_X(x)e^{r(\|x\|_F^2-TP)}P_{Y|HX}(y|h,x)^{\frac{1}{1+\rho}}dx = c_0\mathcal{N}\left(0,\left(1+\rho+\frac{\frac{P}{n_t}\|h\|^2}{1-2r\frac{P}{n_t}}\right)I_T\right) \quad (4.79)$$

Where

$$c_0 = \frac{e^{-rT\frac{P}{n_t}}}{\sqrt{1-2r\frac{P}{n_t}}^T}\sqrt{1+\rho}^T\sqrt{2\pi}^{\frac{\rho T}{1+\rho}} \quad (4.80)$$

The outer integral of (4.64) is, with P_Y as the Gaussian piece of (4.79),

$$c_0^{1+\rho}\int P_Y^{1+\rho}dy = c_0^{1+\rho}\int \frac{1}{\sqrt{2\pi v}^{(1+\rho)T}}e^{-\frac{\|y\|^2}{2v\frac{1}{1+\rho}}} = c_0^{1+\rho}c_1 \quad (4.81)$$

Where

$$c_1 = \left(\sqrt{2\pi v}^{\rho T}\sqrt{1+\rho}^T\right)^{-1} \quad (4.82)$$

And v is the same variance as (4.72). Putting the constants together, the error exponent is

$$E_0(\rho, P_{X_{OD}}, r) = -\mathbb{E}\log(c_0^{1+\rho}c_1) \quad (4.83)$$

$$= \mathbb{E}\log\left(\frac{\sqrt{2\pi v}^{\rho T}\sqrt{1+\rho}^T}{\sqrt{2\pi}^{\rho T}\sqrt{1+\rho}^{T+\rho T}}\sqrt{1-2r\frac{P}{n_t}}^{T(1+\rho)}e^{rT\frac{P}{n_t}(1+\rho)}\right) \quad (4.84)$$

$$= rT\frac{P}{n_t}(1+\rho)+\frac{T}{2}\log\left(1-2r\frac{P}{n_t}\right)+\mathbb{E}\frac{\rho T}{2}\log\left(1-2r\frac{P}{n_t}+\frac{\frac{P}{n_t}\|H\|^2}{1+\rho}\right) \quad (4.85)$$

Which matches expression (4.62) in the proposition. Note this expression is only valid when $0 \le r \le \frac{1}{2\frac{P}{n_t}}$. □

4.3.2 Comparison of Error Exponents

We're interested whether the i.i.d. Gaussian input or the orthogonal design input give a larger random coding exponent. It turns out that the orthogonal design exponent is larger for all ρ, r, and P, as we show now.

We will show $E_0(\rho, P_{X_G}, r) - \mathbb{E}_0(\rho, P_{OD}, r) \leq 0$. Expanding this difference:

$$E_0(\rho, P_{X_G}, r) - \mathbb{E}_0(\rho, P_{OD}, r) = rTP(1+\rho)\left(1 - \frac{1}{n_t}\right) + \frac{1}{2}(n_t T + \rho n_t T - \rho T - T)\log\left(1 - 2r\frac{P}{n_t}\right) \quad (4.86)$$

We can simplify this to a non-negative coefficient and a polynomial in n_t:

$$= -\frac{T(1+\rho)}{n_t \log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)}\left(n_t^2 - n_t\left(1 + \frac{rP}{\log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)}\right) + \frac{rP}{\log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)}\right) \quad (4.87)$$

Factoring the polynomial gives

$$= -\frac{T(1+\rho)}{n_t \log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)}\left((n_t - 1)\left(n_t - \frac{rP}{\log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)}\right)\right) \quad (4.88)$$

The factor $n_t - 1$ is always non-negative (i.e. there is always at least one transmit antenna), so to show $E_0(\rho, P_{X_G}, r) - \mathbb{E}_0(\rho, P_{OD}, r) \leq 0$, we need to show the second factor is also non-negative for all $0 \leq r \leq \frac{1}{2\frac{P}{n_t}}$:

$$n_t \geq \frac{rP}{\log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)} \quad (4.89)$$

To show this, we use the inequality

$$\log\left(\frac{1}{1-x}\right) \geq x \quad (4.90)$$

Applying this inequality to (4.89) gives

$$\frac{rP}{\log\left(\frac{1}{\sqrt{1-2r\frac{P}{n_t}}}\right)} = \frac{2rP}{\log\left(\frac{1}{1-2r\frac{P}{n_t}}\right)} \leq \frac{2rP}{2r\frac{P}{n_t}} = n_t \quad (4.91)$$

The inequality used is strict unless $r = 0$. These results are summarized in the following proposition.

Proposition 12. *For all $\rho \in [0,1]$, $P \geq 0$, $r \in [0, \frac{1}{2\frac{P}{n_t}}]$, $E_0(\rho, P_{X_G}, r) \leq E_0(\rho, P_{OD}, r)$ with equality iff $r = 0$.*

Summarizing, we have shown that the random coding exponent for an orthogonal design input distribution is larger than the i.i.d. Gaussian input, hence it gives a tighter achievability bound using the method of error exponents. Furthermore, the orthogonal design random coding exponent is strictly larger when $r \neq 0$. Hence this classical method shows that orthogonal designs should have a better finite blocklength preformance. This result is analyzed from the perspective of dispersion in much more detail in the next chapter.

Chapter 5

Orthogonal Designs Minimize Achievable Dispersion

In Chapter 4, we saw that for any capacity achieving input distribution (CAID) X, the conditional variance

$$V(P) = \frac{1}{T}\mathrm{Var}[i(X;Y,H)|X] \tag{5.1}$$

was achievable in the sense that there exists an (n, M, ϵ) code for the MISO block fading channel satisfying

$$\log M(nT, \epsilon) \geq nTC(P) - \sqrt{nTV(P)}Q^{-1}(\epsilon) + o(\sqrt{n}) \tag{5.2}$$

Where $C(P)$ is the capacity of the channel. In this chapter, we answer the question: over all CAIDs X (recall from Chapter 3 that X is not unique), which one minimizes the conditional variance for the $n_t \times T$ MISO channel? This minimizer will give us the tightest bound in (5.2), and will also give insight to the best coding schemes for the MISO channel.

5.1 Minimizing the Dispersion

We start from the expression for the conditional variance. From Proposition 3 in Chapter 3, the conditional variance has the form

$$\frac{1}{T}\mathrm{Var}(i(X;Y,H)|X) = V_1(P) - \frac{\chi_2}{n_t^2 T}\mathrm{Var}(\|X\|_F^2) \tag{5.3}$$

Where V_1 is independent of the CAID X and χ_2 is a non-negative constant. In this form, the dependence of the dispersion on the input distribution is explicit: in order to minimize the dispersion, we maximize $\mathrm{Var}(\|X\|_F^2)$ over the set of CAIDs. We can expand $\mathrm{Var}(\|X\|_F^2)$ as a sum of covariances which will be easier to deal with. This is captured in the following definition, then use to define V_min as the minimal disperison over the set of CAIDs.

Definition 5. For the MISO channel with n_t transmit antennas and coherence time T we define

$$v^*(n_t, T) \triangleq \max_{P_X : I(X;Y,H)=C} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \sum_{k=1}^{T} \sum_{l=1}^{T} \rho_{ikjl}^2 \tag{5.4}$$

where

$$\rho_{ikjl}^2 = \frac{\mathrm{Cov}(X_{ik}^2, X_{jl}^2)}{\mathrm{Var}(X_{11}^2)} \tag{5.5}$$

The notation ρ_{ikjl} is appropriate since whenever X is jointly Gaussian, ρ_{ikjl}^2 is the squared correlation coefficient between X_{ik} and X_{jl}. However, there are non-jointly Gaussian CAIDs where this isn't the case. For instance, when $n_t = T = 2$, for $v \sim \mathrm{Ber}(1/2)$ and A, B i.i.d. $\mathcal{N}(0, P/2)$, the following achieves capacity

$$X = \begin{bmatrix} A & -(-1)^v B \\ B & (-1)^v A \end{bmatrix} \tag{5.6}$$

Here, the correlation coefficient between X_{11} and X_{22} is 0, however (5.5) gives $\rho_{1122}^2 = 1$. Now V_{min}, the minimal dispersion, is given in terms of $v^*(n_t, T)$.

Proposition 13. *The minimal dispersion of an $n_t \times T$ block-fading MISO channel is given by*

$$V_{min} \triangleq \inf_{X\text{-}caid} \frac{1}{T} \mathrm{Var}[i(X;Y,H)|X] = V_1(P) - \frac{2\chi_2 P^2}{n_t^4 T} v^*(n_t, T) \tag{5.7}$$

where V_1 and χ_2 are from (3.51)*.*

Proof. The only term that depends on X in (3.51) is $\mathrm{Var}(\|X\|_F^2)$. We can expand this as a sum of covariance terms:

$$\mathrm{Var}(\|X\|_F^2) = \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \sum_{k=1}^{T} \sum_{l=1}^{T} \mathrm{Cov}(X_{ik}^2, X_{jl}^2) \tag{5.8}$$

Writing this in terms of the ρ_{ikjl}^2, and using $Var(X_{ik}^2) = 2(P/n_t)^2$ from (5.5) yields

$$\mathrm{Var}(\|X\|_F^2) = 2\left(\frac{P}{n_t}\right)^2 \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \sum_{k=1}^{T} \sum_{l=1}^{T} \rho_{ikjl}^2 \tag{5.9}$$

Maximizing this term over the set of CAIDs gives $2(P/n_t)^2 v^*(n_t, T)$, and putting this into the expression for the conditional variance from Proposition 3 gives V_{min} above. □

Therefore intuitively, minimizing dispersion is equivalent to *maximizing* the amount of correlation amongst the entries of X when X is jointly Gaussian. In a sense, this

asks for the capacity achieving input distribution having the least amount of randomness. Next we must characterize $v^*(n_t, T)$. The manifold of CAIDs is not a particularly nice manifold to optimize over, one must account for all the independence constraints on the rows and columns, the covariance constraints on the 2×2 minors, positive definite and symmetric constraints, etc. Our strategy instead will be to given an upper bound on $v^*(n_t, T)$, then show that for most of the pairs (n_t, T), the upper bound is tight. The crux of the upper bound is the following simple lemma.

Lemma 14. *Let $(A_1, \ldots, A_n)$ and $(B_1, \ldots, B_n)$ be i.i.d. random vectors that may have arbitrary correlation between them, then*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(A_i, B_j) \leq n\sigma^2 \tag{5.10}$$

With equality iff $\sum_{i=1}^n A_i = \sum_{i=1}^n B_i$ almost surely.

Proof. Simply use the fact that covariance is a bilinear function and apply the Cauchy-Schwarz inequality:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(A_i, B_j) = \mathrm{Cov}(\sum_{i=1}^{n} A_i, \sum_{j=1}^{n} B_j) \tag{5.11}$$

$$\leq \sqrt{\mathrm{Var}(\sum_{i=1}^{n} A_i)\mathrm{Var}(\sum_{j=1}^{n} B_j)} \tag{5.12}$$

$$= \sqrt{(n\mathrm{Var}(A_1))(n\mathrm{Var}(B_1))} \tag{5.13}$$

$$= n\sigma^2 \tag{5.14}$$

We have equality in Cauchy-Schwarz when $\sum_{i=1}^n A_i$ and $\sum_{i=1}^n B_i$ are propositional, and since these sums have the same distribution, the constant of proportionality must be 1, so we have equality iff $\sum_{i=1}^n A_i = \sum_{i=1}^n B_i$. □

We will use this lemma shortly to upper bound $\mathrm{Var}(\|X\|_F^2)$. But before stating the main theorem of the section, we review orthogonal designs.

5.1.1 Orthogonal Designs

Historical Introduction

In this section we will give some relevant background on orthogonal designs, since they will play a large role in this work. For some historical background, Hurwitz was interested the existence of a "composition formula" for positive integers (r, s, n) [14]

$$(x_1^2 + \cdots + x_r^2)(y_1^2 + \cdots + y_s^2) = (z_1^2 + \cdots + z_n^2) \tag{5.15}$$

Where each z_i is a bilinear form in the x_i's and y_i's. For example, a $(2,2,2)$ composition formula is

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) = (x_1y_1 - x_2y_2)^2 + (x_1y_2 + x_2y_1)^2 \tag{5.16}$$

Let x, y, z denote the r, n, s length vector of the x_i, y_i, z_i's respectively, then the condition that z_i is a bilinear form of the x_i's and y_i's means that z can be written as $z = Ay$ where the entries of A $(n \times s)$ are some linear combinations of x_i's. In which case (5.15) is restated as

$$z^T z = y^T A^T A y = x^T x y^T y \tag{5.17}$$

Which must hold for all indeterminants in y, so the existence condition reduces to the existence of an $n \times s$ matrix A with entries that are linear combinations of the x_i's such that

$$A^T A = \sum_{i=1}^{r} x_i^2 I_s \tag{5.18}$$

Thus yielding (5.15) as desired

$$\sum_{i=1}^{n} z_i^2 = \left(\sum_{i=1}^{r} x_i^2\right)\left(\sum_{i=1}^{s} y_i^2\right) \tag{5.19}$$

Hurwitz-Radon Families

A real $n \times n$ orthogonal design of size k is defined to be an $n \times n$ matrix A with entries given by linear forms in $x_1, \ldots, x_k$ and coefficients in $\mathbb{R}$ satisfying

$$A^T A = \left(\sum_{i=1}^{k} x_i^2\right) I_n \tag{5.20}$$

In other words, all columns of A have squared Euclidean norm $\sum_{i=1}^{k} x_i^2$, and all columns are pairwise orthogonal. Orthogonal designs may be represented as the sum $A = \sum_{i=1}^{k} x_i V_i$ where $\{V_1, \ldots, V_k\}$ is a collection of $n \times n$ real matrices satisfying Hurwitz-Radon conditions:

$$\begin{aligned} V_i^T V_i &= I_n \\ V_i^T V_j + V_j^T V_i &= 0 \quad i \neq j \end{aligned} \tag{5.21}$$

The main theorem on Hurwitz-Radon families gives the largest k such that a family satisfying the above conditions exists, as stated in the following theorem from [6, 12].

Theorem 15 (Radon-Hurwitz)**.** *There exists a family of $n \times n$ real matrices $\{V_1, \ldots, V_k\}$*

satisfying (5.21) iff $k \leq \rho(n)$, where

$$\rho(2^a b) = 8\left\lfloor \frac{a}{4} \right\rfloor + 2^{a \bmod 4}, \qquad a, b \in \mathbb{Z},\, b\text{–odd}. \tag{5.22}$$

In particular, $\rho(n) \leq n$ and $\rho(n) = n$ only for $n = 1, 2, 4, 8$.

So the maximal size of a $n \times n$ orthogonal design is the *Hurwitz-Radon number* $\rho(n)$. As in [16], we can generalize the definition of orthogonal designs to be a rectangular $n \times k$ matrix A (we will assume $n \geq k$) with indeterminants that are linear forms in $x_1, \ldots, x_k$, and A satisfies (5.20). These non-square orthogonal designs are again constructed by a Hurwitz-Radon family $\{V_1, \ldots, V_k\}$ with $V_i \in \mathbb{R}^{n \times n}$ of size k (5.21) via

$$A = [V_1 x \ \cdots \ V_k x] \tag{5.23}$$

Where $x = [x_1, \ldots, x_k]$ is the vector of indeterminates. It follows immediately from this construction that (5.20) is satisfied.

We use this classical theorem to prove the main theorem of this section, which upper bounds the conditional variance and gives conditions for when this bound is tight. The proof and following discussion show that when the bound is tight, *full rate orthogonal designs* achieve the bound, and i.i.d Gaussian inputs preform strictly worse.

5.1.2 Main Theorem on Minimizing Achievable Dispersion

The main theorem of the chapter states that, for dimensions where orthogonal designs exist, the dispersion is minimized if and only if the input is an orthogonal design distribution.

Theorem 16. *For any pair of positive integers n_t, T we have*

$$v^*(T, n_t) = v^*(n_t, T) \leq n_t T \min(n_t, T)\,. \tag{5.24}$$

Furthermore, the bound (5.24) is tight if and only if $n_t \leq \rho(T)$ or $T \leq \rho(n_t)$.

Proof. $v^*(n_t, T) = v^*(T, n_t)$ follows from the symmetry to transposition of CAID-conditions on X (see Proposition 1) and symmetry to transposition of (5.4). From now on, without loss of generality we assume $n_t \leq T$.

For the upper bound, since the rows and columns of X are i.i.d, we can apply Lemma 14 to the rows (or columns) of X, giving

$$\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{T}\sum_{l=1}^{T} \rho^2_{ijkl} = \frac{1}{\mathrm{Var}(X^2_{11})} \sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{T}\sum_{l=1}^{T} \mathrm{Cov}(X^2_{ik}, X^2_{jl}) \tag{5.25}$$

$$\leq \sum_{i=1}^{n_t}\sum_{j=1}^{n_t} T \tag{5.26}$$

$$= n_t^2 T \tag{5.27}$$

We can apply Lemma 14 instead to the columns of X to obtain the bound $T^2 n_t$, hence the smallest upper bound is $n_t T \min(n_t, T)$. Note that (5.25) implies that if X achieves the bound (5.24), then removing the last row of X achieves (5.24) as an $(n_t - 1) \times T$ design. In other words, if (5.24) is tight for $n_t \times T$ then it is tight for all $n_t' \leq n_t$.

With this observation in mind, take $n_t = \rho(T)$ and a maximal Hurwitz-Radon family $\{V_i, i = 1, \ldots, n_t\}$ of $T \times T$ matrices. Let $\xi \sim \mathcal{N}(0, I_T)$ be i.i.d. normal row-vector. Consider

$$X = \left[V_1^T \xi^T \ \cdots V_{n_t}^T \xi^T\right]^T \tag{5.28}$$

The definition of orthogonal design (5.21) implies that rows of X satisfy conditions (3.3)-(3.4). Thus X is capacity achieving. On the other hand, in representation (5.28) the matrix $V_j^T V_i$ contains the correlation coefficients between rows i and j of X, since $\mathbb{E}[(\xi V_j)^T (\xi V_i)] = V_j^T V_i$, so

$$\|V_j^T V_i\|_F^2 = \sum_{k=1}^{T} \sum_{l=1}^{T} \rho_{ikjl}^2 \tag{5.29}$$

Therefore we can represent the sum of squared correlation coefficients as

$$\sum_{i,j,k,l} \rho_{ijkl}^2 = \operatorname{tr}\left(\left(\sum_{i=1}^{n_t} V_i V_i^T\right)^2\right) \tag{5.30}$$

Since V_i are orthogonal, $V_i V_i^T = I_T$ and hence the trace above equals $n_t^2 T$, matching (5.24).

Conversely, suppose X is capacity achieving and attains (5.24). Let $(X_1, \ldots, X_T)$ be the first row of X, we will show that

$$XX^T = \left(\sum_{i=1}^{T} X_i^2\right) I_{n_t} \ \ a.s. \tag{5.31}$$

By definition, this relation means that X is an orthogonal design, and hence $n_t \leq \rho(T)$ or $T \leq \rho(n_t)$ by Theorem 15. From Lemma 14, the equality condition from the Cauchy-Schwarz inequality tells us that the bound (5.24) is tight iff for all $j \in \{1, \ldots, n_t\}$,

$$\|R_j\|^2 = \sum_{i=1}^{T} X_i^2 \ \ a.s. \tag{5.32}$$

Where R_j is the j-th row of X, and $\|\cdot\|$ is the Euclidean norm. So the Euclidean norms of all rows must be almost surely equal. Hence the diagonal entries of XX^T, which are the random variables $R_i R_i^T$, are all $\sum_{i=1}^{T} X_i^2$. In order for (5.31) to hold, we must show that the off diagonal entries $R_i R_j^T = 0$ a.s. for $i \neq j$.

If a CAID X achieves the bound, then any left-rotation of X also achieves the

bound, i.e. UX achieves the bound for any $n_t \times n_t$ orthogonal matrix U. To see this, notice that left rotations do not change the channel statistics, since H is rotationally invariant by assumption:

$$Y = H(UX) + Z = (HU)X + Z \sim HX + Z \tag{5.33}$$

Let $u_1, \ldots, u_{n_t}$ be the rows of any $n_t \times n_t$ orthogonal matrix U. Take the orthogonal matrix that mixes row i and j, and leaves the other rows untouched, i.e.

$$u_i = [0, \ldots, \frac{1}{\sqrt{2}}, \ldots, \quad \frac{1}{\sqrt{2}}, \ldots 0] \tag{5.34}$$

$$u_j = [0, \ldots, \frac{1}{\sqrt{2}}, \ldots, -\frac{1}{\sqrt{2}}, \ldots 0] \tag{5.35}$$

And all other rows u_k are the standard basis vectors with a 1 in the k-th position and 0's elsewhere. The matrix UX has row i as $u_i X$ and row j as $u_j X$, which have Euclidean norm:

$$\|u_i X\|^2 = \sum_{k=1}^{n_t} \sum_{l=1}^{n_t} u_{ik} u_{il} \langle R_k, R_l \rangle = \sum_{i=1}^{T} X_i^2 + \langle R_i, R_j \rangle \tag{5.36}$$

$$\|u_j X\|^2 = \sum_{k=1}^{n_t} \sum_{l=1}^{n_t} u_{jk} u_{jl} \langle R_k, R_l \rangle = \sum_{i=1}^{T} X_i^2 - \langle R_i, R_j \rangle \tag{5.37}$$

Where $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product. Since UX achieves the bound, we must have the almost sure equality condition

$$\|u_1 X\|^2 = \|u_2 X\|^2 \tag{5.38}$$

And hence we see that almost surely

$$\langle R_i, R_j \rangle = -\langle R_i, R_j \rangle \tag{5.39}$$

And therefore $\langle R_i, R_j \rangle = 0$ almost surely. Summarizing, if X is a CAID and achieves the bound (5.24), almost surely we have

$$R_i R_i^T = \sum_{i=1}^{T} X_i^2 \tag{5.40}$$

$$R_i R_j^T = 0, \quad i \neq j \tag{5.41}$$

And thus almost surely (5.31) holds. □

Remark 3. This proof shows that, even if a CAID X is a non-jointly Gaussian, for

example

$$X = \begin{bmatrix} A & -(-1)^v B \\ B & (-1)^v A \end{bmatrix} \tag{5.42}$$

We still know that if X achieves the bound (5.24), then it is an orthogonal design, i.e. that all rows have the same norm almost surely, and all pairs of rows are orthogonal almost surely. In a way, this show that orthogonal designs are very naturally tied to the MISO block fading channel.

Remark 4. Note that the input distribution where all entries are i.i.d. Gaussian certainly does not satisfy

$$XX^T = \sum_{i=1}^{T} X_i^2 I_{n_t} \quad a.s. \tag{5.43}$$

Where $(X_1, \dots, X_T)$ is the first row of X. Hence this input is strictly worse than an orthogonal design input.

Remark 5. The condition for tightness in Theorem 16 is that $n_t \leq \rho(T)$ (assuming $n_t \leq T$), one may ask, how often is this satisfied? From the definition of the ρ function in (5.22), we see that $\rho(n)$ is increasing, which means that for any number of transmit antennas n_t, eventually the bound will be tight. Often the coherence time is much larger than the number of antennas, so this in this regime, the bound will very likely be tight.

Remark 6. Elementary results on orthogonal designs show that the conditions for tightness of (5.24) are satisfied if and only if a full rate real orthogonal design of dimensions $n_t \times T$ or $T \times n_t$ exists, cf. [16] or [8, Proposition 4]. Consequently, each full-rate orthogonal design yields a CAID X that achieves minimal dispersion. Some examples (ξ_j are i.i.d. $\mathcal{N}(0,1)$) for the $n_t = T = 4$ and the $n_t = 4, T = 3$ case are

$$X = \sqrt{\frac{P}{4}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 & \xi_4 \\ -\xi_2 & \xi_1 & -\xi_4 & \xi_3 \\ -\xi_3 & \xi_4 & \xi_1 & -\xi_2 \\ -\xi_4 & -\xi_3 & \xi_2 & \xi_1 \end{bmatrix} \tag{5.44}$$

$$X = \sqrt{\frac{P}{4}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ -\xi_2 & \xi_1 & -\xi_4 \\ -\xi_3 & \xi_4 & \xi_1 \\ -\xi_4 & -\xi_3 & \xi_2 \end{bmatrix}$$

5.1.3 Dimensions Where Orthogonal Designs Do Not Exist

For pairs (n_t, T) where $n_t > \rho(T)$, an orthogonal design does not exist, yet we can still ask: which capacity achieving input distribution still minimizes the dispersion? The minimizer won't be an orthogonal design, but won't be an i.i.d. Gaussian CAID

either (as shown in the next section). Hence we'll get some new object. The question of exactly which distributions minimize the conditional variance in these cases is still open. The following gives a construction of input distributions that achieve a smaller dispersion than i.i.d. Gaussian in these cases.

Note that for the designs with entries $\pm\xi_j$, with ξ_j being independent Gaussians, computation of the sum (5.4) is simplified:

$$\sum_{ijkl} \rho_{ikjl}^2 = \sum_{t=1}^{d} (\ell_t)^2 \,, \tag{5.45}$$

where ℓ_t is the number of times $\pm\xi_t$ appears in the description of X. By this observation and the remark after Proposition 1 we can obtain lower bounds on $v^*(n_t, T)$ for $n_t > \rho(T)$ via the following *truncation construction*:

1. Take $T' > T$ such that $\rho(T') \geq n_t$ and let X' be a corresponding $\rho(T') \times T'$ full-rate orthogonal design (with entries $\pm\xi_1, \ldots \pm\xi_{T'}$).

2. Choose an $n_t \times T$ submatrix of X' maximizing the sum of squares of the number of occurrences of each of ξ_j, cf. (5.45).

As an example of this method, by truncating a 4×4 design (5.44) we obtain the following 2×3 and 3×3 submatrices:

$$X = \sqrt{\frac{P}{3}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ -\xi_2 & \xi_1 & \xi_4 \\ -\xi_3 & -\xi_4 & \xi_1 \end{bmatrix} \quad X = \sqrt{\frac{P}{2}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ -\xi_2 & \xi_1 & \xi_4 \end{bmatrix} \tag{5.46}$$

By independent methods we were able to show that designs (5.46) are dispersion-optimal out of all jointly Gaussian CAIDs, and attain $v^*(3,3) = 21$ and $v^*(2,3) = 10$. Note that in these cases the bound (5.24) is not tight, illustrating the "only if" part of Theorem 16.

Orthogonal designs were introduced into communication theory by Tarokh et al [16] as a natural generalization of Alamouti's scheme [1]. In cases when full-rate designs do not exist, there have been various suggestions as to what could be the best solution, e.g. [8]. Thus for non full-rate designs the property of minimizing dispersion (such as (5.46)) could be used for selecting the best design for cases $n_t > \rho(T)$.

5.1.4 Numerical Values for Minimal Dispersion

Our current knowledge about v^* is summarized in Table 5.1. The lower bounds for cases not handled by Theorem 16 were computed by truncating the 8x8 orthogonal design [16, (5)]. Based on the evidence from $2 \times T$ and 3×3 we *conjecture* this construction to be optimal.

Finally, returning to the original question of the minimal delay required to achieve capacity, see (1.2), we calculate the value of $\frac{V_{min}}{C^2}$ in Table 5.2.

Table 5.1: Values for $v^*(n_t, T)$

$n_t \setminus T$	1	2	3	4	5	6	7	8
1	1	2	3	4	5	6	7	8
2		8	10	16	18	24	26	32
3			21	36	[39,45)	[46,54)	[57,63)	72
4				64	[68,80)	[80,96)	[100,112)	128
5					[89,125)	[118,150)	[155,175)	200
6						[168,216)	[222,252)	288
7							[301,343)	392
8								512

Note: Table is symmetric about diagonal; intervals $[a, b)$ mark entries for which dispersion-optimal input is unknown.

Table 5.2: Values of $\frac{V_{min}}{C^2}$ (when known) at $SNR = 20\ dB$

$n_t \setminus T$	1	2	3	4	5	6	7	8
1	0.38	0.60	0.82	1.05	1.27	1.49	1.72	1.94
2	0.35	0.39	0.52	0.60	0.73	0.82	0.94	1.03
3	0.38	0.42	0.45	0.49				0.79
4	0.42	0.43	0.44	0.45				0.69
5	0.46	0.48						0.64
6	0.50	0.51						0.62
7	0.54	0.55						0.61
8	0.59	0.59	0.59	0.60	0.60	0.60	0.61	0.61

From the proof of Theorem 16 it is clear that Telatar's i.i.d. Gaussian (as in (3.1)) is never dispersion optimal, unless $n_t = 1$ or $T = 1$. Indeed, for Telatar's input $\rho_{ikjl} = 0$ unless $(i, k) = (j, l)$. Thus embedding even a single Alamouti block (3.2) into an otherwise i.i.d. $n_t \times T$ matrix X strictly improves the sum (5.4).

We note that the value of $\frac{V}{C^2}$ entering (1.2) can be quite sensitive to the suboptimal choice of the design. For example, for $n_t = T = 8$ and $SNR = 20\ dB$ estimate (1.2) shows that one needs

- around 600 channel inputs (that is 600/8 blocks) for the optimal 8×8 orthogonal design, or
- around 850 channel inputs for Telatar's i.i.d. Gaussian design

in order to achieve 90% of capacity. This translates into a 40% longer delay (or battery spent in running the decoder) with unoptimized transmitter.

Thus, curiously even in cases where pure multiplexing (that is maximizing transmission rate) is needed – as is often the case in modern cellular networks – transmit diversity enters the picture by enhancing the finite blocklength fundamental limits. We remind, however, that our discussion pertains only to cases when the transmitter (base-station) is equipped with more antennas than the receiver (user equipment).

5.2 Linear Decoders for Orthogonal Designs

One major advantage about using an orthogonal design in a MISO system is that the decoder can decouple all data stream by using a linear transformation. The following theorem shows that this is possible if and only if

Theorem 17. *Suppose X is an $n_t \times T$ full rate orthogonal design with entries $d_1, \ldots, d_T$ for the channel $Y = hX + Z$, then there exists a $T \times T$ matrix B such that*

$$\frac{1}{\|h\|} YB = \|h\| d + \tilde{Z}_i \tag{5.47}$$

Where $d = [d_1, \ldots, d_T]$ and $\tilde{Z}_i \sim \mathcal{N}(0, I_T)$, and B is a $T \times T$ orthogonal design with entries that are linear combinations of $\{0, h_1, \ldots, h_{n_t}\}$. Furthermore, if X is any input that contains more than T independent symbols, then no such B exists.

Note that the i-th entry of the transformed received vector $\frac{1}{\|h\|} YB$ only depends on d_i, hence B decouples the data streams.

Proof. Represent hX as $hX = dA$, where A is a $T \times T$ matrix with entries as linear forms of h and $d = [d_1, \ldots, d_T]$. This can be done for any matrix X given that X has entries that are linear forms of d, simply by solving for the a_{ij}'s entry-wise. Now suppose X is a full rate orthogonal design, i.e. satisfies $XX^T = \sum_{i=1}^{T} d_i^2 I_{n_t}$. Then

$$dAA^T d^T = hXX^T h^T \tag{5.48}$$

$$= \left(\sum_{i=1}^{T} d_i^2 \right) \left(\sum_{i=1}^{n_t} h_i^2 \right) \tag{5.49}$$

$$= d \left(\sum_{i=1}^{n_t} h_i^2 I_T \right) d^T \tag{5.50}$$

Since the left and right hand side are equal as quadratic forms, $AA^T = \sum_{i=1}^{n_t} h_i^2$, hence A is an orthogonal design, and

$$YB = hXB + ZB = dAA^T + ZB = \|h\|^2 d + ZB \tag{5.51}$$

Finally, note that $\frac{1}{\|h\|} B$ is an orthogonal matrix, and since and the distribution $\mathcal{N}(0, I_T)$ is invariant under orthogonal transformations, $\frac{1}{\|h\|} ZB \sim \mathcal{N}(0, I_T)$.

Conversely, suppose X contains l data symbols $\{d_1, \ldots, d_l\}$ where $l > T$. Again we can find A such that $hX = dA$ (again, simply by solving entry-wise equations), where A has dimension $T \times l$. If a linear decoding matrix B of dimension $l \times T$ were to exist, it would satisfy

$$dAB = d\|h\|^2 \implies AB = \|h\|^2 I_T \tag{5.52}$$

But since $l > T$, the determinant of the left hand side is zero while the right hand size is non zero, hence such a B cannot exist. □

Example: Take the $n_t = 3, T = 4$ case with CAID

$$X = \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 & \xi_4 \\ -\xi_2 & \xi_1 & -\xi_4 & \xi_3 \\ -\xi_3 & \xi_4 & \xi_1 & -\xi_2 \end{bmatrix} \tag{5.53}$$

Which is a full rate orthogonal design. Solving $hX = dA$ for A where $h = [h_1, h_2, h_3]$ and $d = [\xi_1, \xi_2, \xi_3, \xi_4]$ yields

$$A = \begin{bmatrix} h_1 & h_2 & h_3 & 0 \\ -h_2 & h_1 & 0 & -h_3 \\ -h_3 & 0 & h_1 & h_2 \\ 0 & h_3 & -h_2 & h_1 \end{bmatrix} \tag{5.54}$$

This is a 4×4 orthogonal design of size 3 in h_1, h_2, h_3. Upon receiving $Y = HX + Z = dA + Z$, the decoder computes

$$R \triangleq Y \frac{A^T}{\|h\|} = \frac{dAA^4}{\|h\|} + \frac{ZA^T}{\|h\|} = d\|h\| + \tilde{Z} \tag{5.55}$$

Since A is an orthogonal design, $AA^T = \left(\sum_{i=1}^{3} h_i^2\right) I_4 = \|h\|^2 I_4$, and since $\frac{1}{\|h\|} A^T$ is an orthogonal matrix, the additive noise $\tilde{Z}$ is still white. Hence the independent data streams are completely decoupled at the receiver, since $R_i = \|h\|\xi_i + \tilde{Z}_i$.

Appendix A

Proof

Appendix B

Figures

Bibliography

[1] Siavash M Alamouti. A simple transmit diversity technique for wireless communications. *IEEE J. Select. Areas Commun.*, 16(8):1451–1458, Oct. 1998.

[2] Ezio Biglieri, John Proakis, and Shlomo Shamai. Fading channels: Information-theoretic and communications aspects. *Information Theory, IEEE Transactions on*, 44(6):2619–2692, 1998.

[3] Harald Cramér. Über eine eigenschaft der normalen verteilungsfunktion. *Mathematische Zeitschrift*, 41(1):405–414, 1936.

[4] R. L. Dobrushin. Mathematical problems in the Shannon theory of optimal coding of information. In *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, volume 1, pages 211–252, Berkeley, CA, USA, 1961.

[5] Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

[6] Adolf Hurwitz. Über die komposition der quadratischen formen. *Math. Ann.*, 88(1):1–25, 1922.

[7] Erik G Larsson, Fredrik Tufvesson, Ove Edfors, and Thomas L Marzetta. Massive mimo for next generation wireless systems. *arXiv preprint arXiv:1304.6690*, 2013.

[8] Xue-Bin Liang. Orthogonal designs with maximal rates. *IEEE Trans. Inform. Theory*, 49(10):2468–2503, Oct. 2003.

[9] Y. Polyanskiy, H.V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory*, 56(5):2307–2359, 2010.

[10] Y. Polyanskiy and S. Verdu. Finite blocklength methods in information theory (tutorial). In *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, July 2013.

[11] Yury Polyanskiy. *Channel coding: non-asymptotic fundamental limits*. Princeton University, 2010.

[12] J Radon. Lineare scharen orthogonaler matrizen. In *Abh. Sem. Hamburg*, volume 1, pages 1–14. Springer, 1922.

[13] Hariharan Rahul, Swarun Kumar, and Dina Katabi. Megamimo: scaling wireless capacity with user demands. 2012.

[14] Daniel B Shapiro. *Compositions of quadratic forms*, volume 33. Walter de Gruyter, 2000.

[15] V. Strassen. Asymptotische Abschätzungen in Shannon's Informationstheorie. In *Trans. 3d Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962.

[16] Vahid Tarokh, Hamid Jafarkhani, and A Robert Calderbank. Space-time block codes from orthogonal designs. *IEEE Trans. Inform. Theory*, 45(5):1456–1467, July 1999.

[17] Emre Telatar. Capacity of multi-antenna Gaussian channels. *Eur. trans. telecom.*, 10(6):585–595, 1999.