Approximation of Capacity for ISI channels with 3-level Output Quantization

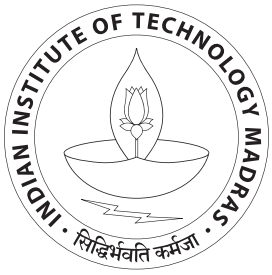
A Project Report

submitted by

SUHAS S KOWSHIK

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY
AND
BACHELOR OF TECHNOLOGY



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

June 2016

THESIS CERTIFICATE

This is to certify that the thesis entitled **Approximation of Capacity for ISI channels with 3-level Output Quantization**, submitted by **Suhas S Kowshik**, to the Indian Institute of Technology Madras, for the award of the degree of **Master of Technology and Bachelor of Technology**, is a bona fide record of the research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Andrew Thangaraj
Research Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date:

ACKNOWLEDGEMENTS

I would like to thank Prof. Andrew Thangaraj for guiding me through this project and helping me realize the importance of intuitive understanding of the problem at hand. I would also like to express my gratitude towards Prof. Radhakrishna Ganti for giving useful insights that helped me in this work. Special thanks to Mr. Arijit Mondal for clarifying any queries that I had on the problem. It would be unfair if I do not mention my friends Dheeraj and Sundara Rajan with whom I have had wonderful discussions which have immensely benefited me.

Last but never the least, I would like to thank my parents, Subramanya and Pushpa, for their continuous support throughout my life; my sister, Sahana, for bugging me with all sorts of questions in pre-university Math and Physics over the past year that has helped me in strengthening by basics; and the Almighty.

ABSTRACT

KEYWORDS: Capacity, ISI Channel, Quantization

In this work we consider approximating capacity of Inter-Symbol Interference (ISI) channels with 3-level quantization at the output under an average-power constrained input. This work is motivated by the recent results on ISI channels with a 1-bit (2-level) quantization at the output under a similarly constrained input [Ganti *et al.*, 2015]. Since the exact capacity of such systems are difficult to characterize, we consider an approximation in which the output concurs with the exact channel output up to a probability of error. In this approximation, there is no additive noise but the absolute value of the ISI channel output is constrained to be away from the decision points of the quantizer by a certain threshold . The capacity under this approximation is calculated using convex optimization and involves standard Gibbs distribution. We use Markovian schemes under a zero forcing input to show that it approaches the approximate capacity. We show that practical coding schemes can be developed for ISI channels with 3-level output quantization using the methods developed for the approximate ISI channel and discuss some signaling methods for actual channel particularly with 2-level quantization. We also give a possible achievable rate for the exact channel and discuss about using source simulation to demonstrate that the rate is achievable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

IITM	Indian Institute of Technology Madras
ISI	Inter-Symbol Interference
AWGN	Additive White Gaussian Noise
DTFT	Discrete Time Fourier Transform
SNR	Signal to Noise Ratio
BER	Bit Error Rate

NOTATION

$\mathbb{N}$	Set of Natural numbers
$\mathbb{R}$	Field of Real numbers
diag$(\cdot)$	Diagonal matrix with the argument vector as the principal diagonal
tr$(\cdot)$	Trace of the argument matrix
$[N]$	$\{n \in \mathbb{N} : 1 \leq n \leq N\}$
$\mathbb{P}(\cdot)$	Probability mass function
$\mathbb{E}[\cdot]$	Expectation of the argument random variable/vector

CHAPTER 1

INTRODUCTION

We often encounter ISI channels with Additive White Gaussian Noise (AWGN) in practice. Either a finite input alphabet constraint or an average-input power constraint is usually used based on the applications. Recently, in applications like optical or intra-chip [Harwood *et al.*, 2007] or millimeter wave [Sun *et al.*, 2014][Alkhateeb *et al.*, 2014] communications, output quantization of ISI channel has been looked at due to limitations in a/d conversion at high speeds. The transmitters of some of these systems can be fairly complex and can operate at high powers and hence the channel input may not have serious quantization limits.

Motivated due to the above applications, a noisy ISI channel with average-power constrained continuous input and 1-bit (2-level) output quantization was considered in [Ganti *et al.*, 2015]. Apart from this work, the available literature mostly deals with continuous input/output or a finite input alphabet and continuous output alphabet [Shamai and Laroia, 1996][Sadeghi *et al.*, 2009]. The case of quantized output with AWGN and no ISI has been dealt with in [Singh *et al.*, 2009]. The design of quantizers for maximizing information rate for the ISI case has been considered in [Zeitler *et al.*, 2012]. In the context of millimeter wave communications, ISI case has been briefly discussed in [Mo and Heath, 2014].

In this work, we consider the 3-level quantization of a noisy ISI channel with average-power continuous input with methods similar to those in [Ganti *et al.*, 2015]. The exact capacity of noisy ISI channel with 3-level output quantization is

difficult to characterize explicitly. So we consider an approximation to the ISI channel model wit 3-level output quantization similar to the one used in [Ganti *et al.*, 2015]. This approximation does not have additive noise, but the noiseless output of the channel is constrained to have a certain minimum distance (threshold) from the transition regions of the quantizer. Because of this constraint, the quantized outputs of approximate channel and actual channel match upto a probability of error that can be controlled by the threshold.

The problem of computing exact capacity for the approximate channel involves solving a quadratic program. We consider a special but useful case where we impose certain constraint on the quantization points and the threshold. In the case, it is somewhat easier to solve this optimization problem. For the general case we provide an algorithm and give an intuitive reasoning for why it works. We also provide a Markovian achievable scheme under zero forcing input that approaches approximate capacity.

Since the approximate channel output matches the actual channel output upto a probability of error, a coding scheme used over approximate channel can be augmented with a standard error control code to derive a practical coding scheme for the actual channel. Using this, we describe a signaling method for the actual ISI channel with noise but with 2-level output quantization since it is easier to analyze. From this signaling method, we claim that a particular rate is achievable. We aim to demonstrate this using source simulation techniques where the objective is to show that an arbitrarily low bit error rate is achievable. But we haven't been successful in this regard due to non one-one nature of the source simulator. Instead we propose and use a typical set theory based method which works, but, due to computational restrictions, there is severe loss of information rate and we are able

to show that only a fraction of the proposed rate is achievable. Given enough computational resources it might possible to use/improve this method to show that the proposed rate is achievable.

1.1 System model

1.1.1 2-level Quantizer

In [Ganti *et al.*, 2015], a discrete time finite length ISI channel with average-power constrained continuous input and 1-bit (2-level) quantized output is considered as shown in Fig 1.1.

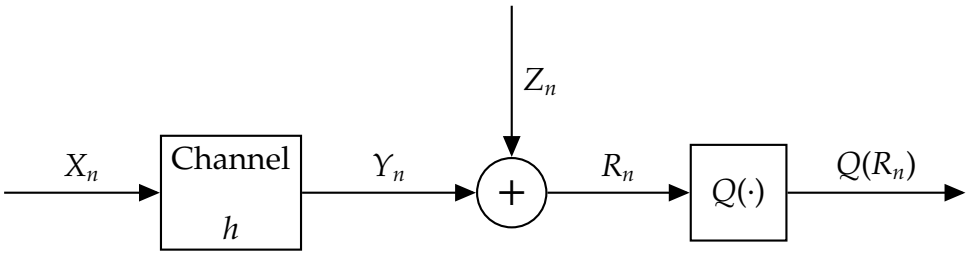


Figure 1.1: ISI channel with 2-level quantized output

The input to the channel is denoted $X = \{X_n \in \mathbb{R}, 0 \leq n \leq N-1\}$. The channel impulse response is of length L and denoted $h = \{h_n \in \mathbb{R}, 0 \leq n \leq L-1\}$. The output of the channel which is the convolution of input with the channel impulse response is denoted $Y = \{Y_n, 0 \leq n \leq N+L-2\}$ which is given by

$$Y_n = \sum_{k=0}^{L-1} h_k X_{n-k} \tag{1.1}$$

The channel h is assumed to be constant. Also, all signals are assumed to be zero outside their specified ranges. Further, it is assumed that $N >> L$ so that the

output vector is of length N. So the convolution in (1.1) can be represented in matrix notation as $Y = \tilde{M}_h X$. The entries of $N \times N$ matrix $\tilde{M}_h$ are either 0 or one of the channel taps h_k and X, Y are column vectors of length N. Since the matrix $\tilde{M}_h$ is not circulant, the $N \times N$ circulant matrix M_h whose first column is equal to h is considered under the observation that for large N with fixed L, $\tilde{M}_h$ behaves like the circulant matrix M_h in the sense that $\lim_{n\to\infty} ||M_h - \tilde{M}_h|| = 0$ where $||\cdot||$ is the matrix norm. In [Ganti *et al.*, 2015] and in this work circular convolution $Y = M_h X$ is assumed for simplicity.

Independent and identically distributed zero-mean Gaussian noise of variance σ^2, denoted Z_n is added to Y_n to get an intermediate signal $R_n = Y_n + Z_n$. This is quantized by a 2-level quantizer $Q(\cdot)$ to obtain the channel output $Q(R) = \{Q(R_n), 0 \le n \le N-1\}$. The quantizer is defined as:

$$Q(x) = \begin{cases} +1, & x \ge 0 \\ -1, & x < 0 \end{cases} \tag{1.2}$$

The average power of the input is constrained to be at most P:

$$\mathbb{E}[||X||^2] = \sum_{n=0}^{N-1} \mathbb{E}[|X_n|^2] \le NP \tag{1.3}$$

.

The goal in [Ganti *et al.*, 2015] was to approximate the mutual information rate $\frac{1}{N} I(X; Q(R))$ and provide computable expressions or bounds.

The approximate ISI channel is shown in figure 1.2. In this model, there is no noise but the convolution output is constrained to be greater than a threshold δ in absolute value.

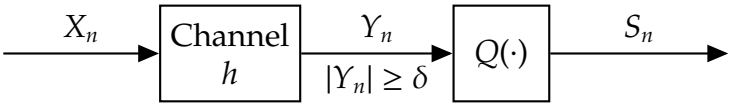


Figure 1.2: Approximate ISI channel with 2-level quantized output

This constraint provides justification for ignoring the noise because under this constraint, the output of the actual model $Q(R_n)$ is approximated by $S_n = Q(Y_n)$ up to a probability of error lesser than or equal to $\boldsymbol{Q}(\delta/\sigma)$, where $\boldsymbol{Q}(x) = \int_x^\infty \frac{1}{2\pi} e^{u^2/2} du$ is the standard Q function. So coding schemes for the approximate model can be used in actual model with an error control coding for the approximation error $\boldsymbol{Q}(\delta/\sigma)$. Even though there is a loss in information rate due to additional error control coding, the approximation is useful since the capacity $\frac{1}{N} I(X;S)$, where $S = \{S_n, 0 \leq n \leq N-1\}$, under the constraints (1.3) and $|Y_n| \geq \delta$ has computable expressions and bounds. Also, the techniques used for computing approximate capacity can be used to get intuition on signaling and coding methods for output-quantized ISI channels [Ganti *et al.*, 2015].

1.1.2 3-level Quantizer

We consider a discrete time finite length ISI channel with average-power constrained continuous input and 3-level quantized output. The exact channel model is similar to the one in figure 1.1 with the output quantizer $Q_d(\cdot)$, given a real number $d \geq 0$, defined as:

$$Q_d(x) = \begin{cases} +1, & x \geq d \\ 0, & -d \leq x < d \\ -1, & x < -d \end{cases} \tag{1.4}$$

The approximate ISI channel is shown in figure 1.3. In the approximate ISI channel, there is no noise. However, the quantities $|Y_n - d|$ and $|Y_n + d|$ are constrained to be greater than or equal to the threshold δ.

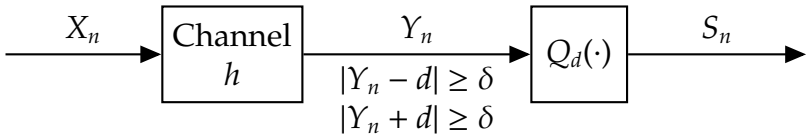


Figure 1.3: Approximate ISI channel with 3-level quantized output

Under the constraints $|Y_n \pm d| \geq \delta$, it is easily seen that the output of the actual model $Q_d(R_n)$ is approximated by $S_n = Q_d(Y_n)$ up to probability of error lesser than or equal to $2Q(\delta/\sigma)$ [1].

Similar to the 2-level case, we consider only circular convolution: $Y = M_h X$ where M_h is a circulant matrix with first column as h. Also, henceforth, the indices of all vectors/sequences of length N range from 1 to N instead of 0 to $N - 1$.

[1] Assuming equally likely quantized outputs, the probability of error is lesser than or equal to $\frac{4}{3}Q(\delta/\sigma)$

CHAPTER 2

2-LEVEL OUTPUT QUANTIZATION

In this chapter, we review some of results on approximate ISI channel with 2-level output quantization from [Ganti *et al.*, 2015].

2.1 Approximate capacity

The capacity of approximate ISI channel in figure 1.2 is given by

$$C_{N,\delta}(P) = \sup_{\substack{\mathbb{E}[\|X\|^2] \leq NP \\ |Y_n| \geq \delta}} \frac{I(X,S)}{N} = \frac{H(S)}{N} \tag{2.1}$$

where X and S are the column vectors of input and quantized output respectively, $I(\cdot,\cdot)$ is the mutual information and $H(\cdot)$ is the entropy. The last equality in (2.1) is because the vector S is a deterministic function of the input vector X in the absence of any noise in the approximate channel. Since $S \in \{+1,-1\}^N$, we have $C_{N,\delta}(P) \leq 1$.

First, the power of the input sequence X required for a given output sequence S is bounded. For a given output sequence $S = s \in \{-1,+1\}^N$, the constraint $|Y_n| \geq \delta$ can be written as $|Y_n| = s_n Y_n \geq \delta$. Hence the minimum energy required to obtain a given output sequence s is given by:

$$\varepsilon(s) = \min_{\text{diag}(s)M_h x \geq \delta \mathbf{1}} \|x\|^2 \tag{2.2}$$

where diag(s) is an $N \times N$ diagonal matrix with s on the principal diagonal and

1 denotes the all-1 column matrix. The constraints of this optimization problem are linear and the feasible set for x is the intersection of hyperplanes, which is convex. So this optimization problem is a convex optimization problem with linear constraints and hence strongly dual. Further,

$$\mathbb{E}[||X||^2] = \sum_{s\in\{-1,1\}^N} \mathbb{P}(S=s)\mathbb{E}[||X||^2|S=s] \geq \sum_{s\in\{-1,1\}^N} \mathbb{P}(S=s)\varepsilon(s) \tag{2.3}$$

Since the input X has an average-power constraint, this implies

$$\sum_{s\in\{-1,1\}^N} \mathbb{P}(S=s)\varepsilon(s) \leq NP \tag{2.4}$$

Letting $\varepsilon_{\min} = \min_s \varepsilon(s)$, $\varepsilon_{\max} = \max_s \varepsilon(s)$ and $\bar{\varepsilon} = \frac{1}{2^N}\sum_s \varepsilon(s)$, under the linear constraints (2.4), it is known [Jaynes, 1957] that the Gibbs distribution maximizes the entropy $H(S)$ for $\varepsilon_{\min} \leq NP \leq \varepsilon_{\max}$. The optimal Gibbs distribution is given by

$$\mathbb{P}(S=s) = \frac{e^{-\beta\varepsilon(s)}}{Z},\ s \in \{-1,1\}^N \tag{2.5}$$

where Z is the normalizing constant and β is the unique value for which (2.4) is met with equality. The maximum entropy is given by $H(S) = \beta NP + \ln(Z)$. The capacity for different ranges of NP is given in table 2.1.

Table 2.1: Capacity of Approximate ISI channel

Range of NP	$C_{N,\delta}(P)$	Remarks
$NP < \varepsilon_{\min}$	0	No state if feasible
$\varepsilon_{\min} \leq NP < \bar{\varepsilon}$	$\frac{1}{\ln 2}(\beta P + \frac{\ln(Z)}{N})$	β is such that (2.4) is met with equality
$NP \geq \bar{\varepsilon}$	1	$\beta = 0$, uniform distribution

It is now required to compute $\varepsilon(s), \forall s \in \{-1,1\}^N$.

2.1.1 Diagonally dominant channels

Definition 2.1.1 (Row-diagonally dominant matrix [Ganti *et al.*, 2015])**.** An $N \times N$ matrix $A = (a_{ij})$ is said to be row-diagonally dominant or simply diagonally dominant if $|a_{ii}| \geq \sum_{i \neq j} |a_{ij}|, \forall i \in [N]$

The channels h for which the matrix $(M_h M_h^T)^{-1}$ exists and is diagonally dominant are called diagonally dominant channels [Ganti *et al.*, 2015]. It is also assumed that $\left((M_h M_h^T)^{-1}\right)_{ii} \geq 0, \forall i \in [N]$. We re-state some of the lemmas from [Ganti *et al.*, 2015]; the reader is referred to it for the proofs.

Lemma 2.1.1 ([Ganti *et al.*, 2015])**.** *When the matrix $(M_h M_h^T)^{-1}$ is row-diagonally dominant, $\varepsilon(s)$ for $s \in \{-1, 1\}^N$ is achieved at x^* that satisfies the equality constraints*

$$\text{diag}(s) M_h x^* = \delta \mathbf{1} \tag{2.6}$$

Hence, $\varepsilon(s) = \|x^*\|^2 = \delta^2 s^T G s$ where $G = (M_h M_h^T)^{-1}$. Also note that the optimal x^* is actually the zero-forced input: $x^* = \delta M_h^{-1} s$ (zero-forced because there is channel inversion).

Lemma 2.1.2 ([Ganti *et al.*, 2015])**.** *The mean energy for diagonally dominant channels is given by*

$$\bar{\varepsilon} = \delta^2 \text{tr}((M_h^T)^{-1} M_h^{-1}) \tag{2.7}$$

Let the discrete time Fourier transform (DTFT) of the channel h be

$$f(\lambda) = \sum_{k=0}^{L-1} h_k e^{jk\lambda}. \tag{2.8}$$

Since M_h is circulant, it can be seen that [Ganti *et al.*, 2015]

$$\bar{\varepsilon} = \delta^2 \text{tr}((M_h^T)^{-1} M_h^{-1}) = \delta^2 \sum_{k=1}^{N} \frac{1}{\left| f(\frac{2\pi k}{N}) \right|^2}. \tag{2.9}$$

Letting $\overline{P}_h = \lim_{N\to\infty} \bar{\varepsilon}/N$, which is the minimum average power required for capacity of 1-bit, and using standard arguments, it is seen in [Ganti *et al.*, 2015] that (see [Gray, 2006])

$$\overline{P}_h \to \frac{\delta^2}{2\pi} \int_0^{2\pi} \frac{1}{|f(\lambda|^2} d\lambda \tag{2.10}$$

To summarize, for diagonally dominant channels, the approximate ISI channel capacity in the case of large N is given by

$$C_{N,\delta}(P) = \begin{cases} 1, & P \geq \overline{P}_h \\ \beta P + \frac{\ln(Z)}{N}, & \underline{P}_h \leq P < \overline{P}_h \\ 0, & P < \underline{P}_h \end{cases} \tag{2.11}$$

where $\underline{P}_h = \lim_{N\to\infty} \varepsilon_{\min}/N$.

2.2 Achievable schemes

In achievable schemes, an information/message sequence $B \in \{-1, 1\}^N$ with a well-chosen probability distribution is encoded into a channel input x that satisfies the constraint $|y_n \geq \delta|$. The rate of transmission over the channel is $H(B)/N$. In [Ganti *et al.*, 2015], two achievable schemes are considered: Zero-forcing with Gibbs distribution and Zero-forcing with Markov input. Both are briefly described below. The reader is referred to [Ganti *et al.*, 2015] for details.

2.2.1 Zero-forcing with Gibbs distribution

Let $b = \{b_i\}_{i=1}^{N} \in \{-1, 1\}^N$ be an instance of the information sequence B. The input x to the channel is chosen as $x = \delta M_h^{-1} b$, which implies that the output of the ISI channel is $y = M_h x = \delta b$. So the output of the quantizer s equals b. For a diagonally dominant channel, as seen from lemma 2.1.1 and section 2.1, a Gibbs distribution on B along with the choice $x = \delta M_h^{-1} b$ as the channel input results in a capacity achieving scheme. However, sampling from Gibbs distribution is known to be exponentially complex in N. Hence b is sampled from a Markov chain instead which is described in the next section.

2.2.2 Zero-forcing with Markov input

Similar to the previous case, the channel input is chosen as $x = \delta M_h^{-1} b$ where b is the information sequence. But now, b is sampled from a 2-state Markov chain as shown in Figure 2.1. The transition matrix is given by

$$T = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{pmatrix} \tag{2.12}$$

where $0 \leq \alpha \leq 1$. Note that $s_n = b_n$ and the information rate of this scheme is $H(B) = H_2(\alpha) = -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$.

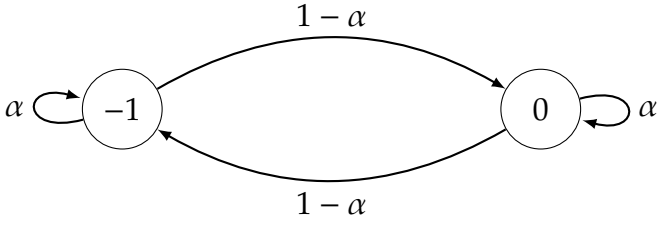


Figure 2.1: 2-State Markov chain

The average transmit power is given by (see [Ganti *et al.*, 2015])

$$P_{zm}(\alpha) = \frac{\delta^2}{N}\text{tr}\left(\mathbb{E}(BB^T)(M_h^T)^{-1}M_h^{-1}\right) = \frac{\delta^2}{N}\text{tr}\left(R(M_h^T)^{-1}M_h^{-1}\right) \tag{2.13}$$

where the correlation matrix $R = \mathbb{E}(BB^T)$ is given by $R_{ij} = (2\alpha - 1)^{|i-j|}$. When N is large, the Toeplitz matrix R can be approximated by a circulant ([Gray, 2006]) and

$$P_z m(\alpha) \to \frac{\delta^2}{2\pi}\int_0^{2\pi} \frac{1}{|f(\lambda)|^2}\left(\frac{2(1-\beta\cos(\lambda))}{1+\beta^2-2\beta\cos(\lambda)} - 1\right)d\lambda \tag{2.14}$$

where $\beta = 2\alpha - 1$. The value of α is to be chosen so as to maximize $H_2(\alpha)$ subject to the constraint $P_z m(\alpha) \leq P$. Thus the rate achieved at power P, denoted $R_m(P)$ is given by

$$R_m(P) = \max_{\alpha: P_{zm}(\alpha)\leq P} H_2(\alpha). \tag{2.15}$$

Finally, the channel $(1, \epsilon), |\epsilon| < 1$ is considered under the zero-forcing with Markov input. The achievable rate is found to be (see [Ganti *et al.*, 2015])

$$R_m(P) = \begin{cases} 1, & P \geq \overline{P}_h \\ H_2\left(\frac{1}{2} + \frac{1}{2\epsilon}\frac{1-P\delta^{-2}(1-\epsilon^2)}{1+P\delta^{-2}(1-\epsilon^2)}\right), & \underline{P}_h \leq P < \overline{P}_h \\ 0, & P < \underline{P}_h \end{cases} \tag{2.16}$$

where $\overline{P}_h = \frac{\delta^2}{1-\epsilon^2}$ and $\underline{P}_h = \frac{\delta^2}{(1+\epsilon)^2}$.

It is observed that the capacity $C_{N,\delta}(P)$ and the achievable rate $R_m(P)$ match at $\overline{P}_h$ and $\underline{P}_h$. The plots of $C_{N,\delta}(P)$ and $R_m(P)$ versus the normalized energy P/δ^2 are also given in [Ganti *et al.*, 2015].

CHAPTER 3

3-LEVEL OUTPUT QUANTIZATION

In this chapter, we present our work on computing the capacity of the approximate ISI channel (figure 1.3) with average-power constrained continuous input and 3-level output quantization.

3.1 Approximate capacity

The capacity of the approximate ISI channel in figure 1.3 is given by

$$C_{N,\delta}(P) = \sup_{\substack{\mathbb{E}[\|X\|^2] \leq NP \\ |Y_n \pm d| \geq \delta}} \frac{I(X,S)}{N} = \frac{H(S)}{N} \tag{3.1}$$

where X and S are the column vectors of input and quantized output respectively, $I(\cdot,\cdot)$ is the mutual information and $H(\cdot)$ is the entropy. The last equality in equation (3.1) is because the vector S is a deterministic function of the input vector X in the absence of any noise in the approximate channel. Since $S \in \{+1, 0, -1\}^N$, we have $C_{N,\delta}(P) \leq \log_2(3)$.

3.1.1 Description of the 3-level quantizer

The 3-level quantizer $Q_d(\cdot)$ described in equation (1.4) can also be described in terms of two 2-level quantizers. Let $s = Q_d(y)$ where y is the intermediate output of the

channel before quantization. Define two 2-level quantized outputs as follows:

$$s_1 = \begin{cases} +1, & y \geq d \\ -1, & y < d \end{cases}$$
$$s_2 = \begin{cases} +1, & y \geq -d \\ -1, & y < -d \end{cases} \tag{3.2}$$

Now s can defined as

$$s = \begin{cases} +1, & s_1 = +1 \ \& \ s_2 = +1 \\ 0, & s_1 = -1 \ \& \ s_2 = +1 \\ -1, & s_1 = -1 \ \& \ s_2 = -1 \end{cases} \tag{3.3}$$

It can be easily seen that this description is same as that in equation (1.4). With this description, we have:

$$\begin{aligned} |y - d| \geq \delta \Leftrightarrow s_1(y - d) \geq \delta \\ |y + d| \geq \delta \Leftrightarrow s_2(y + d) \geq \delta \end{aligned} \tag{3.4}$$

Thus the constraints become linear.

3.1.2 Power constraint

We bound the power of the input sequence/vector X required for a given output sequence/vector S. Given $S = s \in \{1, 0, -1\}^N$, we have the constraints $|Y_n - d| = s_1(n)(Y_n - d) \geq \delta$ and $|Y_n + d| = s_2(n)(Y_n + d) \geq \delta$. So, the minimum energy $\varepsilon(s)$ required for a given output sequence s is given by the following optimization

problem

$$\varepsilon(s) = \min_{\substack{x \\ \text{diag}(s_1)(M_h x - d\mathbf{1}) \geq \delta\mathbf{1} \\ \text{diag}(s_2)(M_h x + d\mathbf{1}) \geq \delta\mathbf{1}}} \left(||X||^2 \right) \tag{3.5}$$

where $s_1, s_2 \in \{-1, 1\}^N$, for $i \in [N]$

$$\begin{aligned} s_1(i) &= \begin{cases} +1, & (M_h x)(i) \geq d \\ -1, & (M_h x)(i) < d \end{cases} \\ s_2(i) &= \begin{cases} +1, & (M_h x)(i) \geq -d \\ -1, & (M_h x)(i) < -d \end{cases} \end{aligned} \tag{3.6}$$

and $s(i)$ is described according to (3.3). The inequality constraints are linear and the feasible set for x is the intersection of hyperplanes, and thus convex. Hence the above optimization problem is a convex optimization problem. Further, similar to the 2-level case, we have

$$\mathbb{E}[||X||^2] = \sum_{s \in \{-1,0,1\}^N} \mathbb{P}(S = s)\mathbb{E}[||X||^2 | S = s] \geq \sum_{s \in \{-1,0,1\}^N} \mathbb{P}(S = s)\varepsilon(s) \tag{3.7}$$

Since the input X has an average-power constraint, this implies

$$\sum_{s \in \{-1,0,1\}^N} \mathbb{P}(S = s)\varepsilon(s) \leq NP \tag{3.8}$$

3.1.3 Entropy maximization

Letting $\varepsilon_{\min} = \min_s \varepsilon(s)$, $\varepsilon_{\max} = \max_s \varepsilon(s)$ and $\bar{\varepsilon} = \frac{1}{3^N} \sum_s \varepsilon(s)$, under the linear constraints (3.8), it is known [Jaynes, 1957] that the Gibbs distribution maximizes

the entropy $H(S)$ for $\varepsilon_{\min} \leq NP \leq \varepsilon_{\max}$. The optimal Gibbs distribution is given by

$$\mathbb{P}(S = s) = \frac{e^{-\beta\varepsilon(s)}}{Z},\ s \in \{-1, 0, 1\}^N \tag{3.9}$$

where Z is the normalizing constant and β is the unique value for which (3.8) is met with equality. The maximum entropy and capacity are given by

$$\begin{aligned} H(S) &= \beta NP + \ln(Z) \\ C_{N,\delta}(P) &= \beta P + \frac{\ln(Z)}{N}. \end{aligned} \tag{3.10}$$

It is easy to see that for $NP < \varepsilon_{\min}$ there exists no valid probability distribution. For $NP = \bar{\varepsilon}$, it is known [Jaynes, 1957] [Conrad, 2013] that $\beta = 0$ and Gibbs distribution is the uniform distribution on $\{-1, 0, 1\}^N$. The capacity for different ranges of NP is given in table 3.1.

Table 3.1: Capacity of Approximate ISI channel with 3-level output quantization

Range of NP	$C_{N,\delta}(P)$	Remarks
$NP < \varepsilon_{\min}$	0	No state if feasible
$\varepsilon_{\min} \leq NP < \bar{\varepsilon}$	$\frac{1}{\ln 2}(\beta P + \frac{\ln(Z)}{N})$	β is such that (3.8) is met with equality
$NP \geq \bar{\varepsilon}$	$\log_2(3)$	$\beta = 0$, uniform distribution

It is now required to compute $\varepsilon(s), \forall s \in \{-1, 0, 1\}^N$.

3.2 Solution of the optimization problem

We begin by defining Strongly diagonally dominant channels.

Definition 3.2.1 (Strongly diagonally dominant channel)**.** A discrete time, finite tap channel h is said to be strongly diagonally dominant if the $N \times N$ channel

matrix M_h is such that $\forall J \subset [N]$, the inverse of the matrix obtained by deleting the rows and columns of $(M_h M_h^T)$ corresponding to index set J is diagonally dominant.

We also have the following lemma:

Lemma 3.2.1. *Let P be an $N \times N$ matrix such that P^{-1} is row-diagonally dominant (with $(P^{-1})_{ii} > 0, \forall i \in [N]$). Also $\forall J \subset [N]$, let the matrix obtained by deleting the rows and columns of P corresponding to J, denoted Q_J be invertible. Then Q_J^{-1} is row diagonally dominant.*

The reader is referred to appendix A for the proof.

3.2.1 The optimization problem

The optimization problem (3.5) is restated here:

$$\varepsilon(s) = \min_{\substack{x \\ \text{diag}(s_1)(M_h x - d\mathbf{1}) \geq \delta \mathbf{1} \\ \text{diag}(s_2)(M_h x + d\mathbf{1}) \geq \delta \mathbf{1}}} \left(||x||^2 \right). \tag{3.11}$$

Given $s \in \{-1, 0, 1\}^N$, $s_1, s_2 \in \{-1, 1\}^N$ are defined through

$$s(i) = \begin{cases} +1, & s_1(i) = +1 \ \& \ s_2(i) = +1 \\ 0, & s_1(i) = -1 \ \& \ s_2(i) = +1 \\ -1, & s_1(i) = -1 \ \& \ s_2(i) = -1 \end{cases} \tag{3.12}$$

3.2.2 The dual problem and the solution

The Lagrangian function of the optimization problem is

$$L(x, \lambda_1, \lambda_2) = \|x\|^2 + \lambda_1^T\Big(\delta\mathbf{1} - \text{diag}(s_1)(M_h x - d\mathbf{1})\Big) + \lambda_2^T\Big(\delta\mathbf{1} - \text{diag}(s_2)(M_h x + d\mathbf{1})\Big) \quad (3.13)$$

where $\lambda_1, \lambda_2 \in \mathcal{R}^N$ are the Lagrange multipliers.

Since the primal problem is convex with linear constraints, strong duality holds. So, if x^* is primal optimal and $(\lambda_1^*, \lambda_2^*)$ is dual optimal then $\inf_x L(x, \lambda_1^*, \lambda_2^*) = L(x^*, \lambda_1^*, \lambda_2^*)$, and this x^* can be found by setting the gradient of the Lagrangian function with respect to x to zero; we get $x^* = \frac{1}{2}\Big(M_h^T \text{diag}(s_1)\lambda_1^* + M_h^T \text{diag}(s_2)\lambda_2^*\Big)$. Also by complementary slackness we have, for $i \in [N]$, $\lambda_1^*(i)\Big(\delta\mathbf{1} - \text{diag}(s_1)(M_h x^* - d\mathbf{1})\Big)(i) = 0$ and $\lambda_2^*(i)\Big(\delta\mathbf{1} - \text{diag}(s_2)(M_h x^* + d\mathbf{1})\Big)(i) = 0$. Now it can be shown that $\lambda_1^*(i)\lambda_2^*(i) = 0$ since the corresponding constraints cannot be simultaneously active (as long as $d > \delta$).

Now it is to be found for which $i \in [N]$ the multipliers have to be zero. Suppose for some i we have $s(i) = 1$ then the second constraint becomes strict inequality and hence, by complimentary slackness, $\lambda_2^*(i) = 0$. Similarly, if $s(i) = -1$ then $\lambda_1^*(i) = 0$. If $s(i) = 0$ then, intuitively, the minimum energy input x would be such that $y(i) = 0$. So we set $\lambda_1^*(i)\lambda_2^*(i) = 0$ but we do not yet know where each of these multipliers are non-zero.

Let $J_0 = \{i \in [N] : s(i) = 0\}$, $J_1 = \{i \in [N] : s(i) = 1\}$ and $J_{-1} = \{i \in [N] : s(i) = -1\}$. So if $i \in J_1$ then $\lambda_2^*(i) = 0$, and if $i \in J_{-1}$ then $\lambda_1^*(i) = 0$. We now need to find the

optimal values of remaining multipliers. The Lagrangian dual is

$$\begin{aligned} L(\lambda_1, \lambda_2) =& \delta(\lambda_1^T + \lambda_2^T)\mathbf{1} + d(\lambda_1^T s_1 - \lambda_2^T s_2) \\ & - \frac{1}{4}\Big(\lambda_1^T \text{diag}(s_1) M_h M_h^T \text{diag}(s_1)\lambda_1 + \lambda_1^T \text{diag}(s_1) M_h M_h^T \text{diag}(s_2)\lambda_2 \\ & + \lambda_2^T \text{diag}(s_2) M_h M_h^T \text{diag}(s_1)\lambda_1 + \lambda_2^T \text{diag}(s_2) M_h M_h^T \text{diag}(s_2)\lambda_2\Big) \end{aligned} \tag{3.14}$$

We must also know the indices in J_0 where the multipliers are non zero. But we have the following lemma.

Lemma 3.2.2. *Given $s \in \{-1, 0, 1\}^N$ and channel matrix M_h, there exists $d_0 > 0$ such that $\forall d > d_0$, we have $\lambda_1(i) = 0 = \lambda_2(i), \forall i \in J_0$.*

The reader is referred to the end of appendix B section B.3 for the proof.

So, in this section, we only consider the case of $d >> \delta$ such that $\lambda_1(i) = 0 = \lambda_2(i), \forall i \in J_0$. For a more general discussion one can refer to appendix B where we present an algorithm for finding the indices in J_0 where the multipliers are non-negative.

We can simplify this by removing the zero values of multipliers. Let

$$\begin{aligned} A &= \text{diag}(s_1) M_h M_h^T \text{diag}(s_1) \\ B &= \text{diag}(s_1) M_h M_h^T \text{diag}(s_2) \\ C &= \text{diag}(s_2) M_h M_h^T \text{diag}(s_1) \\ D &= \text{diag}(s_2) M_h M_h^T \text{diag}(s_2) \end{aligned} \tag{3.15}$$

With this notation, we get

$$\begin{aligned} L(\lambda_1, \lambda_2) =& \delta(\lambda_1^T)^{J_1}\mathbf{1}^{J_1} + \delta(\lambda_2^T)^{J_{-1}}\mathbf{1}^{J_{-1}} + d((\lambda_1^T)^{J_1}s_1^{J_1} - (\lambda_2^T)^{J_{-1}}s_2^{J_{-1}}) \\ & - \frac{1}{4}\Big((\lambda_1^T)^{J_1}A^{J_1J_1}\lambda_1^{J_1} + (\lambda_1^T)^{J_1}B^{J_1J_{-1}}\lambda_2^{J_{-1}} \\ & + (\lambda_2^T)^{J_{-1}}C^{J_{-1}J_1}\lambda_1^{J_1} + (\lambda_2^T)^{J_{-1}}D^{J_{-1}J_{-1}}\lambda_2^{J_{-1}}\Big) \end{aligned} \tag{3.16}$$

where the superscripts denote the sub-matrix corresponding to the rows and/or columns represented by them. It can also be seen that $A^{J_1J_1} = \text{diag}(s_1)^{J_1J_1}(M_hM_h^T)^{J_1J_1}\text{diag}(s_1)^{J_1J_1}$ since $\text{diag}(s_1)$ is a diagonal matrix; similarly for B, C and D.

Setting the gradient of the Lagrangian dual with respect to $\lambda_1^{J_1}$ and $\lambda_2^{J_{-1}}$ to zero to find the optimal values of the multipliers, we get

$$\begin{aligned} &\begin{pmatrix} \text{diag}(s_1)^{J_1}\, P^{J_1J_1}\, \text{diag}(s_1)^{J_1} & \text{diag}(s_1)^{J_1}\, P^{J_1J_{-1}}\, \text{diag}(s_2)^{J_{-1}} \\ \text{diag}(s_2)^{J_{-1}}\, P^{J_{-1}J_1}\, \text{diag}(s_1)^{J_1} & \text{diag}(s_2)^{J_{-1}}\, P^{J_{-1}J_{-1}}\, \text{diag}(s_2)^{J_{-1}} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_1} \\ \lambda_2^{J_{-1}} \end{pmatrix} \\ &= 2\begin{pmatrix} \delta\mathbf{1}^{J_1} + d\mathbf{s_1}^{J_1} \\ \delta\mathbf{1}^{J_{-1}} - d\mathbf{s_2}^{J_{-1}} \end{pmatrix} \end{aligned} \tag{3.17}$$

where $P = (M_hM_h^T)$. This can also be written as

$$\begin{aligned} &\begin{pmatrix} P^{J_1J_1}\, \text{diag}(s_1)^{J_1} & P^{J_1J_{-1}}\, \text{diag}(s_2)^{J_{-1}} \\ P^{J_{-1}J_1}\, \text{diag}(s_1)^{J_1} & P^{J_{-1}J_{-1}}\, \text{diag}(s_2)^{J_{-1}} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_1} \\ \lambda_2^{J_{-1}} \end{pmatrix} \\ &= 2\begin{pmatrix} \delta\mathbf{s_1}^{J_1} + d\mathbf{1}^{J_1} \\ \delta\mathbf{s_2}^{J_{-1}} - d\mathbf{1}^{J_{-1}} \end{pmatrix} \end{aligned}. \tag{3.18}$$

Note that $i \in J_1 \implies s_1(i) = +1$ and $i \in J_{-1} \implies s_2(i) = -1$. Hence we have

$$\begin{pmatrix} P^{J_1 J_1} & -P^{J_1 J_{-1}} \\ P^{J_{-1} J_1} & -P^{J_{-1} J_{-1}} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_1} \\ \lambda_2^{J_{-1}} \end{pmatrix} = 2 \begin{pmatrix} (\delta + d)\mathbf{1}^{J_1} \\ (-\delta - d)\mathbf{1}^{J_{-1}} \end{pmatrix} = 2(\delta + d) \begin{pmatrix} \mathbf{1}^{J_1} \\ -\mathbf{1}^{J_{-1}} \end{pmatrix} \tag{3.19}$$

Assuming that the matrix in the left hand side of the above equation to be invertible, we have the optimal values of the multipliers unless these values are negative. In the next section we give sufficient conditions for non-negativity of the solution to equation (3.19): **diagonally dominant channels**.

3.2.3 Conditions for non-negativity: diagonally dominant channels

Let $Z = \begin{pmatrix} P^{J_1 J_1} & -P^{J_1 J_{-1}} \\ P^{J_{-1} J_1} & -P^{J_{-1} J_{-1}} \end{pmatrix}$ and let Q be the matrix obtained from P by removing the rows and columns corresponding to indices in J_0. Let $\lambda = \begin{pmatrix} \lambda_1^{J_1} \\ \lambda_2^{J_{-1}} \end{pmatrix}$, $W = 2(\delta + d) \begin{pmatrix} \mathbf{1}^{J_1} \\ -\mathbf{1}^{J_{-1}} \end{pmatrix}$ and $N = \begin{pmatrix} I_{J_1} & \mathbf{0} \\ \mathbf{0} & -I_{J_{-1}} \end{pmatrix}$ where I_{J_1} and$I_{J_{-1}}$ are identity matrices of dimensions $|J_1|$ and$|J_{-1}|$ respectively. The matrices Z, W and N are dependent on the output state s through the sets J_0, J_1 and J_{-1}.

We can write Z as $Z = UQU^T N$ where U is a permutation matrix (and hence unitary). So the equation (3.19) becomes

$$UQU^T N\lambda = W \tag{3.20}$$

So, if Q is invertible then

$$\lambda = NUQ^{-1}U^TW \tag{3.21}$$

It can be seen that $UQ^{-1}U^T = \begin{pmatrix} (Q^{-1})^{J_1J_1} & (Q^{-1})^{J_1J_{-1}} \\ (Q^{-1})^{J_{-1}J_1} & (Q^{-1})^{J_{-1}J_{-1}} \end{pmatrix}$.

So if $Q^{-1} = (t_{ij})$,

$$\begin{aligned} i \in J_1 &\implies \lambda_1(i) = 2(\delta + d) \sum_{j \in [N] \setminus J_0} t_{ij}(-1)^{\mathbb{1}[j \in J_{-1}]} \\ i \in J_{-1} &\implies \lambda_2(i) = -2(\delta + d) \sum_{j \in [N] \setminus J_0} t_{ij}(-1)^{\mathbb{1}[j \in J_{-1}]} \end{aligned} \tag{3.22}$$

where $\mathbb{1}[x \in A]$ is 1 if $x \in A$ and 0 otherwise .

We want the multipliers to be non-negative for any J_0, J_1 and J_{-1}. Suppose $i \in J_1$. We want $\lambda_1(i) \geq 0$. It is sufficient if $\sum_{j \in [N] \setminus J_0} t_{ij}(-1)^{\mathbb{1}[j \in J_{-1}]} \geq 0$. Similarly, for $i \in J_{-1}$, it is sufficient if $\sum_{j \in [N] \setminus J_0} t_{ij}(-1)^{\mathbb{1}[j \in J_{-1}]} \leq 0$. So if Q^{-1} is diagonally dominant (with diagonal entries being non-negative) then the above sufficient conditions are satisfied and hence $\lambda_1^{J_1}$ and $\lambda_2^{J_{-1}}$ are non-negative. So if P satisfies the hypothesis of lemma 3.2.1 then Q^{-1} is diagonally dominant and hence the optimal multipliers in (3.22) are non-negative.

Further, using (3.17) and $x^* = \frac{1}{2}\left(M_h^T \text{diag}(s_1)\lambda_1^* + M_h^T \text{diag}(s_2)\lambda_2^*\right)$ we get the following under-determined system of equations for x^*:

$$\begin{aligned} \text{diag}(s_1)M_h^{J_1[]}x^* &= \delta \mathbf{1}^{J_1} + d s_1^{J_1} \\ \text{diag}(s_2)M_h^{J_{-1}[]}x^* &= \delta \mathbf{1}^{J_{-1}} - d s_2^{J_{-1}} \end{aligned} \tag{3.23}$$

where $M_h^{J[]}$ denotes the sub-matrix corresponding to rows with indices J and all

columns. This can re-written as

$$\begin{pmatrix} M_h^{J_1[]} \\ M_h^{J_{-1}[]} \end{pmatrix} x^* = \begin{pmatrix} \delta s_1^{J_1} + d\mathbf{1}^{J_1} \\ \delta s_2^{J_{-1}} - d\mathbf{1}^{J_{-1}} \end{pmatrix}. \tag{3.24}$$

It is well known that for an under-determined system $Ax = b$ with (AA^T) invertible, the least norm solution is given by $x = A^T(AA^T)^{-1}b$. Letting $\bar{H} = \begin{pmatrix} M_h^{J_1[]} \\ M_h^{J_{-1}[]} \end{pmatrix}$ and $\bar{s} = \begin{pmatrix} \delta s_1^{J_1} + d\mathbf{1}^{J_1} \\ \delta s_2^{J_{-1}} - d\mathbf{1}^{J_{-1}} \end{pmatrix}$, we have

$$x^* = \bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{s}. \tag{3.25}$$

Note that $\bar{H}\bar{H}^T = UQU^T$ and from equations (3.21) and (3.18) we have

$$\begin{pmatrix} (\lambda_1^*)^{J_1} \\ (\lambda_2^*)^{J_{-1}} \end{pmatrix} = 2\begin{pmatrix} \mathrm{diag}(s_1)^{J_1} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s_2)^{J_{-1}} \end{pmatrix} UQ^{-1}U^T\bar{s}. \tag{3.26}$$

So clearly $\bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{s} = \frac{1}{2}\Big(M_h^T\mathrm{diag}(s_1)\lambda_1^* + M_h^T\mathrm{diag}(s_2)\lambda_2^*\Big)$.

Therefore the solution to the optimization problem (3.11) is the least norm solution to (3.24) which is given by (3.25). Hence the optimal energy for the case of large d is given by

$$\varepsilon(s) = \bar{s}^T(\bar{H}\bar{H}^T)^{-1}\bar{s} = (\delta + d)^2\tilde{s}^TQ^{-1}\tilde{s} \tag{3.27}$$

where Q is the sub-matrix obtained from $M_hM_h^T$ by removing the rows and columns corresponding to J_0 and $\tilde{s}$ is the column vector obtained by removing the entries corresponding to J_0 from the quantized output sequence/vector s.

3.3 Zero forcing input

It can be seen that, for output sequence s such that $J_0 = \emptyset$, the equation (3.25) for minimum energy input reduces to

$$x^* = (\delta + d)M_h^{-1}s. \tag{3.28}$$

This is called the zero forced input because it involves channel inversion. In this section we consider only the zero-forced input to the channel as an approximation to the minimum energy input given by equation (3.25). The zero forced input x to the channel is given by

$$x = (\delta + d)M_h^{-1}s. \tag{3.29}$$

So the approximate minimum energy, denoted again by $\varepsilon(s)$ is given by

$$\varepsilon(s) = ||x||^2 = (\delta + d)^2 s^T G s \tag{3.30}$$

where $G = (M_h M_h^T)^{-1}$. We characterize the mean energy $\bar{\varepsilon}$ for zero forcing in the following lemma.

Lemma 3.3.1. *The mean energy for zero forcing is given by*

$$\bar{\varepsilon} = \frac{2}{3}(\delta + d)^2 \mathrm{tr}((M_h^T)^{-1}M_h^{-1}) \tag{3.31}$$

where $\mathrm{tr}(A)$ *denotes the trace of the matrix* A*.*

Proof. For $s \in \{-1, 0, 1\}^N$, the energy for zero forcing given by (3.30) can be expanded as

$$\varepsilon(s) = (\delta + d)^2 (\sum_{i=1}^{N} G_{ii} + \sum_{i,j,i\neq j} G_{ij}s_i s_j). \tag{3.32}$$

Hence,

$$\bar{\varepsilon} = \frac{1}{3^N} \sum_{s \in \{-1,0,1\}^N} \varepsilon(s) = (\delta + d)^2 \frac{1}{3^N} \left(\sum_{i=1}^{N} G_{ii}(2 \times 3^{N-1}) + \sum_{i,j,i \neq j} G_{ij} \left(\sum_{s \in \{-1,0,1\}^N} s_i s_j \right) \right). \quad (3.33)$$

Note that, for any $i, j, i \neq j$, we have $\sum_{s \in \{-1,0,1\}^N} s_i s_j = 0$. Therefore

$$\bar{\varepsilon} = \frac{2}{3}(\delta + d)^2 \sum_{i=1}^{N} G_{ii} = \frac{2}{3}(\delta + d)^2 \mathrm{tr}((M_h^T)^{-1} M_h^{-1}). \quad (3.34)$$

□

We can now characterize the average power $\overline{P}_h = \lim_{N \to \infty} \bar{\varepsilon}/N$ which is the minimum average power needed for capacity of $\log_2(3)$ bits in the case of zero forcing, in terms of the discrete time Fourier transform (DTFT) of the channel h. Let the DTFT of the channel be

$$f(\lambda) = \sum_{k=0}^{L-1} h_k e^{jk\lambda}. \quad (3.35)$$

Since M_h is circulant, it can be seen that

$$\bar{\varepsilon} = \frac{2}{3}(\delta + d)^2 \mathrm{tr}((M_h^T)^{-1} M_h^{-1}) = \frac{2}{3}(\delta + d)^2 \sum_{k=1}^{N} \frac{1}{\left| f(\frac{2\pi k}{N}) \right|^2}. \quad (3.36)$$

Using standard arguments, it can be shown that [Gray, 2006]

$$\overline{P}_h \to \frac{2}{3} \frac{(\delta + d)^2}{2\pi} \int_0^{2\pi} \frac{1}{|f(\lambda)|^2} d\lambda. \quad (3.37)$$

So, for large N, the maximum information rate of approximate ISI channel which are strongly diagonally dominant in the case of large d, under zero forcing,

is given by

$$R_{N,\delta}(P) = \begin{cases} \log_2(3), & P \geq \overline{P}_h \\ \beta P + \frac{\ln(Z)}{N}, & \underline{P}_h \leq P < \overline{P}_h \end{cases} \tag{3.38}$$

where $\underline{P}_h = \lim_{N\to\infty} \varepsilon_{\min}/N = 0$. The case $P < \underline{P}_h$ does not arise.

3.4 Achievable schemes

In this section, we consider achievable schemes for the approximate ISI channel which is strongly diagonally dominant. In achievable schemes, a message/information sequence $B \in \{-1, 0, 1\}^N$ with a carefully chosen distribution is encoded into a channel input x that satisfies the constraints $|y_n \pm d| \geq \delta$. The rate of transmission over the approximate ISI channel is $H(B)/N$.

3.4.1 Gibbs distribution

Let $b = \{b_i\}_{i=1}^N \in \{-1, 0, 1\}^N$ be a sample of the information sequence B. Choose the input to the channel according to (3.25): $x = \bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{b}$ (with $\bar{H}$ defined according to b). It can be verified that the output of the quantizer s equals b. As seen before in section 3.1.3, a Gibbs distribution on B and the choice $x = \bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{b}$ as the input to the channel results in a capacity achieving scheme. Hence for strongly diagonally dominant channels, in the case of large d, the scheme $x = \bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{b}$ is optimal when b is sampled from the Gibbs distribution. But this input x is non-trivial in the sense that the matrix $\bar{H}$ itself depends on b. So we consider the zero forcing input in the subsequent discussions since it just involves channel matrix inversion.

Zero forcing with Gibbs distribution

As before, let $b = \{b_i\}_{i=1}^{N} \in \{-1, 0, 1\}^N$ be a sample of the information sequence B. In this case we choose the input to the channel as

$$x = (\delta + d)M_h^{-1}b \tag{3.39}$$

which is the zero forcing input. The output of the ISI channel is $y = M_h x = (\delta + d)b$. So y satisfies the constraints $|y \pm d| \geq \delta$. Hence the output of the quantizer s equals b. As seen in section 3.1.3, a Gibbs distribution on B gives the maximum information rate. This could be less than capacity since the energy for zero forcing (3.30) is greater than or equal to the optimal energy (3.27) for each output sequence $s\{-1, 0, 1\}^N$. But it can be seen through simulations that both these schemes have very close information rate for the case of large d.

However, b is a sequence of length N and we are dealing with large N, and it is known that sampling from a Gibbs distribution has exponential complexity in N. So it is impractical to use this scheme. In the next subsection, we consider the case of zero forcing with the information sequence b sampled from a simple 3-state Markov chain.

3.4.2 Zero forcing with Markov input

Here, we choose $x = (\delta+d)M_h^{-1}b$, where b is the information sequence. The sequence b is sampled from a 3-state Markov chain shown in figure 3.1 with the transition

matrix

$$T = \begin{pmatrix} \alpha & \frac{\beta}{2} & \frac{\beta}{2} \\ \frac{\beta}{2} & \alpha & \frac{\beta}{2} \\ \frac{\beta}{2} & \frac{\beta}{2} & \alpha \end{pmatrix} \tag{3.40}$$

where $0 \leq \alpha \leq 1$ and $\alpha+\beta = 1$. Note that the quantizer output $s(n) = b(n), 1 \leq n \leq N$ and the achievable rate of this scheme is

$$\begin{aligned} H(\alpha, \beta/2, \beta/2) &= -\left(\alpha \log_2(\alpha) + (\beta/2)\log_2(\beta/2) + (\beta/2)\log_2(\beta/2)\right) \\ &= -\left(\alpha \log_2(\alpha) + (1-\alpha)\log_2\left(\frac{1-\alpha}{2}\right)\right) \end{aligned} \tag{3.41}$$

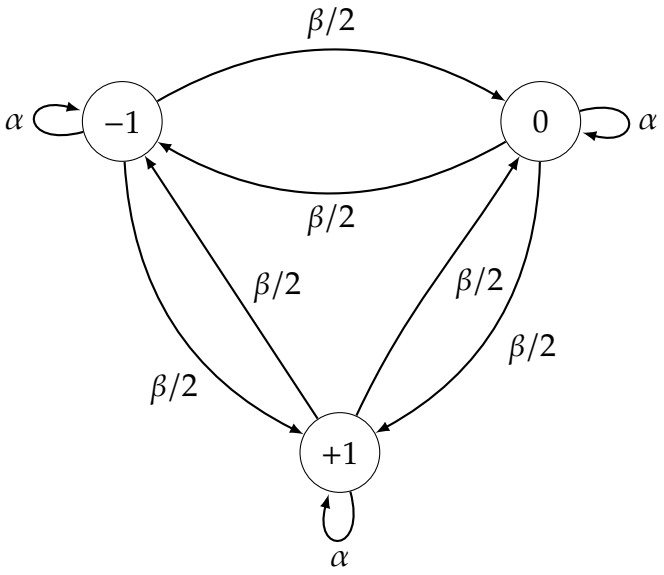


Figure 3.1: 3-State Markov chain

The average transmit power, denoted $P_{zm}(\alpha)$, is given by

$$\begin{aligned} P_{zm}(\alpha) &= \frac{1}{N}\mathbb{E}[\|X\|^2] = \frac{(\delta+d)^2}{N}\mathbb{E}[B^T(M_h^{-1})^T M_h^{-1} B] \\ &= \frac{(\delta+d)^2}{N}\mathrm{tr}(\mathbb{E}[BB^T](M_h^{-1})^T M_h^{-1}). \end{aligned} \tag{3.42}$$

Let $\gamma = 1 - \frac{3\beta}{2}$. The eigenvalues of T are $(1, \gamma, \gamma)$. Using the eigenvalue decom-

position, it can be shown that

$$T^d = \frac{1}{3}\begin{pmatrix} 1+2\gamma^d & 1-\gamma^d & 1-\gamma^d \\ 1-\gamma^d & 1+2\gamma^d & 1-\gamma^d \\ 1-\gamma^d & 1-\gamma^d & 1+2\gamma^d \end{pmatrix}. \tag{3.43}$$

Using $(1/3, 1/3, 1/3)$ as the initial distribution and T^d, we get

$$\mathbb{E}[b(n)b(n+d)] = \frac{2}{3}\gamma^d. \tag{3.44}$$

Hence the correlation matrix $R = \mathbb{E}[BB^T]$ is given by $R_{ij} = \frac{2}{3}\gamma^{|i-j|}$. Therefore,

$$P_{zm}(\alpha) = \frac{(\delta+d)^2}{N}\text{tr}(R(M_h^{-1})^T M_h^{-1}). \tag{3.45}$$

Note that R is a Toeplitz matrix. When N is large, it can be approximated by a circulant matrix [Gray, 2006] and

$$P_{zm}(\alpha) \to \frac{2}{3}\frac{(\delta+d)^2}{2\pi}\int_0^{2\pi} \frac{1}{|f(\lambda)|^2}\left(\frac{2(1-\gamma\cos(\lambda))}{1+\gamma^2-2\gamma\cos(\lambda)} - 1\right)d\lambda \tag{3.46}$$

where $f(\lambda) = \sum_{k=0}^{L-1} h_k e^{jk\lambda}$ is the DTFT of the channel h. It is required to choose α so as to maximize the entropy $H(\alpha, \beta/2, \beta/2)$. The rate achieved at power P, denoted $R_m(P)$, is given by

$$R_m(P) = \max_{\alpha: P_{zm}(\alpha) \le P} H(\alpha, \beta/2, \beta/2) \tag{3.47}$$

3.5 Numerical results for $(1, \epsilon)$ channel

In this section, we evaluate the capacity of approximate ISI channel and the rate achieved by the Markov scheme in section 3.4.2 for the $(1, \epsilon)$ channel. For numerical evaluation, we choose $\delta = 0.3$.

The channel $(1, \epsilon), |\epsilon| \leq 1$ is strongly diagonally dominant if $|\epsilon| < \sqrt{\frac{3}{2}} - 1$ (refer appendix C). We consider $0 \leq \epsilon \leq \sqrt{\frac{3}{2}} - 1$ only. For this channel,

$$f(\lambda) = 1 + \epsilon e^{j\lambda}. \tag{3.48}$$

From (3.37), the minimum average power required for zero forcing is given by

$$\overline{P}_h = \frac{2}{3} \frac{(\delta + d)^2}{1 - \epsilon^2}. \tag{3.49}$$

The transmit power required for the 3-state Markov scheme (3.46) is

$$P_{zm}(\alpha) = \frac{2}{3} \frac{(\delta + d)^2}{1 - \epsilon^2} \left(\frac{1 - \epsilon\gamma}{1 + \epsilon\gamma} \right) \tag{3.50}$$

where $\gamma = 1 - 3\beta/2 = (3\alpha - 1)/2$. Since $0 \leq \alpha \leq 1$, we have

$$P_{zm}(\alpha) \geq \frac{2}{3} \frac{(\delta + d)^2}{(1 + \epsilon)^2} = \underline{P}_h. \tag{3.51}$$

Hence the problem of maximum entropy for the 3-state Markov chain translates to

$$R_m(P) = \max_{\alpha \geq \frac{1}{3} + \frac{2}{3\epsilon} \left(\frac{1 - (3/2)P(\delta+d)^{-2}(1-\epsilon^2)}{1 + (3/2)P(\delta+d)^{-2}(1-\epsilon^2)} \right)} H(\alpha, \beta/2, \beta/2). \tag{3.52}$$

Note that the constraint in the above equation can also be written as $\alpha \geq \frac{1}{3} + \frac{2}{3\epsilon} \left(\frac{\overline{P}_h - P}{\overline{P}_h + P} \right)$.

The solution of the above optimization problem is given by

$$R_m(P) = \begin{cases} \log_2(3), & P \geq \overline{P}_h \\ H\left(\alpha = \frac{1}{3} + \frac{2}{3\epsilon}\left(\frac{\overline{P}_h - P}{\overline{P}_h + P}\right), \beta/2, \beta/2\right), & \underline{P}_h \leq P < \overline{P}_h \\ 0, & P < \underline{P}_h = \frac{2}{3}\frac{(\delta+d)^2}{(1+\epsilon)^2} \end{cases} \tag{3.53}$$

It can be observed that $R_{N,\delta}(P)$ from (3.38) and the achievable rate $R_m(P)$ match at $\overline{P}_h$

3.5.1 Simulation

In Figure 3.2, the approximate capacity $C_{N,\delta}(P)$, the maximum information rate under zero forcing $R_{N,\delta}(P)$ and the achievable rate of Markov scheme $R_m(P)$ are plotted as a function of the normalized power $P/(\delta + d)^2$ for $\epsilon = 0.1$, $\delta = 0.3$, $d = 50$ and $N = 10$. The same are plotted for $\epsilon = 0.2$ in Figure 3.3. It can be observed that, as ϵ increases, the gap between the optimal scheme and the zero forcing scheme increases.

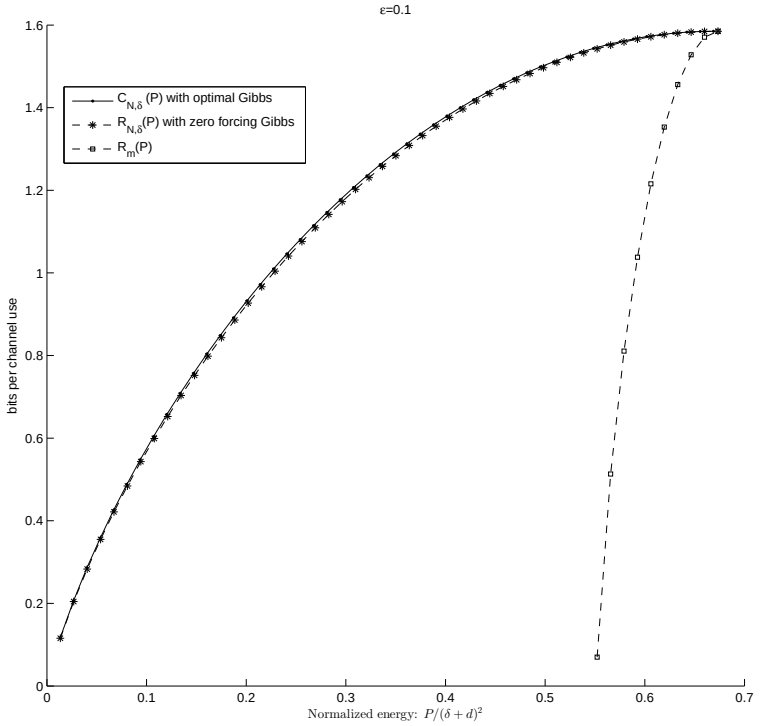


Figure 3.2: $C_{N,\delta}(P)$, $R_{N,\delta}(P)$ and $R_m(P)$ versus normalized $P/(\delta + d)^2$ for $\epsilon = 0.1$

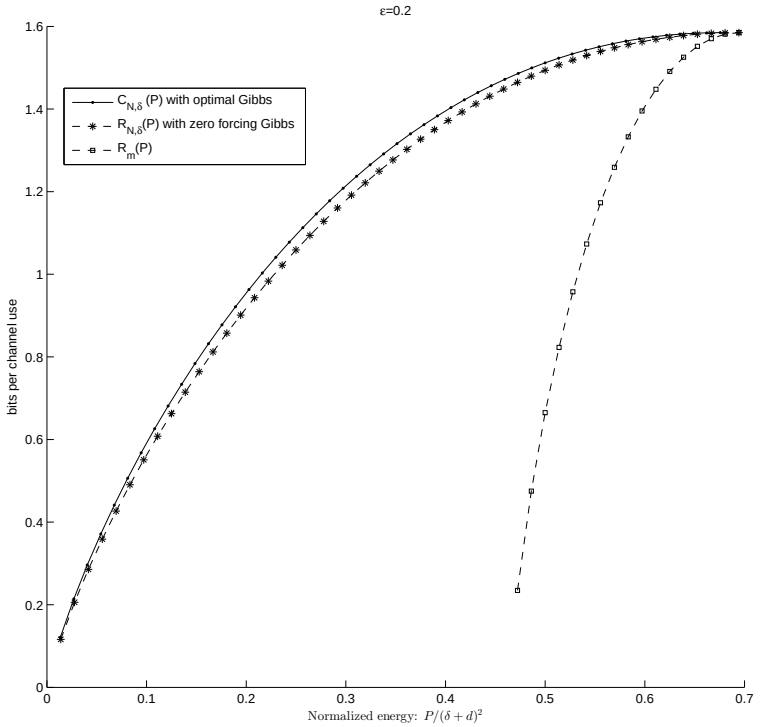


Figure 3.3: $C_{N,\delta}(P)$, $R_{N,\delta}(P)$ and $R_m(P)$ versus normalized $P/(\delta + d)^2$ for $\epsilon = 0.2$

CHAPTER 4

SOURCE SIMULATION

So far, we have characterized the capacity of approximate ISI channel with average power constrained continuous input and quantized output. Using the approximate channel as a precoder we can come up with signaling methods for the actual channel with AWGN noise. In this chapter, we look at how to do signaling and discuss about simulating the entire system so as to be able to prove that a certain rate is achievable.

4.1 Signaling for approximate ISI channel

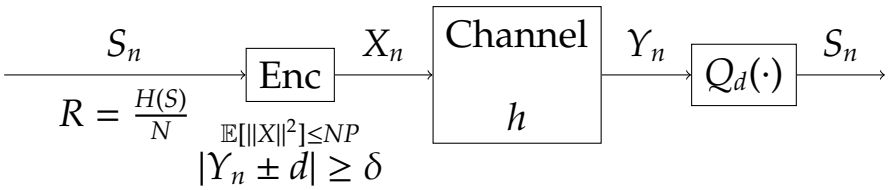


Figure 4.1: Signaling for approximate ISI channel with 3-level quantizer

In Figure 4.1, the input to the approximate ISI channel is the optimum minimum energy input (3.25) for a particular input vector/sequence $s \in \{-1, 0, 1\}^N$. Given s sampled according to the optimal Gibbs distribution and channel h, the encoder looks up for the optimal x^* according to (3.25) and that is the input to the approximate ISI channel. It can be seen that the output after quantization is indeed s and the achievable rate is the $H(S)/N$ which is the capacity of the approximate ISI

channel. So given a sequence of iid message bits (with equal probability of being 0 or 1), power P and the channel h, if we can encode in to the alphabet $\{-1, 0, 1\}^N$ such that the outputs $s \in \{-1, 0, 1\}^N$ are according to the Gibbs distribution then this is a capacity achieving scheme. A similar argument holds for the approximate ISI channel with 2-level quantized output as well. In the next section, we discuss signaling method for the actual ISI channel with a 2-level quantized output since the analysis is slightly simpler compared to the 3-level case. Nonetheless, a similar argument holds for the latter case as well.

4.2 Signaling for actual ISI channel with noise

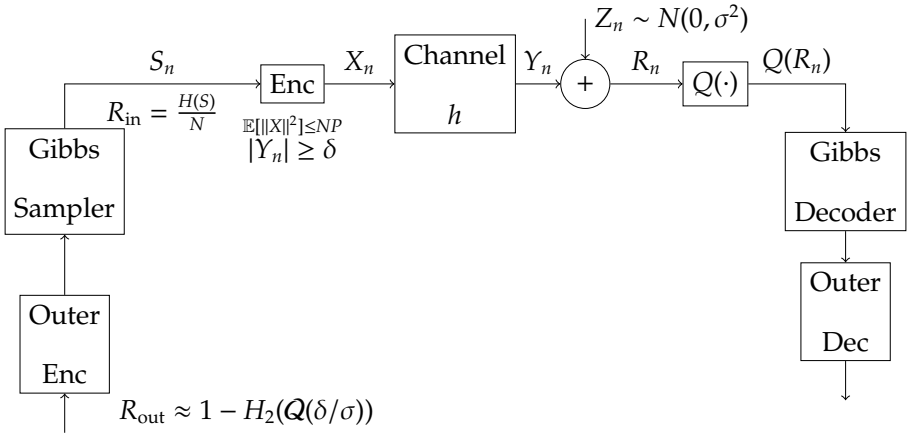


Figure 4.2: Signaling for actual ISI channel with noise

The idea here is similar to the signaling for the approximate ISI channel: given a channel h and power constraint P and a sequence of equally likely iid message bits, sample $S = s \in \{-1, 1\}^N$ from the optimal Gibbs distribution for the corresponding approximate ISI channel with 2-level quantization and encode it into the optimal x^* (2.6), and give it as input to the actual channel. Due to noise, the output sequence after quantization need not be s. Decoding it back to the bit sequence, we need to

a use an error control code to correct the errors. From Figure 4.2, it can be seen that the Gibbs sampler sees a binary symmetric channel with probability of error $p \leq H(\delta/\sigma)$ (but interleaving might be necessary since the channel has memory). The capacity of a binary symmetric channel with probability of error p is $1 - H_2(p)$. Hence an outer error correcting code of rate $R_{\text{out}} = 1 - H_2(Q(\delta/\sigma))$ can be used. This seems to imply that an overall rate of $R_{\text{out}}H(S)/N = 1 - H_2(Q(\delta/\sigma))C_{N,\delta}(P)$ should be achievable. So, if we can simulate this system and be able to achieve bit error rate (BER) as low as we want for some SNR (signal to noise ratio) then we can say that the above rate is achievable. In the next section, we have a brief discussion regarding source simulation and suggest a method to simulate the above system. But due to computational complexity of the method we are not able to get to the required rate.

4.3 Source simulation

In this a section, we describe a well-known method of simulating an arbitrary discrete probability distribution from a sequence of iid Bernoulli 1/2 (Ber(1/2)) random bits.

Let the target distribution be $\{p_1, p_2, \ldots, p_k\}$ with samples $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$. Let $\{X_n\}_{n\in\mathbb{N}}$, $X_n \in \{0, 1\}$ be iid Ber(1/2). Let $U = \sum_{n=1}^{\infty} X_n/2^n$. Note that $U \sim \text{Unif}[0, 1]$. Let $U_m = \sum_{k=1}^{m} X_k/2^k, m \geq 1$. Let $c_0 = 0$ and $c_j = \sum_{i=1}^{j} p_i, 1 \leq j \leq k$ where c_j are values of the cumulative distribution function of the target distribution. We have the following algorithm [Romik, 1999]:

Algorithm 1 Algorithm to sample from a given discrete distribution

loop:
Sample $X_1, X_2, \ldots$ one at a time
if $\exists m$:[$\exists j$ such that $c_{j-1} < U_m < U_m + \frac{1}{2^m} < c_j$] **then**
    Output α_j
    break
else
    goto *loop*

We also have the following theorem from [Romik, 1999]:

Theorem 4.3.1. *If $H(p_1, p_2, \ldots, p_k)$ denote the entropy of the target distribution, then the number of iid* Ber(1/2) *samples N required in the source simulation algorithm 1 satisfies*

$$H(p_1, p_2, \ldots, p_k) \leq \mathbb{E}[N] \leq H(p_1, p_2, \ldots, p_k) + 4 \tag{4.1}$$

But there is a serious issue with this method of sampling: there exists many input sequences that leads the same sample. Essentially, the sampler is not one-one. This leads to problem while decoding. We haven't been able to successfully tackle this problem. In the next subsection, we present a different approach based on the theory of typical sequences.

4.3.1 Source simulation: typical sequences

Here, instead of sampling from the full Gibbs distribution over 2^N sample space, we choose top k highly probable samples; we set the probability of other samples to 0 and normalize to get a new distribution. Call this distribution Gibbs-1. From

the sample space of Gibbs-1 (only non-zero probability samples are considered), form iid sequences of length m. Now we have a joint probability distribution on the space of m length sequences of vectors sampled from Gibbs-1; call this sample space Γ. From the theory of typical sequences and typical sets, for large enough m, given $\epsilon > 0$, there is a set $A_m^{(\epsilon)} \subset \Gamma$ such that

$$2^{-m(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_m) \leq 2^{-m(H(X)-\epsilon)}, (x_1, x_2 \ldots, x_m) \in A_m^{(\epsilon)} \tag{4.2}$$

where $X \sim$ Gibbs-1. So the elements in the typical set have almost uniform distribution. We also have,

$$\mathbb{P}(A_m^{(\epsilon)}) \geq 1 - \epsilon \tag{4.3}$$

for large m. So for the sequence $(x_1, x_2, \ldots, x_m) \in \Gamma$, the probability that it belongs to typical set is very high.

Now, choose top 2^n probable sequences from $A_m^{(\epsilon)}$, create a uniform distribution on them and encode using n bits. This implies that bits are iid Ber(1/2). Ideal, we would like to have 2^n as close to $|A_m^{(\epsilon)}|$ as possible. But depending on the computational feasibility, we may have to compromise.

So, now, given a sequence of iid equally likely message bits, we can consider n at a time and sample from the new uniform distribution on 2^n sequences. Clearly, this encoding is one-one. Each element of the sequence can be encoded into optimal x^* and sent through the channel (as described in section 4.2). But while decoding we may have errors, and the received vectors may not be in the non-zero sample space of Gibbs-1. Even if they are, not every sequence of length m of vectors from Gibbs-1 is in the sampled uniform distribution. So we can use a nearest neighbor search to find valid vectors from Gibbs-1 and then the valid sequences from the uniform

distribution. The latter is computationally expensive and thus puts a limitation on n. Finally ,we can decode them into message bits. Since we also have an error control code in place, we may have to use interleaving to make the coded bits close to iid, but this is fine. In the next subsection, we give some numerical results.

4.3.2 Numerical results

For the simulation, we chose $[1, \epsilon]$ channel with $N = 9$, $\epsilon = 0.2$, $\delta = 0.3$. The value of other parameters and results are given below.

$$
\begin{aligned}
&k = 20,\ m = 5,\ n = 10 \\
&P = 0.08 \\
&\text{Entropy of Gibbs distribution} = 8.11 \\
&\text{Entropy of Gibbs-1} \approx 4 \\
&\text{Entropy of the transmitted vectors in } \{-1, 1\}^N = 3.35 \\
&\Gamma = 3.2 \times 10^6,\ \text{size of typical set} = 1049760 \\
&\text{SNR} = 22\text{db},\ R_{\text{out}} = 0.9,\ 1 - H_2(Q(\delta/\sigma)) = 0.9998 \\
&\text{For SNR=22db, BER} = 0 \text{ for more than } 10^6 \text{ message bits} \\
&\text{Hence, a rate of around } 0.4 \times C_{N,\delta}(P)(1 - H_2(Q(\delta/\sigma)) \text{ was achieved}
\end{aligned}
\tag{4.4}
$$

So, it can be seen that the huge loss in rate is mainly due to the loss in entropy in Gibbs-1. It remains to be seen if we can improve upon this method or come up with better methods to perform source simulation and decoding in the presence of noise.

For example, using the Gibbs sampler according to algorithm 1, a Viterbi algorithm could be used at the decoder assuming that the receiver knows what was the

length of the message bits that was used to sample each of the samples. Again, the issue is to use a suitable metric. Even when there is no noise, there can different sequences of message bits which are of same length that can give the same sample. So tackling the decoding part could be an interesting problem for future study.

CHAPTER 5

CONCLUDING REMARKS

The capacity of approximate ISI channel with average power constrained continuous input and 3-level quantized output was characterized using Gibbs distribution. Markovian achievable schemes that approach capacity were described. More general and better Markovian achievable schemes could be studied in future. Also, it would be interesting to generalize the output quantization to any arbitrary level.

Further, signaling methods for the actual ISI channel were studied and a possible achievable rate was proposed. A well-known sampler was described but it was observed that it wasn't one-one and hence could not be decoded. A typical set based method to sample from an approximate Gibbs distribution was proposed and decoding became possible; but there was a severe loss in information rate. It would interesting to come up with source simulation techniques for which decoding in the presence of noise is possible without a substantial loss in information rate. These could be topics for further study.

APPENDIX A

PROOF OF LEMMA 3.2.1

In this chapter, we provide a proof of lemma 3.2.1. This is inspired from the discussion in [adam W , http://math.stackexchange.com/users/43193/adam w]. The lemma is restated here for convenience:

Lemma A.0.2. *Let P be an $N \times N$ matrix such that P^{-1} is row-diagonally dominant (with $(P^{-1})_{ii} > 0, \forall i \in [N]$). Also $\forall J \subset [N]$, let the matrix obtained by deleting the rows and columns of P corresponding to J, denoted Q_J, be invertible. Then Q_J^{-1} is row diagonally dominant.*

Proof. It is enough to prove the lemma for $J \subset [N]$ which are singleton matrices because for any other J, we can construct Q_J by removing one row and corresponding column at a time, and applying the lemma repeatedly we can conclude that the result holds for any $J \subset [N]$

Without loss of generality, assume that $J = \{N\}$, i.e. Q_J is obtained by deleting the last row and last column of P. So

$$P = \begin{pmatrix} Q_J & b \\ c^T & d \end{pmatrix} \tag{A.1}$$

where b, c are column vectors of length $N-1$ and d is a scalar. Let P^{-1} be given by

$$P^{-1} = \begin{pmatrix} E & f \\ g^T & h \end{pmatrix} \tag{A.2}$$

where $f,\ g$ are column vectors of length $N-1$ and $h>0$ is a scalar. So we have

$$\begin{pmatrix} Q_J & b \\ c^T & d \end{pmatrix}^{-1} = \begin{pmatrix} E & f \\ g^T & h \end{pmatrix} \tag{A.3}$$

which implies

$$\begin{pmatrix} E & f \\ g^T & h \end{pmatrix}\begin{pmatrix} Q_J & b \\ c^T & d \end{pmatrix} = \begin{pmatrix} I_{N-1} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{A.4}$$

where I_{N-1} is the identity matrix of size $(N-1)\times(N-1)$. Now left multiplying equation (A.4) with $\begin{pmatrix} I_{N-1} & -f\frac{1}{h} \\ \mathbf{0}^T & 1 \end{pmatrix}$ we get

$$\begin{pmatrix} E-f\frac{1}{h}g^T & \mathbf{0} \\ g^T & h \end{pmatrix}\begin{pmatrix} Q_J & b \\ c^T & d \end{pmatrix} = \begin{pmatrix} I_{N-1} & -f\frac{1}{h} \\ \mathbf{0}^T & 1 \end{pmatrix}. \tag{A.5}$$

Thus we have

$$\left(E-f\frac{1}{h}g^T\right)Q_J = I_{N-1} \Longleftrightarrow Q_J^{-1} = E-f\frac{1}{h}g^T. \tag{A.6}$$

So $(Q_J^{-1})_{ij} = E_{ij} - \frac{f_i g_j}{h}$. Since P^{-1} is row-diagonally dominant, we have

$$\begin{aligned} &E_{ii} - \sum_{j=1, j\neq i}^{N-1} |E_{ij}| - |f_i| \geq 0, i \in \{1,2,\cdots,N-1\} \\ &h - \sum_{j=1}^{N-1} |g_j| \geq 0 \Longrightarrow \frac{1}{h}\sum_{j=1}^{N-1}|g_j| \leq 1. \end{aligned} \tag{A.7}$$

Therefore, for $i \in \{1, 2, \cdots, N-1\}$, we have

$$\begin{aligned}
&(Q_J^{-1})_{ii} - \sum_{j=1, j\neq i}^{N-1} |(Q_J^{-1})_{ij}| = E_{ii} - \frac{f_i g_i}{h} - \sum_{j=1, j\neq i}^{N-1} \left| E_{ij} - \frac{f_i g_j}{h} \right| \\
&\geq E_{ii} - \frac{|f_i g_i|}{h} - \sum_{j=1, j\neq i}^{N-1} |E_{ij}| - |f_i| \sum_{j=1, j\neq i}^{N-1} \frac{|g_j|}{h} \\
&= E_{ii} - \sum_{j=1, j\neq i}^{N-1} |E_{ij}| - |f_i| \sum_{j=1}^{N-1} \frac{|g_j|}{h} \\
&\geq E_{ii} - \sum_{j=1, j\neq i}^{N-1} |E_{ij}| - |f_i| \geq 0.
\end{aligned} \tag{A.8}$$

Hence Q_J^{-1} is row-diagonally dominant. □

APPENDIX B

OPTIMIZATION PROBLEM OF 3-LEVEL QUANTIZER

In this chapter we discuss the optimization problem (3.11) in generality and give proof of lemma 3.2.2.

B.1 The optimization problem

Let x and y denote the input and output vectors of length N respectively for the channel h with $N \times N$ matrix M_h. Let $s \in \{-1, 0, 1\}^N$ denote the quantized output. Let $\{+d, -d\}$ be the decision points. Let $s_1, s_2 \in \{-1, 1\}^N$ be 2-level quantizers defined according to (3.12). Given $s \in \{-1, 0, 1\}^N$ and $0 < \delta < d$, the aim is to solve the optimization problem (3.11) which is restated here for convenience:

$$\varepsilon(s) = \min_{\substack{x \\ \text{diag}(s_1)(M_h x - d\mathbf{1}) \geq \delta\mathbf{1} \\ \text{diag}(s_2)(M_h x + d\mathbf{1}) \geq \delta\mathbf{1}}} \left(\|x\|^2 \right) \tag{B.1}$$

where $\mathbf{1}$ is the all-1 column vector of length N.

B.2 The dual problem and the solution

The Lagrangian function of the optimization problem is

$$L(x, \lambda_1, \lambda_2) = \|x\|^2 + \lambda_1^T \left(\delta\mathbf{1} - \text{diag}(s_1)(M_h x - d\mathbf{1}) \right) + \lambda_2^T \left(\delta\mathbf{1} - \text{diag}(s_2)(M_h x + d\mathbf{1}) \right) \tag{B.2}$$

where $\lambda_1, \lambda_2 \in \mathcal{R}^N$ are the Lagrange multipliers.

Since the primal problem is convex with linear constraints, strong duality holds. So, if x^* is primal optimal and $(\lambda_1^*, \lambda_2^*)$ is dual optimal then $\inf_x L(x, \lambda_1^*, \lambda_2^*) = L(x^*, \lambda_1^*, \lambda_2^*)$, and this x^* can be found by setting the gradient of the Lagrangian function with respect to x to zero; we get $x^* = \frac{1}{2}\Big(M_h^T \text{diag}(s_1)\lambda_1^* + M_h^T \text{diag}(s_2)\lambda_2^*\Big)$. Also by complementary slackness we have, for $i \in [N]$, $\lambda_1^*(i)\Big(\delta\mathbf{1} - \text{diag}(s_1)(M_h x^* - d\mathbf{1})\Big)(i) = 0$ and $\lambda_2^*(i)\Big(\delta\mathbf{1} - \text{diag}(s_2)(M_h x^* + d\mathbf{1})\Big)(i) = 0$. Now it can be shown that $\lambda_1^*(i)\lambda_2^*(i) = 0$ since the corresponding constraints cannot be simultaneously active (as long as $d > \delta$).

Now it is to be found for which $i \in [N]$ the multipliers have to be zero. Suppose for some i we have $s(i) = 1$ then the second constraint becomes strict inequality and hence, by complimentary slackness, $\lambda_2^*(i) = 0$. Similarly, if $s(i) = -1$ then $\lambda_1^*(i) = 0$. If $s(i) = 0$ then, intuitively, the minimum energy input x would be such that $y(i) = 0$. So we set $\lambda_1^*(i)\lambda_2^*(i) = 0$ but we do not yet know where each of these multipliers are non-zero.

Let $J_0 = \{i \in [N] : s(i) = 0\}$, $J_1 = \{i \in [N] : s(i) = 1\}$ and $J_{-1} = \{i \in [N] : s(i) = -1\}$. So if $i \in J_1$ then $\lambda_2^*(i) = 0$, and if $i \in J_{-1}$ then $\lambda_1^*(i) = 0$. We now need to find the optimal values of remaining multipliers. The Lagrangian dual is

$$\begin{aligned} L(\lambda_1, \lambda_2) =& \delta(\lambda_1^T + \lambda_2^T)\mathbf{1} + d(\lambda_1^T s_1 - \lambda_2^T s_2) \\ & - \frac{1}{4}\Big(\lambda_1^T \text{diag}(s_1) M_h M_h^T \text{diag}(s_1)\lambda_1 + \lambda_1^T \text{diag}(s_1) M_h M_h^T \text{diag}(s_2)\lambda_2 \\ & + \lambda_2^T \text{diag}(s_2) M_h M_h^T \text{diag}(s_1)\lambda_1 + \lambda_2^T \text{diag}(s_2) M_h M_h^T \text{diag}(s_2)\lambda_2\Big) \end{aligned} \tag{B.3}$$

We must also know the indices in J_0 where the multipliers are non zero. Let us assume that we know this OR let us choose these indices arbitrarily. Let

$J_0^a = \{i \in J_0 : \lambda_1(i) \neq 0\}$, $J_0^b = \{i \in J_0 : \lambda_2(i) = 0 = \lambda_1(i)\}$ and $J_0^c = \{i \in J_0 : \lambda_2(i) \neq 0\}$. Also, let $J_{m,n} = J_m \cup J_n,\ m, n \in \{-1, 0, 1\}$ and $J_{0,n}^x = J_0^x \cup J_n,\ n \in \{-1, 0, 1\},\ x \in \{a, b, c\}$. With this notation, we get

$$\begin{aligned} L(\lambda_1, \lambda_2) =& \delta(\lambda_1^T)^{J_{0,1}^a}\mathbf{1}^{J_{0,1}^a} + \delta(\lambda_2^T)^{J_{0,-1}^c}\mathbf{1}^{J_{0,-1}^c} + d((\lambda_1^T)^{J_{0,1}^a}s_1^{J_{0,1}^a} - (\lambda_2^T)^{J_{0,-1}^c}s_2^{J_{0,-1}^c}) \\ & - \frac{1}{4}\Big((\lambda_1^T)^{J_{0,1}^a}A^{J_{0,1}^a J_{0,1}^a}\lambda_1^{J_{0,1}^a} + (\lambda_1^T)^{J_{0,1}^a}B^{J_{0,1}^a J_{0,-1}^c}\lambda_2^{J_{0,-1}^c} \\ & + (\lambda_2^T)^{J_{0,-1}^c}C^{J_{0,-1}^c J_{0,1}^a}\lambda_1^{J_{0,1}^a} + (\lambda_2^T)^{J_{0,-1}^c}D^{J_{0,-1}^c J_{0,-1}^c}\lambda_2^{J_{0,-1}^c}\Big) \end{aligned} \tag{B.4}$$

where the superscripts denote the sub-matrix corresponding to the rows and/or columns represented by them. It can also be seen that $A^{J_{0,1}^a J_{0,1}^a} = \text{diag}(s_1)^{J_{0,1}^a J_{0,1}^a}(M_h M_h^T)^{J_{0,1}^a J_{0,1}^a}\text{diag}(s_1)^{J_{0,1}^a J_{0,1}^a}$ since $\text{diag}(s_1)$ is a diagonal matrix; similarly for B, C and D.

Setting the gradient of the Lagrangian dual with respect to $\lambda_1^{J_{0,1}^a}$ and $\lambda_2^{J_{0,-1}^c}$ to zero to find the optimal values of the multipliers, we get

$$\begin{aligned} &\begin{pmatrix} \text{diag}(s_1)^{J_{0,1}^a}\, P^{J_{0,1}^a J_{0,1}^a}\, \text{diag}(s_1)^{J_{0,1}^a} & \text{diag}(s_1)^{J_{0,1}^a}\, P^{J_{0,1}^a J_{0,-1}^c}\, \text{diag}(s_2)^{J_{0,-1}^c} \\ \text{diag}(s_2)^{J_{0,-1}^c}\, P^{J_{0,-1}^c J_{0,1}^a}\, \text{diag}(s_1)^{J_{0,1}^a} & \text{diag}(s_2)^{J_{0,-1}^c}\, P^{J_{0,-1}^c J_{0,-1}^c}\, \text{diag}(s_2)^{J_{0,-1}^c} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_{0,1}^a} \\ \lambda_2^{J_{0,-1}^c} \end{pmatrix} \\ &= 2\begin{pmatrix} \delta\mathbf{1}^{J_{0,1}^a} + d\mathbf{s_1}^{J_{0,1}^a} \\ \delta\mathbf{1}^{J_{0,-1}^c} - d\mathbf{s_2}^{J_{0,-1}^c} \end{pmatrix} \end{aligned} \tag{B.5}$$

where $P = (M_h M_h^T)$. This can also be written as

$$\begin{aligned} &\begin{pmatrix} P^{J_{0,1}^a J_{0,1}^a}\, \text{diag}(s_1)^{J_{0,1}^a} & P^{J_{0,1}^a J_{0,-1}^c}\, \text{diag}(s_2)^{J_{0,-1}^c} \\ P^{J_{0,-1}^c J_{0,1}^a}\, \text{diag}(s_1)^{J_{0,1}^a} & P^{J_{0,-1}^c J_{0,-1}^c}\, \text{diag}(s_2)^{J_{0,-1}^c} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_{0,1}^a} \\ \lambda_2^{J_{0,-1}^c} \end{pmatrix} \\ &= 2\begin{pmatrix} \delta\mathbf{s_1}^{J_{0,1}^a} + d\mathbf{1}^{J_{0,1}^a} \\ \delta\mathbf{s_2}^{J_{0,-1}^c} - d\mathbf{1}^{J_{0,-1}^c} \end{pmatrix} \end{aligned} \tag{B.6}$$

Using the fact that $[i \in J_1 \implies s_1(i) = 1]$, $[i \in J_0^a \implies s_1(i) = -1]$, $[i \in J_0^c \implies$

$s_2(i) = 1]$ and $[i \in J_{-1} \implies s_2(i) = -1]$, the above equation can be expanded further and we get

$$\begin{pmatrix} -P^{J_0^a J_0^a} & P^{J_0^a J_1} & P^{J_0^a J_0^c} & -P^{J_0^a J_{-1}} \\ -P^{J_1 J_0^a} & P^{J_1 J_1} & P^{J_1 J_0^c} & -P^{J_1 J_{-1}} \\ -P^{J_0^c J_0^a} & P^{J_0^c J_1} & P^{J_0^c J_0^c} & -P^{J_0^c J_{-1}} \\ -P^{J_{-1} J_0^a} & P^{J_{-1} J_1} & P^{J_{-1} J_0^c} & -P^{J_{-1} J_{-1}} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_0^a} \\ \lambda_1^{J_1} \\ \lambda_2^{J_0^c} \\ \lambda_2^{J_{-1}} \end{pmatrix} = 2 \begin{pmatrix} (-\delta + d)\mathbf{1}^{J_0^a} \\ (\delta + d)\mathbf{1}^{J_1} \\ (\delta - d)\mathbf{1}^{J_0^c} \\ (-\delta - d)\mathbf{1}^{J_{-1}} \end{pmatrix} \tag{B.7}$$

Assuming that the matrix in the left hand side of the above equation to be invertible, we have the optimal values of the multipliers unless these values are negative. We need to find the conditions on the matrix on the left hand side such that the solutions we get by inversion are non-negative for any given output state s. In the next section, we will show that diagonal dominance of P^{-1} will ensure the non-negativity of $\lambda_1^{J_1}$ and $\lambda_2^{J_{-1}}$, and also provide a proof of lemma 3.2.2. Later, in section B.4, we give an algorithm that gives us the sets J_0^a and J_0^c so that the multipliers $\lambda_1^{J_0^a}$ and $\lambda_2^{J_0^c}$ are non-negative.

B.3 Conditions for non-negativity

Let $Z = \begin{pmatrix} -P^{J_0^a J_0^a} & P^{J_0^a J_1} & P^{J_0^a J_0^c} & -P^{J_0^a J_{-1}} \\ -P^{J_1 J_0^a} & P^{J_1 J_1} & P^{J_1 J_0^c} & -P^{J_1 J_{-1}} \\ -P^{J_0^c J_0^a} & P^{J_0^c J_1} & P^{J_0^c J_0^c} & -P^{J_0^c J_{-1}} \\ -P^{J_{-1} J_0^a} & P^{J_{-1} J_1} & P^{J_{-1} J_0^c} & -P^{J_{-1} J_{-1}} \end{pmatrix}$ and let Q be the matrix obtained from P by removing the rows and columns corresponding to J_0^b. Let $\lambda = \begin{pmatrix} \lambda_1^{J_0^a} \\ \lambda_1^{J_1} \\ \lambda_2^{J_0^c} \\ \lambda_2^{J_{-1}} \end{pmatrix}$, $W = 2\begin{pmatrix} (-\delta + d)\mathbf{1}^{J_0^a} \\ (\delta + d)\mathbf{1}^{J_1} \\ (\delta - d)\mathbf{1}^{J_0^c} \\ (-\delta - d)\mathbf{1}^{J_{-1}} \end{pmatrix}$ and $N = \begin{pmatrix} -I_{J_0^a} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{J_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{J_0^c} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -I_{J_{-1}} \end{pmatrix}$ where $I_{J_0^a}$, $I_{J_0^c}$ I_{J_1} and$I_{J_{-1}}$ are identity matrices of dimensions $|J_0^a|$, $|J_0^c|$ $|J_1|$ and$|J_{-1}|$ respectively. The matrices Z, W and N are dependent on the output state s through the sets J_0, J_1 and J_{-1}.

We can write Z as $Z = UQU^T N$ where U is a permutation matrix (and hence unitary). So the equation (B.7) becomes

$$UQU^T N\lambda = W \tag{B.8}$$

So, if Q is invertible then

$$\lambda = NUQ^{-1}U^T W \tag{B.9}$$

It can be seen that $UQ^{-1}U^T = \begin{pmatrix} (Q^{-1})^{J_0^a J_0^a} & (Q^{-1})^{J_0^a J_1} & (Q^{-1})^{J_0^a J_0^c} & (Q^{-1})^{J_0^a J_{-1}} \\ (Q^{-1})^{J_1 J_0^a} & (Q^{-1})^{J_1 J_1} & (Q^{-1})^{J_1 J_0^c} & (Q^{-1})^{J_1 J_{-1}} \\ (Q^{-1})^{J_0^c J_0^a} & (Q^{-1})^{J_0^c J_1} & (Q^{-1})^{J_0^c J_0^c} & (Q^{-1})^{J_0^c J_{-1}} \\ (Q^{-1})^{J_{-1} J_0^a} & (Q^{-1})^{J_{-1} J_1} & (Q^{-1})^{J_{-1} J_0^c} & (Q^{-1})^{J_{-1} J_{-1}} \end{pmatrix}$.

So if $Q^{-1} = (t_{ij})$,

$$\begin{aligned} i \in J_1 &\implies \lambda_1(i) = 2\Big[\delta \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^a \cup J_{-1}]} + d \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^c \cup J_{-1}]}\Big] \\ i \in J_{-1} &\implies \lambda_2(i) = -2\Big[\delta \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^a \cup J_{-1}]} + d \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^c \cup J_{-1}]}\Big] \end{aligned} \tag{B.10}$$

where $\mathbb{1}[x \in A]$ is 1 if $x \in A$ and 0 otherwise .

We want the multipliers to be non-negative for any J_0, J_1 and J_{-1}. Suppose $i \in J_1$. We want $\lambda_1(i) \geq 0$. It is sufficient if $\sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^a \cup J_{-1}]} \geq 0$ and $\sum_{j=1}^N t_{ij}(-1)^{\mathbb{1}[j \in J_0^c \cup J_{-1}]} \geq 0$. Similarly, for $i \in J_{-1}$, it is sufficient if $\sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^a \cup J_{-1}]} \leq 0$ and $\sum_{j=1}^N t_{ij}(-1)^{\mathbb{1}[j \in J_0^c \cup J_{-1}]} \leq 0$. So if Q^{-1} is diagonally dominant (with diagonal entries being non-negative) then the above sufficient conditions are satisfied and hence $\lambda_1^{J_1}$ and $\lambda_2^{J_{-1}}$ are non-negative. So if P satisfies the hypothesis of lemma 3.2.1 then Q^{-1} is diagonally dominant and hence the optimal multipliers in (B.10) are non-negative.

We also have,

$$\begin{aligned} i \in J_0^a &\implies \lambda_1(i) = 2\Big[\delta \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_1 \cup J_0^c]} - d \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^c \cup J_{-1}]}\Big] \\ i \in J_0^c &\implies \lambda_2(i) = 2\Big[\delta \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^a \cup J_{-1}]} - d \sum_{j \in [N] \setminus J_0^b} t_{ij}(-1)^{\mathbb{1}[j \in J_0^a \cup J_1]}\Big] \end{aligned} \tag{B.11}$$

Again assuming that P satisfies the hypothesis of lemma 3.2.1, Q^{-1} is diagonally dominant and hence the coefficients of both d and δ are non-negative. So for a

given channel for which the matrix $P = M_h M_h^T$ satisfies the hypothesis of lemma 3.2.1 (and hence Q^{-1} is diagonally dominant for every s, J_0^a and J_0^c), given output sequence s and a choice of J_0^a, J_0^c, as d increases, the optimal value of λ_1 and λ_2, corresponding to J_0^a and J_0^c respectively, as calculated from equations (B.11), decrease monotonically. So, given such a channel, there exists a $d_0 > \delta > 0$ such that for any $d > d_0$, for any output sequence s and any choice of J_0^a, J_0^c, the value of multipliers as calculated from (B.11) are negative. This implies that the only valid choice is $J_0^b = J_0$ and $J_0^a = J_0^c = \emptyset$. Thus we have the proof of lemma 3.2.2.

Further, using (B.5) and $x^* = \frac{1}{2}\left(M_h^T \text{diag}(s_1)\lambda_1^* + M_h^T \text{diag}(s_2)\lambda_2^*\right)$ we get the following under-determined system of equations for x^*:

$$\begin{aligned} \text{diag}(s_1) M_h^{J_{0,1}^a[]} x^* &= \delta \mathbf{1}^{J_{0,1}^a} + d s_1^{J_{0,1}^a} \\ \text{diag}(s_2) M_h^{J_{0,-1}^c[]} x^* &= \delta \mathbf{1}^{J_{0,-1}^c} - d s_2^{J_{0,-1}^c} \end{aligned} \tag{B.12}$$

where $M_h^{J[]}$ denotes the sub-matrix corresponding to rows with indices J and all columns. This can re-written as

$$\begin{pmatrix} M_h^{J_{0,1}^a[]} \\ M_h^{J_{0,-1}^c[]} \end{pmatrix} x^* = \begin{pmatrix} \delta s_1^{J_{0,1}^a} + d\mathbf{1}^{J_{0,1}^a} \\ \delta s_2^{J_{0,-1}^c} - d\mathbf{1}^{J_{0,-1}^c} \end{pmatrix}. \tag{B.13}$$

It is well known that for an under-determined system $Ax = b$ with (AA^T) invertible, the least norm solution is given by $x = A^T(AA^T)^{-1}b$. Letting $\bar{H} = \begin{pmatrix} M_h^{J_{0,1}^a[]} \\ M_h^{J_{0,-1}^c[]} \end{pmatrix}$ and $\bar{s} = \begin{pmatrix} \delta s_1^{J_{0,1}^a} + d\mathbf{1}^{J_{0,1}^a} \\ \delta s_2^{J_{0,-1}^c} - d\mathbf{1}^{J_{0,-1}^c} \end{pmatrix}$, we have

$$x^* = \bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{s}. \tag{B.14}$$

Note that $\bar{H}\bar{H}^T = UQU^T$ and from equations (B.9) and (B.6) we have

$$\begin{pmatrix} (\lambda_1^*)^{J_{0,1}^a} \\ (\lambda_2^*)^{J_{0,-1}^c} \end{pmatrix} = 2 \begin{pmatrix} \text{diag}(s_1)^{J_{0,1}^a} & \mathbf{0} \\ \mathbf{0} & \text{diag}(s_2)^{J_{0,-1}^c} \end{pmatrix} UQ^{-1}U^T\bar{s}. \tag{B.15}$$

So clearly $\bar{H}^T(\bar{H}\bar{H}^T)^{-1}\bar{s} = \frac{1}{2}\Big(M_h^T\text{diag}(s_1)\lambda_1^* + M_h^T\text{diag}(s_2)\lambda_2^*\Big)$.

Therefore the solution to the optimization problem (B.1) is the least norm solution to (B.13) which is given by (B.14). Hence the optimal energy is given by

$$\varepsilon(s) = \bar{s}^T(\bar{H}\bar{H}^T)^{-1}\bar{s}. \tag{B.16}$$

B.4 Finding the sets J_0^a and J_0^c

We now give an algorithm to find the sets J_0^a and J_0^c (and $J_0^b = J_0 \setminus (J_0^a \cup J_0^c)$). First, we make some observations. In this section d represents the generic decision point parameter of the quantizer.

Assume that for a given channel matrix M_h, output sequence s and the quantizer decision points $[+d_1, -d_1]$ (i.e. $d = d_1$), we know the the sets J_0^a and J_0^c such that the multipliers corresponding to them as calculated from equation (B.11) are non-negative. When we increase d from d_1, beyond some point say d_2, atleast one of these multipliers may become negative and hence the previous choice of J_0^a and J_0^c are no longer correct. Now for $d \in [d_1, d_2)$ the choices of J_0^a and J_0^c are valid and the corresponding multipliers decrease monotonically as d is increased from d_1 to d_2.

Let $\epsilon > 0$ be small. For $d \in [d_2, d_2 + \epsilon]$, let $\bar{J}_0^a$ and $\bar{J}_0^c$ be a valid choice of the subsets in J_0 where the $\lambda_1(i) \neq 0, i \in \bar{J}_0^a$ and $\lambda_2(i) \neq 0, i \in \bar{J}_0^c$ and ϵ is small enough such that these sets are valid throughout the interval $[d_2, d_2 + \epsilon]$. We claim that

$\bar{J}_0^a \subset J_0^a$ and $\bar{J}_0^c \subset J_0^c$. Although it seems difficult to give a rigorous proof, we provide an explanation for the same and the claims is supported by numerical simulations that we have performed.

For simplicity, let us assume that $\lambda_1(i_1), i_1 \in J_0^a$ is the only multiplier that becomes zero when d is increased from d_1 to d_2. For $d < d_2$ we use J_0^a and J_0^c in equation (B.7) (call this case 1), and for $d \geq d_2$ we use $\bar{J}_0^a$ and $\bar{J}_0^c$ in equation (B.7) (call this case 2). In case 1, as d becomes close to d_2, $\lambda_1(i_1) \approx 0$. Hence the first column of the coefficient matrix in the left hand side of (B.7) can be neglected, and therefore the rows in the left hand side (representing the equations), except for the row corresponding to i_1, are almost equal to the corresponding rows in case 2. Thus by removing the row and column corresponding to $i_1 \in J_0^a$ from case 1 we get equation (B.7) for case 2. Hence $\bar{J}_0^a = J_0^a \setminus i_1$. So, $\bar{J}_0^a \subset J_0^a$. Similarly $\bar{J}_0^c \subset J_0^c$.

Thus if we know valid J_0^a and J_0^c for some decision point parameter d_1, we can find the corresponding sets $\bar{J}_0^a$ and $\bar{J}_0^c$ for any $d > d_1$ as follows:

Algorithm 2 Algorithm to find $\bar{J}_0^a$ and $\bar{J}_0^c$ given J_0^a, J_0^c and d

function F(J_0^a, J_0^b, d, s, M_h)
    $\bar{J}_0^a \leftarrow J_0^a$
    $\bar{J}_0^c \leftarrow J_0^c$
loop:
    Solve equations (B.11)
    if $\exists i \in \bar{J}_0^a$ such that $\lambda_1(i) < 0$ **or** $\exists j \in \bar{J}_0^c$ such that $\lambda_2(j) < 0$ **then**
        $A_1 = \{i \in \bar{J}_0^a : \lambda_1(i) < 0\}$
        $A_2 = \{i \in \bar{J}_0^c : \lambda_2(i) < 0\}$
        $\bar{J}_0^a \leftarrow \bar{J}_0^a \setminus A_1$
        $\bar{J}_0^c \leftarrow \bar{J}_0^c \setminus A_2$
        goto *loop*
        close;
    return $\bar{J}_0^a$, $\bar{J}_0^c$

So if the sets J_0^a and J_0^c are known for the case $d = \delta$, then for any other value of the d, the above algorithm can be used to find these sets.

B.5 The case $d = \delta$

In the case of $d = \delta$, there is no need to decompose the set J_0 and we can find J_0^a and J_0^c. But, given s, for $i \in J_0$, both optimal multipliers can possibly be non-zero since $d-\delta = -\delta+d$. Let $\epsilon > 0$ be small. For $d = \delta+\epsilon$, we know that the $\lambda_1(i)\lambda_2(i) = 0, i \in J_0$. We can choose ϵ small enough that the sets J_0^a and J_0^c remain constant for every $d \in (\delta, \delta + \epsilon]$. The corresponding multipliers satisfy equation (B.11) and hence are continuous functions of d in this interval. Hence, we can impose the condition $\lambda_1(i)\lambda_2(i) = 0$ for the case of $d = \delta$ as well.

In this case, equation (B.4) becomes

$$\begin{aligned} L(\lambda_1, \lambda_2) =& \delta(\lambda_1^T)^{J_{0,1}}\mathbf{1}^{J_{0,1}} + \delta(\lambda_2^T)^{J_{0,-1}}\mathbf{1}^{J_{0,-1}} + d((\lambda_1^T)^{J_{0,1}}s_1^{J_{0,1}} - (\lambda_2^T)^{J_{0,-1}}s_2^{J_{0,-1}}) \\ & - \frac{1}{4}\Big((\lambda_1^T)^{J_{0,1}}A^{J_{0,1}J_{0,1}}\lambda_1^{J_{0,1}} + (\lambda_1^T)^{J_{0,1}}B^{J_{0,1}J_{0,-1}}\lambda_2^{J_{0,-1}} \\ & + (\lambda_2^T)^{J_{0,-1}}C^{J_{0,-1}J_{0,1}}\lambda_1^{J_{0,1}} + (\lambda_2^T)^{J_{0,-1}}D^{J_{0,-1}J_{0,-1}}\lambda_2^{J_{0,-1}}\Big) \end{aligned} \tag{B.17}$$

Setting the gradient of the Lagrangian with respect to $\lambda_1^{J_{0,1}}$ and $\lambda_2^{J_{0,-1}}$ to zero, we get

$$\begin{aligned} \begin{pmatrix} -P^{J_0J_0} & P^{J_0J_1} & P^{J_0J_0} & -P^{J_0J_{-1}} \\ -P^{J_1J_0} & P^{J_1J_1} & P^{J_1J_0} & -P^{J_1J_{-1}} \\ -P^{J_0J_0} & P^{J_0J_1} & P^{J_0J_0} & -P^{J_0J_{-1}} \\ -P^{J_{-1}J_0} & P^{J_{-1}J_1} & P^{J_{-1}J_0} & -P^{J_{-1}J_{-1}} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_0} \\ \lambda_1^{J_1} \\ \lambda_2^{J_0} \\ \lambda_2^{J_{-1}} \end{pmatrix} &= 2 \begin{pmatrix} (-\delta + d)\mathbf{1}^{J_0} \\ (\delta + d)\mathbf{1}^{J_1} \\ (\delta - d)\mathbf{1}^{J_0} \\ (-\delta - d)\mathbf{1}^{J_{-1}} \end{pmatrix} \\ &= 4 \begin{pmatrix} \mathbf{0}^{J_0} \\ (\delta)\mathbf{1}^{J_1} \\ \mathbf{0}^{J_0} \\ (-\delta)\mathbf{1}^{J_{-1}} \end{pmatrix} \end{aligned} \tag{B.18}$$

Simplifying, we get

$$\begin{pmatrix} -P^{J_0J_0} & P^{J_0J_1} & -P^{J_0J_{-1}} \\ -P^{J_1J_0} & P^{J_1J_1} & -P^{J_1J_{-1}} \\ -P^{J_{-1}J_0} & P^{J_{-1}J_1} & -P^{J_{-1}J_{-1}} \end{pmatrix} \begin{pmatrix} \lambda_1^{J_0} - \lambda_2^{J_0} \\ \lambda_1^{J_1} \\ \lambda_2^{J_{-1}} \end{pmatrix} = 4 \begin{pmatrix} \mathbf{0}^{J_0} \\ (\delta)\mathbf{1}^{J_1} \\ -(\delta)\mathbf{1}^{J_{-1}} \end{pmatrix} \tag{B.19}$$

If P^{-1} is diagonally dominant then it can be seen that the multipliers corresponding to the indices J_1 and J_{-1} are non-negative by arguments similar to those used in section B.3. By solving equation (B.19), for each $i \in J_0$ we know the value of $\lambda_1(i) - \lambda_2(i)$ and we also know that $\lambda_1(i)\lambda_2(i) = 0$. So we have

$$\begin{aligned} &\lambda_1(i) - \lambda_2(i) = 0 \text{ then choose } \lambda_1(i) = 0 = \lambda_2(i) \\ &\lambda_1(i) - \lambda_2(i) > 0 \text{ then choose } \lambda_2(i) = 0 \ \& \ \lambda_1(i) > 0 \\ &\lambda_1(i) - \lambda_2(i) < 0 \text{ then choose } \lambda_1(i) = 0 \ \& \ \lambda_2(i) > 0 \end{aligned} \tag{B.20}$$

Thus the optimal multipliers are non-negative and we know the sets J_0^a, J_0^b and J_0^c for the case of $d = \delta$.

APPENDIX C

THE $(1, \epsilon)$ CHANNEL

Consider the $(1, \epsilon)$ channel where $|\epsilon| < 1$. For this channel with $N \geq 3$, $P = (M_h M_h^T)$ is given by

$$P = \begin{pmatrix} 1+\epsilon^2 & \epsilon & 0 & 0 & \cdots & \epsilon \\ \epsilon & 1+\epsilon^2 & \epsilon & 0 & \cdots & 0 \\ 0 & \epsilon & 1+\epsilon^2 & \epsilon & \cdots & 0 \\ 0 & 0 & \epsilon & 1+\epsilon^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \epsilon & 1+\epsilon^2 & \epsilon \\ \epsilon & 0 & 0 & \cdots & \epsilon & 1+\epsilon^2 \end{pmatrix} \tag{C.1}$$

This channel is diagonally dominant if $\epsilon < \sqrt{\frac{3}{2}} - 1$. To see this, P can be written as $P = I + \varepsilon$ where I is the identity matrix of dimension N. If $\varepsilon = (\epsilon_{i,j})$, then

$$\|\varepsilon\|_\infty = \max_{1 \leq i \leq N} \sum_{j=1}^{N} |\epsilon_{ij}| = \epsilon^2 + 2|\epsilon| \tag{C.2}$$

Let $\|\varepsilon\|_\infty < \frac{1}{2} \Leftrightarrow |\epsilon| < \sqrt{\frac{3}{2}} - 1$. Then $P^{-1} = I - \varepsilon + \varepsilon^2 - \varepsilon^3 + ...$ (i.e., the Neumann series converges). We know $\|\varepsilon^k\|_\infty \leq \|\varepsilon\|_\infty^k < \frac{1}{2^k}$. Let $S = -\varepsilon + \varepsilon^2 - \varepsilon^3 + ... = (s_{ij})$. Then for any $i \in [N]$, $1 + s_{ii} - \sum_{j,j \neq i} |s_{ij}| \geq 1 - \sum_{j=1}^{N} |s_{ij}| > 1 - \sum_{k=1}^{\infty} \frac{1}{2^k} > 0$. Hence P^{-1} is diagonally dominant.

We can also prove, independent of lemma 3.2.1, that this channel is strongly diagonally dominant. For any set $J \subset [N]$, if the rows and columns of P corre-

sponding to the indices in J are removed we get a new matrix Q which still has the diagonal entries as $1 + \epsilon^2$. If Q is written as $Q = I + \varepsilon$ then it can be seen that $\|\varepsilon\|_\infty \leq \epsilon^2 + 2|\epsilon|$. Therefore

$$|\epsilon| < \sqrt{\frac{3}{2}} - 1 \implies \|\varepsilon\|_\infty < \frac{1}{2} \tag{C.3}$$

Hence, by a similar argument as above, Q^{-1} is diagonally dominant. Therefore this channel is strongly diagonally dominant for $|\epsilon| < \sqrt{\frac{3}{2}} - 1$.

REFERENCES

adam W (http://math.stackexchange.com/users/43193/adam w) (). Are there any decompositions of a symmetric matrix that allow for the inversion of any submatrix? Mathematics Stack Exchange. URL `http://math.stackexchange.com/q/208021`. URL:http://math.stackexchange.com/q/208021 (version: 2013-12-18).

Alkhateeb, A., **J. Mo**, **N. González-Prelcic**, and **R. W. Heath** (2014). Mimo precoding and combining solutions for millimeter-wave systems. *Communications Magazine, IEEE*, **52**(12), 122–131.

Conrad, K. (2013). Probability distributions and maximum entropy. *retrieved November*, **14**, 2013.

Ganti, R. K., **A. Thangaraj**, and **A. Mondal**, Approximation of capacity for isi channels with one-bit output quantization. *In Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015.

Gray, R. M., *Toeplitz and circulant matrices: A review*. now publishers inc, 2006.

Harwood, M., **N. Warke**, **R. Simpson**, **T. Leslie**, **A. Amerasekera**, **S. Batty**, **D. Colman**, **E. Carr**, **V. Gopinathan**, **S. Hubbins**, *et al.*, A 12.5 gb/s serdes in 65nm cmos using a baud-rate adc with digital receiver equalization and clock recovery. *In Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*. IEEE, 2007.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, **106**(4), 620.

Mo, J. and **R. W. Heath**, High snr capacity of millimeter wave mimo systems with one-bit quantization. *In Information Theory and Applications Workshop (ITA), 2014*. IEEE, 2014.

Romik, D. (1999). Sharp entropy bounds for discrete statistical simulation. *Statistics & probability letters*, **42**(3), 219–227.

Sadeghi, P., **P. O. Vontobel**, and **R. Shams** (2009). Optimization of information rate upper and lower bounds for channels with memory. *Information Theory, IEEE Transactions on*, **55**(2), 663–688.

Shamai, S. and **R. Laroia** (1996). The intersymbol interference channel: Lower bounds on capacity and channel precoding loss. *Information Theory, IEEE Transactions on*, **42**(5), 1388–1404.

Singh, J., **O. Dabeer**, and **U. Madhow** (2009). On the limits of communication with low-precision analog-to-digital conversion at the receiver. *Communications, IEEE Transactions on*, **57**(12), 3629–3639.

Sun, S., **T. S. Rappaport**, **R. W. Heath**, **A. Nix**, and **S. Rangan** (2014). Mimo for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both? *Communications Magazine, IEEE*, **52**(12), 110–121.

Zeitler, G., **A. C. Singer**, and **G. Kramer** (2012). Low-precision a/d conversion for maximum information rate in channels with memory. *Communications, IEEE Transactions on*, **60**(9), 2511–2521.