UNIVERSITY OF CALIFORNIA
Santa Barbara

Communication Transceiver Design with Low-Precision Analog-to-Digital Conversion

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Jaspreet Singh

Committee in Charge:

Professor Upamanyu Madhow, Chair

Professor Shiv Chandrasekaran

Professor Jerry Gibson

Professor Joao Hespanha

Professor Kenneth Rose

December 2009

The Dissertation of
Jaspreet Singh is approved:

Professor Shiv Chandrasekaran

Professor Jerry Gibson

Professor Joao Hespanha

Professor Kenneth Rose

Professor Upamanyu Madhow, Committee Chairperson

November 2009

Communication Transceiver Design with Low-Precision Analog-to-Digital Conversion

Copyright © 2009

by

Jaspreet Singh

To my parents and my brother

Curriculum Vitæ

Jaspreet Singh

EDUCATION

Dec 2009	*Ph.D.* Electrical and Computer Engineering University of California, Santa Barbara
Sep 2009	*M.A.* Statistics University of California, Santa Barbara
Dec 2005	*M.S.* Electrical and Computer Engineering University of California, Santa Barbara
May 2004	*B.Tech.* Electrical Engineering Indian Institute of Technology, Delhi

PUBLICATIONS

Journal Articles

- J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital at the receiver", *IEEE Transactions on Communications, vol. 57, no. 12, December 2009.*
- J. Singh, R. Kumar, U. Madhow, S. Suri, and R. E. Cagley, "Multiple target tracking with binary proximity sensors", *submitted to ACM Transactions on Sensor Networks.*

Conference Proceedings

- J. Singh, P. Sandeep, and U. Madhow, "Multi-gigabit communication: the ADC bottleneck", *(invited paper), Proc. IEEE Intl. Conf. on UltraWideband, Vancouver, Canada, September 2009.*
- J. Singh and U. Madhow, "On block noncoherent communication with low-precision phase quantization at the receiver", *Proc. IEEE Intl. Symp. on Information Theory (ISIT'09), Seoul, Korea, July 2009.*
- J. Singh, A. Saxena, K. Rose, and U. Madhow, "Optimization of correlated source coding for event-based monitoring in sensor networks", *Proc. IEEE Data Compression Conf. (DCC'09), Snowbird, USA, March 2009.*
- P. Sandeep, J. Singh, and U. Madhow, "Signal processing for multi-gigabit communication", *(invited paper), Proc. Information Theory and Applications Workshop (ITA'09), San Diego, USA, February 2009.*

- J. Singh, O. Dabeer and U. Madhow, "Capacity of the discrete-time AWGN channel under output quantization", *Proc. IEEE Intl. Symp. on Information Theory (ISIT'08), Toronto, Canada, July 2008.*

- J. Singh, U. Madhow, R. Kumar, S. Suri, and R. Cagley, "Tracking multiple targets using binary proximity sensors", *Proc. ACM/IEEE Intl. Conf. on Information Processing in Sensor Networks (IPSN'07), Cambridge, USA, April 2007.*

- J. Singh, O. Dabeer, and U. Madhow, "Communication limits with low-precision analog-to-digital conversion at the receiver", *Proc. IEEE Intl. Conf. on Communications (ICC'07), Glasgow, Scotland, June 2007.*

- O. Dabeer, J. Singh, and U. Madhow, "On the limits of communication performance with one-bit analog-to-digital conversion", *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communication (SPAWC'06), Cannes, France, July 2006.*

- J. Singh, O. Dabeer, and U. Madhow, "Signal processing with low-precision A/D conversion: a framework for low-cost gigabit wireless communication", *Poster Presented at IEEE Communication Theory Workshop (CTW'06), Puerto Rico, USA, May 2006.*

Abstract

Communication Transceiver Design with Low-Precision Analog-to-Digital Conversion

by

Jaspreet Singh

As communication systems scale up in speed and bandwidth, the cost and power consumption of high-precision (e.g., 10–12 bits) analog-to-digital converter (ADC) becomes the limiting factor in modern receiver architectures based on digital signal processing. One possible approach to relieve this ADC bottleneck is to employ a low-precision (e.g., 1–4 bits) ADC. This may be suitable for applications requiring limited dynamic range, such as line-of-sight communication using small constellations. However, the drastic reduction of ADC precision raises fundamental questions, at both information-theoretic and algorithmic levels, regarding whether it is even possible to engineer a communication link with such a significant nonlinearity so early in the receiver processing. In this thesis, we present results from our efforts towards answering some of these questions.

We first investigate the Shannon-theoretic limits of communication imposed by the choice of low-precision ADC, for transmission over the ideal real additive white Gaussian noise channel. For an ADC employing K quantization bins (i.e., a precision of $\log_2 K$ bits), we prove that the channel capacity can be achieved using a discrete input distribution with at most $K+1$ support points. A joint optimization over the choice of the input and the quantizer is performed, and the obtained numerical results reveal that at SNR up to 20 dB, the use of 2-3 bit ADC incurs a loss of only about 10-15 % in capacity compared to unquantized observations. Furthermore, we observe that a sensible choice of uniform pulse amplitude modulated input, with quantizer thresholds set to perform maximum likelihood hard decisions, achieves performance close to that attained by an optimal input and quantizer pair.

We then turn our attention to the problem of carrier synchronization using low-precision ADC. We focus on a block noncoherent channel model, wherein the phase rotation caused by a small frequency offset, although a priori unknown, can be approximated as constant over a block of symbols. For M-ary phase shift keyed (M-PSK) inputs, the performance of phase-only quantization, which is attractive due to its ease of implementation, is investigated. The symmetry inherent in the resulting phase-quantized channel model is exploited to obtain low-complexity algorithms for channel capacity computation and block noncoherent demodulation. Numerical results, quantifying the channel capacity, and the uncoded error rates, are obtained for QPSK input with different number of phase quantization sectors and different block lengths. Dithering the constellation is shown to improve the performance in the face of drastic quantization.

Contents

Appendices 96

List of Figures

List of Tables

Chapter 1

Introduction

The last decade has witnessed rapid mass market deployment of cellular and wireless local area network communication systems. This has been propelled by the economies of scale provided by the low-cost integrated circuit implementation of sophisticated digital signal processing (DSP) algorithms that perform the bulk of the receiver functionalities, such as synchronization, channel estimation and equalization, demodulation and decoding. An integral component of such *DSP-centric* receiver architectures is the analog-to-digital converter (ADC), which converts the received analog waveform into the digital domain with a sufficiently high precision (Fig. 1.1). As we look to scale this DSP-centric design philosophy to higher speeds and bandwidths (to achieve data rates of the order of multi-Gigabit per second), the ADC becomes a bottleneck: high-speed high-precision ADC is either not available, or is costly and power-hungry [1]. On the other hand, the continuing progress of Moore's "law" [3] implies that the integrated circuit im-

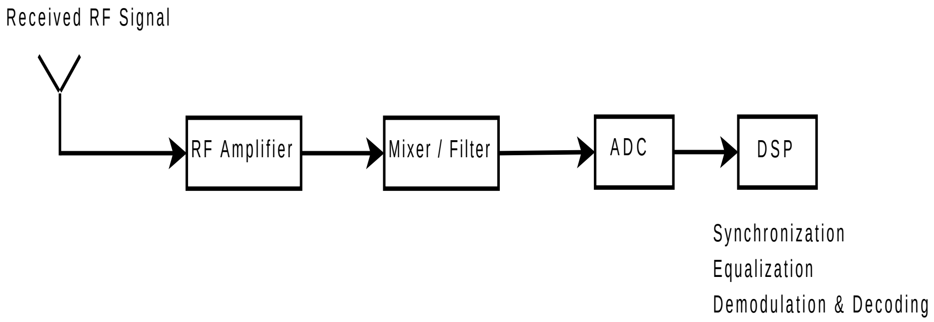

Figure 1.1: Modern DSP-centric receiver design. Does it scale to multi-Gigabit per second speeds ?

plementation of DSP algorithms is expected to continue to scale up in speed and down in cost. It is of interest, therefore, to explore the feasibility of DSP-centic transceiver design with low-precision ADC at the receiver.

The conventional approach to transceiver design, when the available ADC precision is high enough (10–12 bits or more), is to perform the design assuming that the ADC has infinite precision, and to then conduct simulation tests to obtain the algorithmic refinements needed to accommodate the effects of finite precision. This paradigm for design and implementation is predicated on the assumption that the performance with high-precision quantization essentially replicates that with infinite precision. For drastically quantized systems (1–4 bits), this paradigm breaks down, since the effect of such severe quantization is expected to be fundamentally different from that of high-precision quantization. This mandates a comprehensive rethinking of the system design, ranging from a Shannon-theoretic investigation to the design of new algorithms for performing the various receiver operations, with the *starting assumption* that the ADC used at the receiver has low-precision. In this thesis, we present results obtained from our efforts towards building such an understanding of the impact of low-precision ADC.

We focus attention on transmission over the classical bandlimited additive white Gaussian noise (AWGN) channel model. Not only is this model of fundamental significance, it also forms a good approximation for one of the emerging applications for multi-Gigabit communication: short range line-of-sight wireless communication in the 60 GHz mmwave band [4]. Communication in this band must inherently be directional, in order to compensate for the severe free space propagation loss, which scales up as the square of the carrier frequency. Fortunately, the large carrier frequency (and hence the small wavelength) makes it possible to use low-cost antenna arrays in order to synthesize highly directional beams. This cuts down drastically on the multipath, so that a short range directional link operating in this band is well approximated by an AWGN model. Furthermore, the limited dynamic range requirement in an AWGN setting indicates, in the first place, that low-precision ADC may suffice to provide acceptable performance. This is in contrast to transmission over severe fading and dispersive channels, which would necessitate a large receiver dynamic range, unless we perform some precoding at the transmitter. We do not tackle the latter problem in this thesis.

1.1 Contributions

Consider linear modulation over the bandlimited AWGN channel. As a first step, let us assume ideal carrier synchronization (no frequency or phase offset between the local oscillator at the receiver and the incoming carrier wave), and ideal timing synchronization (enabling ideal Nyquist-rate sampling). Under the first assumption, we can separate out the in-phase (I) and quadrature (Q) components, and restrict attention to a real baseband AWGN channel. If the Nyquist-rate samples received over this channel are now quantized drastically, we obtain a real discrete-time memoryless quantized AWGN channel model. As our first problem, we investigate the information-theoretic limits of communication over this channel.

1. The AWGN-Quantized Output Channel

The capacity of the average power constrained discrete-time AWGN channel, along with the fact that a Gaussian input distribution achieves the capacity, is perhaps the most well known result of information theory. Under output quantization, however, we find that the Gaussian input distribution is no longer optimal.

Rather, the capacity can be achieved using a *discrete* input. The main results from our investigation are the following [5].

1. For K-bin (i.e., $\log_2 K$ bits) output quantization, we prove that the input distribution need not have any more than $K + 1$ mass points to achieve the channel capacity. (Numerical computation of optimal input distributions reveals that K mass points are sufficient.) An intermediate result of interest is that, when the channel output is quantized with finite-precision, an average power constraint on the input leads to an implicit peak power constraint, in the sense that an optimal input distribution must have bounded support.

2. For the extreme scenario of 1-bit symmetric quantization, the preceding result is tightened analytically to show that binary antipodal signaling is optimal for any signal-to-noise ratio (SNR). An analytical expression for the channel capacity is also provided. For multi-bit quantizers, tight upper bounds on capacity are obtained using a dual formulation of the channel capacity problem. Near-optimal input distributions that approach these bounds are computed using the cutting-plane algorithm [2].

3. While the preceding results optimize the input distribution for a fixed quantizer, comparison with an unquantized system requires optimization over the choice of the quantizer as well. We numerically obtain optimal 2-bit and 3-

bit symmetric quantizers. From our numerical results, we infer that the use of low-precision ADC incurs a relatively small loss in capacity compared to unquantized observations. For example, at 0 dB SNR, a receiver with 2-bit ADC achieves 95% of the capacity attained with unquantized observations. Even at a moderately high SNR of 20 dB, a receiver with 3-bit ADC achieves 85% of the capacity attained with unquantized observations. This indicates that DSP-centric design based on low-precision ADC is indeed attractive as communication bandwidths scale up, since the small loss in spectral efficiency should be acceptable in this regime. Furthermore, we observe that, for K-bin quantization, a "sensible" choice of standard equiprobable K-level pulse amplitude modulated input, with the ADC thresholds set to implement maximum likelihood hard decisions, achieves performance which is quite close to that obtained by numerical optimization of the quantizer and input distribution.

Given the encouraging nature of these results, our next step is to remove some of the idealizations in the channel model. Towards that, we turn our attention to the problem of carrier synchronization. While the receiver's local oscillator (LO) can be locked to the frequency of the incoming carrier wave using a classical analog feedback loop, we continue with the modern DSP-centric view, in which all of the processing happens at the baseband, and is mostly digital. Thus, we assume that the LO at the receiver employs a fixed frequency, independent of the incoming passband signal.

One possible approach to handle the LO asynchronism is to employ training based methods for explicit estimation and correction of the frequency offset, which if accomplished sufficiently well, takes us back to the *coherent* AWGN channel model considered in our first problem. However, an alternate *noncoherent* approach can eliminate the need for explicit estimation and correction, by exploiting the fact that, in practice, the value of the frequency offset is small enough to assume that the phase after downconversion, although a priori unknown, can be well approximated as constant over a small number of symbols. The classical method to exploit this is to approximate the phase as constant over two symbols and apply differential modulation and demodulation. More recent work has demonstrated that significant performance gains can be obtained by considering larger blocks, at least with unquantized observations. As our second problem, we investigate the impact of low-precision quantization on the performance achievable over the block noncoherent channel, while restricting attention to standard M-ary phase shift keying (M-PSK) input constellations.

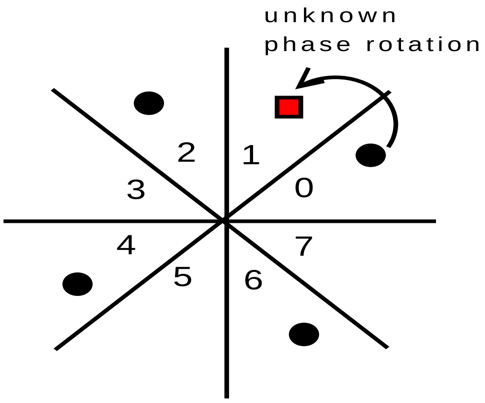

Figure 1.2: QPSK input and 8-sector phase quantization

2. Block Noncoherent Communication with Output Quantization

There can be several ways to quantize a complex-valued received symbol. *Phase quantization*, illustrated in Fig. 1.2, is an attractive option due to its ease of implementation : it eliminates the need for automatic gain control (AGC), and can be implemented using only 1-bit ADCs preceded by analog multipliers (more details are provided later in Chapter 4). Moreover, for PSK inputs, the information is encoded in the phase of the transmitted symbols, so that we can expect phase quantization to perform well. For M-PSK input with K-level uniform phase quantization of the received symbols, we obtain the following results [6, 7].

1. We begin by studying the structure of the input-output relationship of the phase quantized block noncoherent AWGN channel. For the special case when M divides K, we exploit the symmetry inherent in the channel model to derive several results characterizing the output probability distribution

over a block of symbols, both conditioned on the input, and without conditioning. These results are used to obtain a low complexity procedure for computing the capacity of the channel (brute force computation has complexity exponential in the block length L).

2. We also obtain low complexity optimal block noncoherent demodulation rules. These rules are obtained by specializing the existing low complexity procedures for block demodulation with unquantized observations, to our setting with quantized observations. A close analysis of the block demodulator reveals that, depending on the number of quantization sectors, the symmetries inherent in the channel model (which on the one hand help us compute the capacity efficiently) can also have a dire consequence : they can make it impossible to distinguish between the effect of the unknown phase offset and the phase modulation. As a result, we may have two equally likely inputs for certain outputs, irrespective of the block length and the SNR, leading to severe performance degradation. In order to break the undesirable symmetries, we propose a *dithered*-QPSK input scheme, in which we rotate the QPSK constellation across the different symbols in a block.

3. Numerical results are obtained for QPSK input with 8 and 12 sector phase quantization, for different choices of the block length L. We find that 8-

sector quantization, with a dithered-QPSK input, achieves more than 80-85 % of the capacity achieved with unquantized observations (with an identical block length), while with 12-sector quantization, and no dithering, we can get as much as 90-95 % of the unquantized capacity. The corresponding loss in terms of SNR, for fixed capacity, varies between $2-4$ dB for 8-sector quantization, and between $0.5-2$ dB with 12 sectors. In terms of the uncoded symbol error rates (SER), the performance degradation is of the same order. For instance, at SER$=10^{-3}$, the loss for 8 and 12 sector quantization, compared to unquantized observations, is about 4 dB and 2 dB respectively.

1.2 Organization

The rest of this dissertation is organized as follows. In Chapter 2, we provide some background on the different topics related to our work. This includes a brief overview of the ADC technology, survey of the prior work on signal processing with low-precision sampling, and background information on Shannon-theoretic notions. Chapters 3 and 4 present our work on the quantized ideal AWGN channel and the quantized block noncoherent AWGN channel, respectively. Chapter 5 contains our conclusions and directions for future research.

Chapter 2

Background

We start this chapter by highlighting the limitations that hinder the progress of the ADC technology. These limitations and their impacts were illustrated in detail by Walden in a comprehensive state-of-the-art survey conducted in 1999 [1].

2.1 ADC Technology

The results published in Walden's survey are reproduced here in Fig. 2.1. A notable conclusion from the survey was the observation that for large sampling rates (above 2 MS/s), the available ADC precision falls off by 1 bit for every doubling of the sampling rate. This was attributed to *aperture jitter*, the error due to the sample-to-sample variation in the instant of conversion. Another fundamental limitation on the performance of the ADC is imposed by *comparator ambiguity*, which characterizes the uncertainty arising due to the finite speed with

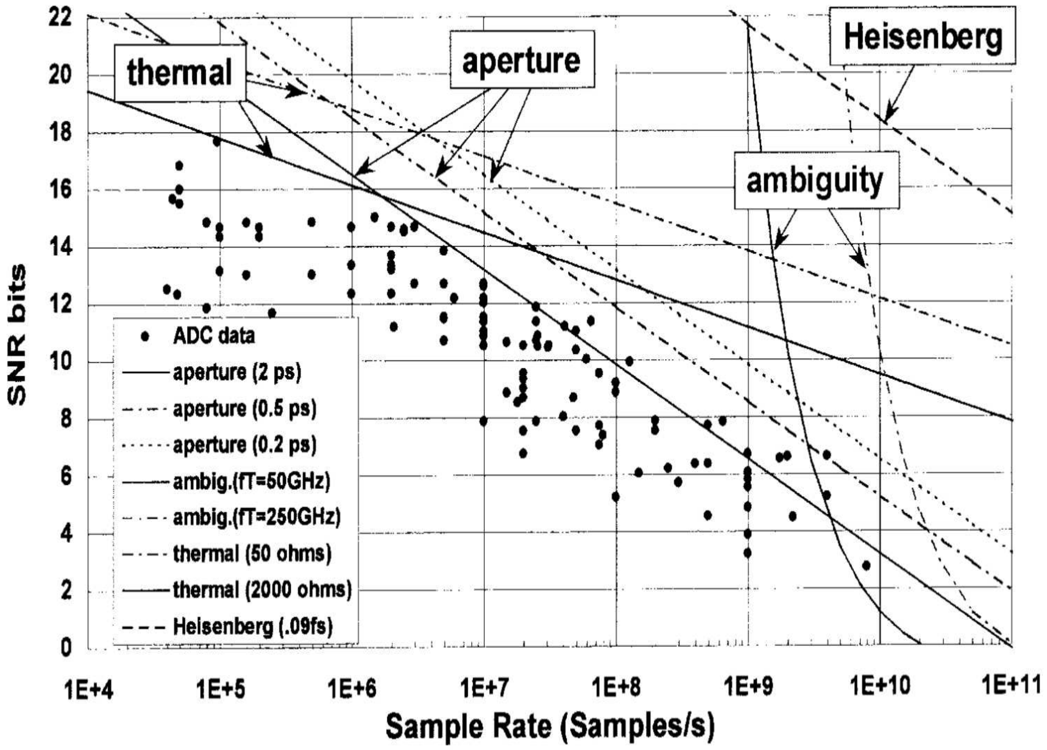

Figure 2.1: ADC technology trends published in Walden's survey [1]. The various curves depict the fundamental limitations on the achievable precision imposed by different non-idealities.

which the comparators used in the ADC can respond to the variations in the input voltage. This ambiguity is governed by the speed of the device technology used to fabricate the ADC. While it places an absolute limit on the sampling rate of the ADC, Walden's analysis showed that it also imposes limits on the achievable precision of the ADC, which falls off rapidly as we move to Gs/s rates. As far as the evolution of the ADC technology over time is concerned, it was observed that the progress was slow, with an average improvement of ~ 1.5 bits for any given sampling frequency over a period of six-eight years.

In addition to the precision and the sampling rate, power dissipation is another key performance measure for ADCs. The power dissipated by an ADC depends on the choice of its architecture. For high-speed applications, the flash architecture is most preferred ([1], see also [8, Chapter 3]) due to its parallel design: to achieve a resolution of N bits, the flash ADC uses $2^N - 1$ comparators sampling the input signal simultaneously. However, the exponential increase in the number of comparators as a function of the ADC precision leads to an exponential increase in the power dissipation as well. Consequently, if we can achieve acceptable communication performance with low-precision ADC, it can lead to significant power savings, while allowing us the liberty to use the fast and simple flash ADC architecture.

2.2 Signal Processing with Low-Precision ADC

Recognizing the limitations imposed by ADC technology, there have been prior efforts in the circuit design community, as well as signal processing and communication communities, to explore the impact of low-precision ADC on communication system design. Most of this work has been in the specific context of Ultrawideband (UWB) systems, designed to operate in the 3.1 to 10.6 GHz band

[9]. In [10], the authors analyze the performance of several different UWB receivers using one-bit ADC, including the matched filter and transmitted reference schemes, as well the use of dither and sigma-delta modulation. The impact of low-precision ADC on the performance of a UWB receiver is studied in [11, 12]. Interference suppression for UWB signaling with one-bit ADC and analog preprocessing is considered in [13]. Decomposition of the UWB signal into parallel frequency channels, using pre-ADC analog components, in order to relax ADC speed requirements is considered in [14, 15, 16]. Methods of moving complexity to the transmitter, and to obtain spatial focusing gains akin to beamforming, by the use of time reversal have been considered in [17, 18]. Finally, a more recent work [19] explores a mixed-signal receiver architecture for designing a 1 Gbps link in the 60 GHz mmwave band using low-precision ADC (4 bits).

The preceding contributions focus on specific applications, and are aimed at devising strategies (possibly analog-centric) that "work" for those scenarios. However, the choice of low-precision ADC for communication system design is clearly going to have fundamental ramifications that go beyond specific applications. The objective of this thesis is to uncover some of these fundamental issues, keeping in mind the long-term goal of devising DSP-centric receiver architectures.

While our emphasis here is on identifying Shannon-theoretic performance limits, there is also prior work on fundamental problems of estimation using low-precision samples that may be relevant for further research on receiver design. Such work includes the use of dither for reconstruction of a signal from its low-precision samples [20, 21, 22], frequency estimation using 1-bit samples [23, 24], study of the choice of the quantization threshold for signal amplitude estimation [25], and signal parameter estimation using 1-bit dithered quantization [26, 27]. The ideas in these papers may be useful for the problems of synchronization, channel estimation and equalization with low-precision ADC.

We now proceed to provide background on the subject of channel capacity, which is the focus of our work in Chapter 3. The relevant background on the problem of carrier synchronization, which we consider in Chapter 4, is provided within that chapter.

2.3 Channel Capacity

The notion of a mathematical model for a noisy communication channel, and its associated capacity, was formally introduced by Shannon in his seminal paper in 1948 [28]. Modeling the channel as a probabilistic communication medium, Shannon defined channel capacity to be the maximum possible rate at which we can transfer data reliably over the channel, and provided an elegant mathematical formula to characterize the capacity. While the result established the absolute limits of communication over a particular channel, it did not provide a constructive coding mechanism which could be used to attain those limits. For a long period of time, the Shannon limit remained elusive, as the performance of practical coding techniques was found to be significantly inferior to what was promised by Shannon's results. This was until 1994, when Berrou et. al dramatically reduced this gap to within a dB of the Shannon limit with the invention of turbo codes [29, 30]. The field of error correction coding has since then been revolutionized, with "turbo-like" codes being developed for a wide variety of channels and rates. The rediscovery of Gallager's low density parity check codes [31] by Mackay [32] in 1999 has also provided an alternative coding approach to attain the Shannon limit. Given these developments, it is of increasing interest to characterize the capacity for different communication channels, hence our interest to study the impact of low-precision ADC on the channel capacity.

2.3.1 Discrete Memoryless Channels

A discrete memoryless channel (DMC) is characterized by a finite set of channel input symbols $\mathcal{X} = \{x_1, \cdots, x_M\}$, a finite set of channel output symbols $\mathcal{Y} = \{y_1, \cdots, y_N\}$, and a transition probability matrix $P = [\mathsf{P}(y_j|x_i)]$, where $\mathsf{P}(y_j|x_i)$ denotes the probability of the received symbol being y_j when the transmitted symbol is x_i. The capacity of this channel (in bits/channel use) is defined to be the maximum possible mutual information between the input and the output.

$$\begin{aligned} C &= \max_{\mathsf{P}_X} I(X;Y) && (2.1) \\ &= \max_{\mathsf{P}_X} \sum_i \mathsf{P}_X(x_i) \sum_j \mathsf{P}(y_j|x_i) \log \frac{\mathsf{P}(y_j|x_i)}{\mathsf{P}_Y(y_j;\mathsf{P}_X)} \quad , && (2.2) \end{aligned}$$

where P_X denotes the probability mass function (PMF) of the channel input X and $\mathsf{P}_Y(\cdot;\mathsf{P}_X)$ denotes the PMF of the channel output Y induced by the input PMF P_X. For simple channel models (e.g., binary and/or symmetric channels), it is usually possible to perform the optimization directly. For complicated channels, the celebrated Blahut-Arimoto algorithm [33, 34] can be employed to perform the optimization in iterative steps that guarantee convergence to the capacity.

2.3.2 Continuous Alphabet Channels

When the channel input and output are allowed to take a continuum of values, rather than a finite discrete set of values, the capacity is given by

$$\begin{aligned} C &= \sup_{F_X} I(X;Y) && (2.3) \\ &= \sup_{F_X} \int\int \mathsf{P}(y|x) \log \frac{\mathsf{P}(y|x)}{\mathsf{P}_Y(y;F_X)} dy \; dF_X(x) \quad , && (2.4) \end{aligned}$$

where $\mathsf{P}(y|x)$ represents the transition density function that defines the channel, F_X is the cumulative distribution function (CDF) of the channel input, and $\mathsf{P}_Y(y;F_X)$ is the output density induced by the input F. When the input and/or the output are discrete, the corresponding integral in (2.4) can be collapsed into a summation, so that (2.2) is a special case of (2.4). The optimization in (2.4) is usually performed under some set of power constraints on the input X.

From Shannon's seminal work in [28] for a continuous-time bandlimited additive white Gaussian noise (AWGN) channel, we know that the capacity for a discrete-time real AWGN channel, under an average power constraint, is achieved by a Gaussian input distribution. This is perhaps the most well-known result of information theory. Shannon also considered the peak power constrained problem (again for the continuous-time channel), and obtained a lower bound on the channel capacity, but could not characterize the optimal input. Smith in [35] showed the rather surprising result that under peak power constraint, the capacity of the discrete-time real AWGN channel is achieved by a *discrete* input distribution, with a finite number of mass points. This was generalized to the discrete-time complex AWGN channel in [36].

The optimality of a discrete input has recently been shown to hold for several other continuous-alphabet channel models as well. This includes the Rayleigh fading channel [37], Rician fading channel [38], vector Gaussian channels [39], noncoherent AWGN channel [40], and the generalization of Smith's results to a bigger class of scalar additive channels [41]. For computation of the capacity for continuous alphabet channels, one possibility is to use the Blahut-Arimoto algorithm after (finely) quantizing the support set of the input. An alternate approach based on linear programming, aimed at finding near-optimal discrete input distributions for channels with continuous alphabets, has been recently proposed in [2].

Moving on to the special case of finite-output channels (which is the subject of interest for our work), there is prior work for both scenarios: when the input alphabet is also finite (which corresponds to a DMC), and when the input is allowed to be continuous. For the DMC, Gallager's classic text [42] shows that for output cardinality K, the capacity can be achieved by putting nonzero probability mass on at most K input points. For continuous input alphabet, Witsenhausen [43] has used Dubins' theorem [44] to show an analogous result for *peak power* constrained input. The key to the proof of our central result reported in Chapter 3, that the *average power* constrained capacity for K-level output quantization can be attained with at most $K+1$ points, is to show that under output quantization, an average power constraint automatically induces a constraint on the peak power of the input. Once we have that, we use Dubins' theorem in a manner analogous to that in Witsenhausen's work.

It is important to mention that there is also prior work on the impact of output quantization on the mutual information achievable with *fixed input distribution* [45, 46, 47]. However, we are not aware of an information-theoretic investigation with output quantization that includes optimization of the input distribution. Another related class of problems that deserves mention relates to the impact of finite-precision quantization on the information-theoretic measure of channel cut-off rate rather than channel capacity (e.g., see [48, 49]).

2.4 Recent Related Work

Since the publication of the preliminary results from this thesis in [50, 51, 52], there have been other related efforts in the communication and information theory literature to investigate the impact of low-precision ADC on system design. Reference [53] considers the impact of overflows on the capacity under output quantization, while reference [54] studies the fundamental communication limits imposed by the precision of the ADC, from an information-theoretic view as well as a theoretical physics view, applying the Heisenberg's uncertainty principle to the process of analog-to-digital conversion. Impact of receiver quantization in the context of fading channels, and multiple input multiple output (MIMO) systems, has been studied in [55, 56] respectively. Reference [57] investigates a scheme in which the received complex signal is quantized using three low-precision ADCs with three different phases (similar in principle to the phase quantization approach considered in Chapter 4), and illustrates the achievable power savings using such an approach, while working with a mean squared error criterion to evaluate the quantization performance.

Chapter 3

The AWGN-Quantized Output Channel

The discrete-time memoryless *AWGN-Quantized Output* (AWGN-QO) channel is

$$Y = \mathsf{Q}\left(X + N\right) \ . \tag{3.1}$$

Here $X \in \mathbb{R}$ is the channel input with cumulative distribution function $F(x)$, $Y \in \{y_1, \cdots, y_K\}$ is the (discrete) channel output, and N is $\mathcal{N}(0, \sigma^2)$ (the Gaussian random variable with mean 0 and variance σ^2). Q maps the real valued input $X + N$ to one of the K bins, producing a discrete output Y. In this work, we only consider quantizers for which each bin is an interval of the real line. The quantizer Q with K bins is therefore characterized by the set of its $(K-1)$ thresholds $\boldsymbol{q} := [q_1, q_2, \cdots, q_{K-1}] \in \mathbb{R}^{K-1}$, such that $-\infty := q_0 < q_1 < q_2 < \cdots < q_{K-1} < q_K := \infty$. The output Y is assigned the value y_i when the quantizer input $(X + N)$ falls in the i^{th} bin, which is given by the interval $(q_{i-1}, q_i]$. The resulting transition probability functions are

$$\begin{aligned} W_i(x) &= \mathsf{P}(Y = y_i | X = x) \\ &= Q\left(\frac{q_{i-1} - x}{\sigma}\right) - Q\left(\frac{q_i - x}{\sigma}\right), \qquad 1 \le i \le K, \end{aligned} \tag{3.2}$$

where $Q(\cdot)$ is the complementary Gaussian distribution function,

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp(-t^2/2)dt \ . \tag{3.3}$$

The Probability Mass Function (PMF) of the output Y, corresponding to the input distribution F is

$$R(y_i; F) = \int_{-\infty}^{\infty} W_i(x) dF(x), \qquad 1 \le i \le K, \tag{3.4}$$

and the input-output mutual information $I(X;Y)$, expressed explicitly as a function of F is

$$I(F) = \int_{-\infty}^{\infty} \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F)} dF(x) \ .^1 \tag{3.5}$$

Under an average power constraint P, we wish to find the capacity of the channel (3.1), given by

$$C = \sup_{F \in \mathcal{F}} I(F), \tag{3.6}$$

where $\mathcal{F} = \left\{ F : \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 dF(x) \le P \right\}$, i.e., the set of all average power constrained distributions on $\mathbb{R}$.

[1] The logarithm is base 2 throughout the chapter, so the mutual information is measured in bits.

Existence of an optimal input: The fact that there exists an input distribution that achieves the capacity C follows by standard function analytic arguments. See Appendix for details.

3.1 Structure of Optimal Inputs

We begin by employing the Karush-Kuhn-Tucker (KKT) optimality condition to show that, even though we have not imposed a peak power constraint on the input, it is automatically induced by the average power constraint. Specifically, a capacity achieving distribution for the AWGN-QO channel (3.1) must have bounded support.

3.1.1 Bounded Support

The KKT optimality condition for an average power constrained channel has been derived in [37]. The mild technical conditions required for it to hold are verified for our channel model in the Appendix. The condition states that an input distribution F^* achieves the capacity C in (3.6) *if and only if* there exists $\gamma \geq 0$ such that

$$\sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F^*)} + \gamma(P - x^2) \leq C \tag{3.7}$$

for all x, with equality if x is in the support [2] of F^*, where the transition probability function $W_i(x)$, and the output probability $R(y_i; F^*)$ are as specified in (3.2) and (3.4), respectively.

The summation on the left-hand side (LHS) of (3.7) is the Kullback-Leibler divergence (or the relative entropy) between the transition PMF $\{W_i(x), i = 1, \ldots, K\}$ and the output PMF $\{R(y_i; F), i = 1, \ldots, K\}$. For convenience, let us denote this divergence function by $d(x; F)$, that is,

$$d(x; F) = \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F)} \ . \tag{3.8}$$

We begin by studying the behavior of this function in the limit as $x \to \infty$.

Lemma 1 *For the AWGN-QO channel* (3.1)*, the divergence function* $d(x; F)$ *satisfies the following properties*

(a) $\lim_{x\to\infty} d(x; F) = -\log R(y_K; F)$.

(b) There exists a finite constant A_0 *such that*

for $x > A_0, d(x; F) < -\log R(y_K; F)$.[3]

[2] The support of a distribution F (or the set of increase points of F) is the set $S_X(F) = \{x : F(x + \epsilon) - F(x - \epsilon) > 0, \ \forall \epsilon > 0\}$.

[3] *The constant* A_0 *depends on the choice of the input* F*. For notational simplicity, we do not explicitly show this dependence.*

Proof : We have

$$\begin{aligned} d(x;F) &= \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i;F)} \\ &= \sum_{i=1}^{K} W_i(x) \log W_i(x) - \sum_{i=1}^{K} W_i(x) \log R(y_i;F) \ . \end{aligned}$$

As $x \to \infty$, the PMF $\{W_i(x), i = 1, \ldots, K\} \to 1(i = K)$, where $1(\cdot)$ is the indicator function. This observation, combined with the fact that the entropy of a finite alphabet random variable is a continuous function of its probability law, gives $\lim_{x\to\infty} d(x;F) = 0 - \log R(y_K;F) = -\log R(y_K;F)$.

Next we prove part (b). For $x > q_{K-1}$, $W_i(x)$ is a strictly decreasing function of x for $i \leq K-1$ and strictly increasing function of x for $i = K$. Since $\{W_i(x)\} \to 1(i = K)$ as $x \to \infty$, it follows that there is a constant A_0 such that $W_i(A_0) < R(y_i;F)$ for $i \leq K-1$ and $W_K(A_0) > R(y_K;F)$. Therefore, it follows that for $x > A_0$,

$$\begin{aligned} d(x;F) &= \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i;F)} \\ &< W_K(x) \log \frac{W_K(x)}{R(y_K;F)} < -\log R(y_K;F). \end{aligned}$$

■

The saturating nature of the divergence function for the AWGN-QO channel, as stated above, coupled with the KKT condition, is now used to prove that a capacity achieving distribution must have bounded support.

Proposition 1 *For the average power constrained AWGN-QO channel* (3.1)*, an optimal input distribution must have bounded support.*

Proof : Let F^* be an optimal input, so that there exists $\gamma \geq 0$ such that (3.7) is satisfied with equality at every point in the support of F^*. We exploit this necessary condition to show that the support of F^* is upper bounded. Specifically, we prove that there exists a finite constant ${A_2}^*$ such that it is not possible to attain equality in (3.7) for any $x > {A_2}^*$.

From Lemma 1, we get $\lim\limits_{x\to\infty} d(x;F^*) = -\log R(y_K;F^*) =: L$. Also, there exists a finite constant A_0 such that for $x > A_0, d(x;F^*) < L$.

We consider two possible cases.

- Case 1: $\gamma > 0$.

  For $x > A_0$, we have $d(x;F^*) < L$.

  For $x > \sqrt{\max\{0,(L - C + \gamma P)/\gamma\}} =: \tilde{A}$, we have $\gamma(P - x^2) < C - L$.

  Defining $A_2^* = \max\{A_0, \tilde{A}\}$, we get the desired result.

- Case 2: $\gamma = 0$.

  Since $\gamma = 0$, the KKT condition (3.7) reduces to

  $$d(x;F^*) \leq C \ , \quad \forall x.$$

  Taking limit $x \to \infty$ on both sides, we get

  $L = \lim\limits_{x\to\infty} d(x;F^*) \leq C.$

Hence, choosing $A_2^* = A_0$, for $x > A_2^*$ we get, $d(x; F^*) < L \leq C$, that is, $d(x; F^*) + \gamma(P - x^2) < C$.

Combining the two cases, we have shown that the support of the distribution F^* has a finite upper bound ${A_2}^*$. Using similar arguments, it can easily be shown that the support of F^* has a finite lower bound ${A_1}^*$ as well, which implies that F^* has bounded support. ■

Remark: For the (unquantized) AWGN channel, we know that the optimal input has a Gaussian distribution, so that the support is unbounded. It is worth checking, therefore, the behavior of the divergence function for the unquantized AWGN channel. It is easy to obtain

$$d_{AWGN}(x) = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right) - \frac{1}{2\ln 2}\left(\frac{P - x^2}{P + \sigma^2}\right) \ ,$$

which does not saturate, but rather goes to ∞ as $x \to \infty$, enabling an equality in the KKT condition even as $x \to \infty$

3.1.2 Optimality of a Discrete Input

In [43], Witsenhausen considered a stationary discrete-time memoryless channel, with a continuous input X taking values on the bounded interval $[A_1, A_2] \subset \mathbb{R}$, and a discrete output Y of finite cardinality K. Using Dubins' theorem [44], it was shown that if the transition probability functions are continuous (i.e., $W_i(x)$

is continuous in x, for each $i = 1, \cdots, K$), then the capacity is achievable by a discrete input distribution with at most K mass points. As stated in Proposition 2 below (proved in the Appendix), this result can be extended to show that, if an *additional* average power constraint is imposed on the input, the capacity is then achievable by a discrete input with at most $K+1$ mass points.

Proposition 2 *Consider a stationary discrete-time memoryless channel with a continuous input X that takes values in the bounded interval $[A_1, A_2]$, and a discrete output $Y \in \{y_1, y_2, \cdots, y_K\}$. Let the channel transition probability function $W_i(x) = \mathsf{P}(Y = y_i | X = x)$ be continuous in x for each i, where $1 \leq i \leq K$. The capacity of this channel, under an average power constraint on the input, is achievable by a discrete input distribution with at most $K+1$ mass points.*

Proof : See Appendix. ■

Proposition 2, coupled with the implicit peak power constraint derived in the previous subsection (Proposition 1), gives us the following result.

Theorem 1 *The capacity of the average power constrained AWGN-QO channel* (3.1) *is achievable by a discrete input distribution with at most $K+1$ points of support.*

Proof : From Proposition 1, we know that an optimal input F^* has bounded support $[{A_1}^*, {A_2}^*]$. Hence, to obtain the capacity in (3.6), we can maximize $I(F)$ over only those average power constrained distributions that have support in $[{A_1}^*, {A_2}^*]$. Since the transition functions $W_i(x)$ are continuous, Proposition 2 guarantees that this maximum is achievable by a discrete input with at most $K+1$ points. ■

Note that our result does not guarantee uniqueness of the capacity achieving input.

3.1.3 Symmetric Inputs for Symmetric Quantization

For our capacity computations ahead, we assume that the quantizer Q employed in (3.1) is symmetric, i.e., its threshold vector $\boldsymbol{q}$ is symmetric about the origin. Given the symmetric nature of the AWGN noise and the power constraint, it seems intuitively plausible that restriction to symmetric quantizers should not be suboptimal from the point of view of optimizing over the quantizer choice in (3.1), although a proof of this conjecture has eluded us. However, once we assume that the quantizer in (3.1) is symmetric, we can restrict attention to only symmetric inputs without loss of optimality, as stated in the following Lemma.[4]

Lemma 2 *If the quantizer in* (3.1) *is symmetric, then, without loss of optimality, we can consider only symmetric inputs for the capacity computation in* (3.6).

[4] A random variable X (with distribution F) is symmetric if X and $-X$ have the same distribution, that is, $F(x) = 1 - F(-x), \forall\, x \in \mathbb{R}$.

Proof: Suppose we are given an input random variable X (with distribution F) that is not necessarily symmetric. Denote the distribution of $-X$ by G (so that $G(x) = 1 - F(-x),\ \forall\ x \in \mathbb{R}$). Due to the symmetric nature of the noise N and the quantizer Q, it is easy to see that X and $-X$ result in the same input-output mutual information, that is, $I(F) = I(G)$. Consider now the following *symmetric* mixture distribution

$$\tilde{F}(x) = \frac{F(x) + G(x)}{2}.$$

Since the mutual information is concave in the input distribution, we get $I(\tilde{F}) \geq \frac{I(F)+I(G)}{2} = I(F)$, which proves the desired result. ■

In the next subsection, we consider the extreme scenario of 1-bit quantization, and tighten the result in Theorem 1 to show that binary antipodal signaling achieves capacity.

3.1.4 1-bit Quantization: Antipodal Signaling is Optimal

With 1-bit symmetric quantization, the channel is

$$Y = \mathrm{sign}(X + N). \tag{3.9}$$

Theorem 1 (Section 3.1.2) guarantees that the capacity of this channel, under an average power constraint, is achievable by a discrete input distribution with at most 3 points. This result is further tightened by the following theorem that shows the optimality of binary antipodal signaling for all SNRs.

Theorem 2 *For the* 1*-bit symmetric quantized channel model* (3.9)*, the capacity is achieved by binary antipodal signaling and is given by*

$$C = 1 - h\left(Q\left(\sqrt{\mathsf{SNR}}\right)\right), \qquad \mathsf{SNR} = \frac{P}{\sigma^2} \ ,$$

where $h(\cdot)$ *is the binary entropy function,*

$$h(p) = -p\log(p) - (1-p)\log(1-p) \ , \quad 0 \leq p \leq 1 \ ,$$

and $Q(\cdot)$ is the complementary Gaussian distribution function shown in (3.3).

Proof : Since Y is binary it is easy to see that

$$H(Y|X) = \mathbb{E}\left[h\left(Q\left(\frac{X}{\sigma}\right)\right)\right] \ ,$$

where $\mathbb{E}$ denotes the expectation operator. Therefore

$$I(X,Y) = H(Y) - \mathbb{E}\left[h\left(Q\left(\frac{X}{\sigma}\right)\right)\right] \ ,$$

which we wish to maximize over all input distributions satisfying $\mathbb{E}[X^2] \leq P$. Since the quantizer is symmetric, we can restrict attention to symmetric input distributions without loss of optimality (cf. Lemma 2). On doing so, we obtain that the PMF of the output Y is also symmetric (since the quantizer and the noise are already symmetric). Therefore, $H(Y) = 1$ bit, and we get

$$C = 1 - \min_{\substack{X \text{ symmetric} \\ \mathbb{E}[X^2] \leq P}} \mathbb{E}\left[h\left(Q\left(\frac{X}{\sigma}\right)\right)\right].$$

Since $h(Q(z))$ is an even function, we get that

$$H(Y|X) = \mathbb{E}\left[h\left(Q\left(\frac{X}{\sigma}\right)\right)\right] = \mathbb{E}\left[h\left(Q\left(\frac{|X|}{\sigma}\right)\right)\right].$$

In the Appendix we show that the function $h(Q(\sqrt{z}))$ is convex in z. Jensen's inequality [58] thus implies

$$H(Y|X) \geq h\left(Q\left(\sqrt{\mathsf{SNR}}\right)\right)$$

with equality iff $X^2 = P$. Coupled with the symmetry condition on X, this implies that binary antipodal signaling achieves capacity and the capacity is

$$C = 1 - h\left(Q\left(\sqrt{\mathsf{SNR}}\right)\right).$$

■

3.2 Capacity Computation

In this section, we consider capacity computation for K-bin symmetric quantization, with $K > 2$. Every choice of the quantizer results in a unique channel model (3.1). This section discusses capacity computation assuming a fixed quantizer only. Optimization over the quantizer choice is performed in Section 3.3.

3.2.1 Cutting-Plane Algorithm

Contrary to the 1-bit case, closed form expressions for optimal input and capacity appear unlikely for multi-bit quantization, due to the complicated expression for mutual information. We therefore resort to the cutting-plane algorithm [2, Sec IV-A] to generate optimal inputs numerically. For channels with continuous input alphabets, the cutting-plane algorithm can, in general, be used to generate nearly optimal discrete input distributions. It is therefore well matched to our problem, for which we already know that the capacity is achievable by a discrete input distribution.

For our simulations, we fix the noise variance $\sigma^2 = 1$, and vary the power P to obtain capacity at different SNRs. To apply the cutting-plane algorithm, we take a fine quantized discrete grid on the interval $[-10\sqrt{P}, 10\sqrt{P}]$, and optimize the input distribution over this grid. Note that Proposition 1 (Section 3.1.1) tells us that an optimal input distribution for our problem must have bounded support, but it does not give explicit values that we can use directly in our simulations. However, on employing the cutting-plane algorithm over the interval $[-10\sqrt{P}, 10\sqrt{P}]$, we find that the resulting input distributions have support sets well within this interval. Moreover, increasing the interval length further does not change these results.

While the cutting-plane algorithm optimizes the distribution of the channel input, a dual formulation of the channel capacity problem, involving an optimization over the output distribution, can alternately be used to obtain easily computable tight upper bounds on the capacity. We discuss these duality-based upper bounds next.

3.2.2 Duality-based Upper Bound

In the dual formulation of the channel capacity problem, we focus on the distribution of the output, rather than that of the input. Specifically, assume a channel with input alphabet $\mathcal{X}$, transition law $W(y|x)$, and an average power constraint P. Then, for every choice of the output distribution $R(y)$, we have the following upper bound on the channel capacity C

$$C \leq U(R) = \min_{\gamma \geq 0} \sup_{x \in \mathcal{X}} [D(W(\cdot|x)||R(\cdot)) + \gamma(P - x^2)] \ , \tag{3.10}$$

where γ is a Lagrange parameter, and $D(W(\cdot|x)||R(\cdot))$ is the divergence between the transition and output distributions. While [59] provides this bound for a discrete channel, its extension to continuous alphabet channels has been established in [60]. For a more detailed perspective on duality-based upper bounds, see [61].

For an arbitrary choice of $R(y)$, the bound (3.10) might be quite loose. Therefore, to obtain a tight upper bound, we may need to evaluate (3.10) for a large number of output distributions and pick the minimum of the resulting upper bounds. This could be tedious in general, especially if the output alphabet is continuous. However, for the channel model we consider, the output alphabet is discrete with small cardinality. For example, for 2-bit quantization, the space of all symmetric output distributions is characterized by a single parameter $\alpha \in (0, 0.5)$. This makes the dual formulation attractive, since we can easily obtain a tight upper bound on capacity by evaluating the upper bound in (3.10) for different choices of α.

It remains to specify how to compute the upper bound (3.10) for a given output distribution R. For our problem, the favorable nature of the divergence function $D(W(\cdot|x)||R(\cdot))$ facilitates a systematic procedure to do this, as discussed next.

Computation of the Upper Bound: For convenience, we denote $d(x) = D(W(\cdot|x)||R(\cdot))$, and $g(x, \gamma) = d(x) + \gamma(P - x^2)$. For symmetric quantizer and symmetric output distribution, the function g is also symmetric in x, so that we need to compute $\min_{\gamma \geq 0} \sup_{x \geq 0} g(x, \gamma)$. Consider first the maximization over x, for a fixed γ. Although we need to perform this maximization over $x \geq 0$, from a practical standpoint, we can restrict attention to a bounded interval $x \in [0, M]$ only. This is justified as follows. From Lemma 1, we know that $\lim_{x \to \infty} d(x) = \log \dfrac{1}{R(y_K)}$. The saturating nature of $d(x)$, coupled with the non-increasing nature of $\gamma(P - x^2)$, implies that for all practical purposes, the search for the supremum of $d(x) + \gamma(P - x^2)$ over $x \geq 0$ can be restricted to $x \in [0, M]$, where M is chosen large enough to ensure that the difference $|d(x) - \log \frac{1}{R(y_K)}|$ is negligible for $x > M$. In our simulations, we take $M = q_{K-1} + 5\sigma$, where q_{K-1} is the largest quantizer threshold, and σ^2 is the noise variance. This choice of M ensures that for $x > M$, the conditional PMF $W_i(x)$ is nearly the same as the unit mass at $i = K$, which consequently makes the difference between $d(x)$ and $\log \frac{1}{R(y_K)}$ negligible for $x > M$, as desired.

We now need to compute $\min\limits_{\gamma \geq 0} \max\limits_{x \in [0,M]} \{g(x,\gamma)\}$. To do this, we quantize the interval $[0, M]$ to generate a fine grid $\{x_1, x_2, \cdots, x_I\}$, and approximate the maximization over $x \in [0, M]$ as a maximization over this quantized grid, so that we need to compute the function $\min\limits_{\gamma \geq 0} \max\limits_{1 \leq i \leq I} g(x_i, \gamma)$. Denoting $r_i(\gamma) := g(x_i, \gamma)$, this becomes $\min\limits_{\gamma \geq 0} \max\limits_{1 \leq i \leq I} r_i(\gamma)$. Hence, we are left with the task of minimizing (over γ) the maximum value of a finite set of functions of γ, which in turn can be done directly using the standard numerical tools (e.g., $fminimax$ in Matlab). Moreover, we note that the function being minimized over γ, i.e. $m(\gamma) := \max\limits_{1 \leq i \leq I} r_i(\gamma)$, is convex in γ. This follows from the observation that each of the functions $r_i(\gamma) = d(x_i) + \gamma(P - {x_i}^2)$ is convex in γ (in fact, affine in γ), so that their pointwise maximum is also convex in γ [62, pp. 81]. The convexity of $m(\gamma)$ guarantees that $fminimax$ provides us the global minimum over γ.

3.2.3 Numerical Example

We compare results obtained using the cutting-plane algorithm with capacity upper bounds obtained using the dual formulation. We consider 2-bit quantiza-

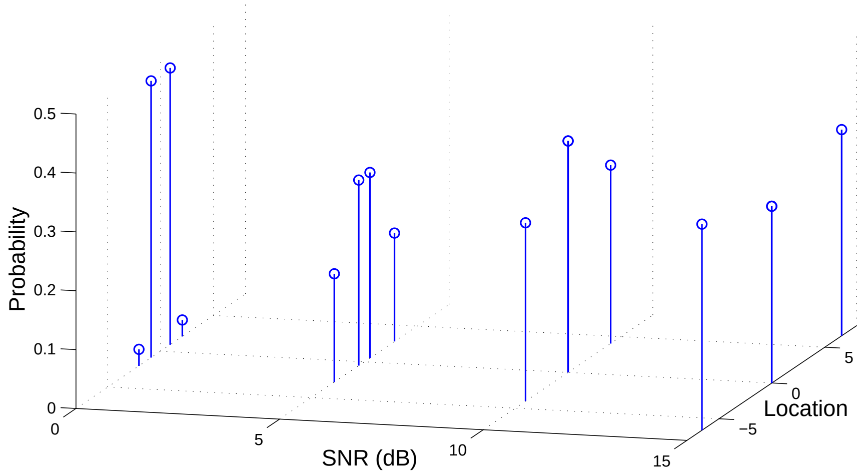

Figure 3.1: Probability mass function of the optimal input generated by the cutting-plane algorithm [2] at various SNRs, for the 2-bit symmetric quantizer with thresholds $\{-2, 0, 2\}$. (noise variance $\sigma^2 = 1$.)

tion, and provide results for the specific choice of quantizer having thresholds at $\{-2, 0, 2\}$.

The input distributions generated by the cutting-plane algorithm at various SNRs (setting $\sigma^2 = 1$) are shown in Figure 3.1, and the mutual information achieved by them is given in Table 3.1. As predicted by Theorem 1 (Section 3.1.2), the support set of the input distribution (at each SNR) has cardinality ≤ 5.

For upper bound computations, we evaluate (3.10) for different symmetric output distributions. For 2-bit quantization, the set of symmetric outputs is characterized by just one parameter $\alpha \in (0, 0.5)$, with the probability distribution on the output being $\{0.5-\alpha, \alpha, \alpha, 0.5-\alpha\}$. We vary α over a fine discrete grid on

$\mathsf{SNR}(dB)$	-5	0	5	10	15	20
Upper Bound	0.163	0.406	0.867	1.386	1.513	1.515
MI	0.155	0.405	0.867	1.379	1.484	1.484

Table 3.1: Duality-based upper bounds on channel capacity, compared with the mutual information (MI) achieved by the distributions generated using the cutting-plane algorithm.

$(0, 0.5)$, and compute the upper bound for each value of α. The least upper bound achieved thus, at a number of different SNRs, is shown in Table 3.1. The small gap between the upper bound and the mutual information (at every SNR) shows the tightness of the obtained upper bounds, and also confirms the near-optimality of the input distributions generated by the cutting-plane algorithm.

It is insightful to verify that the preceding near-optimal input distributions satisfy the KKT condition (3.7). For instance, consider an SNR of 5 dB, for which the input distribution generated by the cutting-plane algorithm has support set $\{-2.86, -0.52, 0.52, 2.86\}$. Figure 3.2 plots, as a function of x, the LHS of (3.7) for this input distribution. (The value of γ used in the plot was obtained by equating the LHS of (3.7) to the capacity value of 0.867, at $x = 0.52$.) The KKT condition is seen to be satisfied (up to the numerical precision of our computations), as the LHS of (3.7) equals the capacity at points in the support set of the input, and is less than the capacity everywhere else. Note that we show the plot for $x \geq 0$ only because it is symmetric in x.

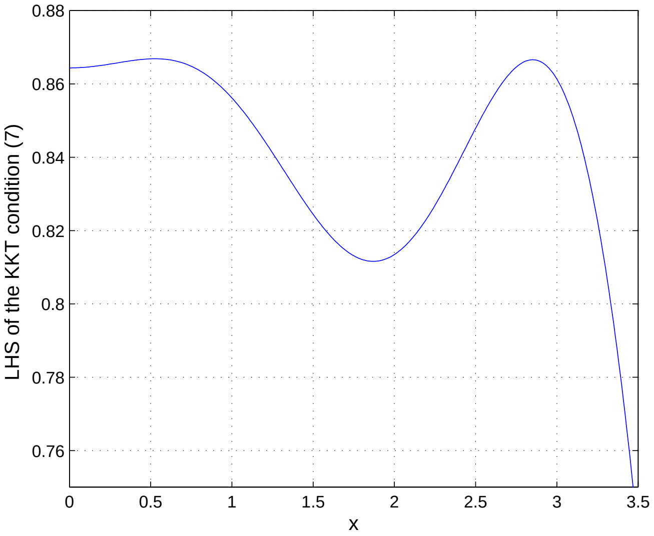

Figure 3.2: The left-hand side of the KKT condition (3.7) for the input distribution generated by the cutting-plane algorithm ($\mathsf{SNR} = 5$ dB). The KKT condition is seen to be satisfied (up to the numerical precision of our computations).

3.3 Quantizer Optimization

Until now, we have addressed the problem of optimizing the input distribution for a fixed output quantizer. In this section, we optimize over the choice of the quantizer, and present numerical results for 2-bit and 3-bit symmetric quantization.

A Simple Benchmark Input-Quantizer Pair: While an optimal quantizer, along with a corresponding optimal input distribution, provides the absolute communication limits for our model, we do not have a simple analytical characterization of their dependence on SNR. From a system designer's perspective, therefore, it is of interest to also examine suboptimal choices that are easy to adapt as a function of

SNR, as long as the penalty relative to the optimal solution is not excessive. Specifically, we take the following input and quantizer pair to be our *benchmark* strategy : for a K-bin quantizer, consider equiprobable, equispaced K-PAM (pulse amplitude modulated) input, with quantizer thresholds chosen to be the mid-points of the input mass point locations. That is, the quantizer thresholds correspond to the ML (maximum likelihood) hard decision boundaries. Both the input mass points and the quantizer thresholds have a simple, well-defined dependence on SNR, and can therefore be adapted easily at the receiver based on the measured SNR. With our K-point uniform PAM input, we have the entropy $H(X) = \log_2 K$ bits for any SNR. Also, it is easy to see that as $\mathsf{SNR} \to \infty$, $H(X|Y) \to 0$ for the benchmark input-quantizer pair. This implies that the benchmark scheme is near-optimal if we operate at high SNR. The issue to investigate therefore is how much gain an optimal quantizer and input pair provides over this benchmark at low to moderate SNR. Note that, for 1-bit symmetric quantization, the benchmark input corresponds to binary antipodal signaling, which has already been shown to be optimal for all SNRs.

As before, we set the noise variance $\sigma^2 = 1$ for convenience. Of course, the results are scale-invariant, in the sense that if both P and σ^2 are scaled by the same factor R (thus keeping the SNR unchanged), then there is an equivalent quantizer (obtained by scaling the thresholds by $\sqrt{R}$) that gives identical performance.

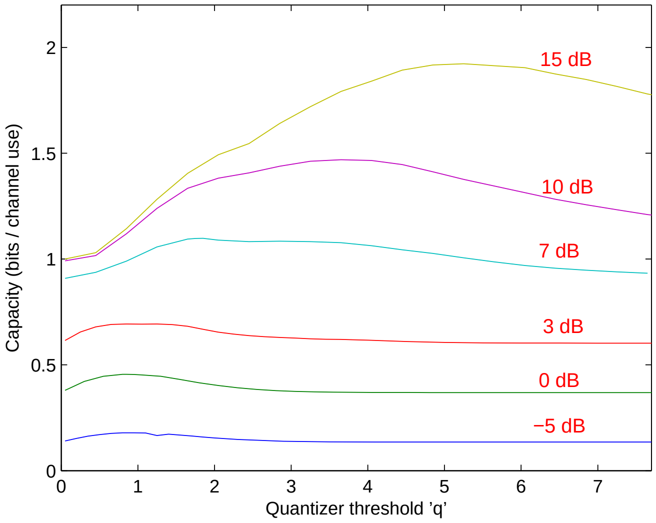

Figure 3.3: 2-bit symmetric quantization : channel capacity (in bits per channel use) as a function of the quantizer threshold q. (noise variance $\sigma^2 = 1$.)

3.3.1 2-Bit Quantization

A 2-bit symmetric quantizer is characterized by a single parameter q, with the quantizer thresholds being $\{-q, 0, q\}$. We therefore employ a brute force search over q to find an optimal 2-bit symmetric quantizer. In Figure 3.3, we plot the variation of the channel capacity (computed using the cutting-plane algorithm) as a function of the parameter q at various SNRs. Based on our simulations, we make the following observations:

- For any SNR, there is an optimal choice of q which maximizes capacity. For the benchmark quantizer (which is optimal at high SNR), q scales as $\sqrt{\mathsf{SNR}}$, hence it is not surprising to note that the optimal value of q we obtain increases monotonically with SNR at high SNR.

- For low SNRs, the variation in the capacity as a function of q is quite small, whereas the variation becomes appreciable as the SNR increases. A practical implication of this observation is that imperfections in Automatic Gain Control (AGC) have more severe consequences at higher SNRs.

- For any SNR, as $q \to 0$ or $q \to \infty$, we approach the same capacity as with 1-bit symmetric quantization (not shown for $q \to \infty$ in the plots for 10 and 15 dB in Figure 3.3). This conforms to intuition: $q = 0$ reduces the 2-bit quantizer to a 1-bit quantizer, while $q \to \infty$ renders the thresholds at $-q$ and q ineffective in distinguishing between two finite valued inputs, so that only the comparison with the quantizer threshold at 0 yields useful information.

Comparison with the Benchmark: In Table 3.2, we compare the performance of the preceding optimal solutions with the benchmark scheme (see the relevant columns for 2-bit ADC). The corresponding plots are shown in Figure 3.5. In addition to being nearly optimal at high SNR, the benchmark scheme is seen to perform fairly well at low to moderate SNR as well. For instance, even at -10 dB SNR, which might correspond to a wideband system designed for very low bandwidth efficiency, it achieves 86% of the capacity achieved with optimal choice of 2-bit quantizer and input distribution. On the other hand, for SNR of 0 dB or above, the capacity is better than 95% of the optimal. These results are encouraging from a practical standpoint, given the ease of implementing the benchmark scheme.

Optimal Input Distributions: It is interesting to examine the optimal input distributions (given by the cutting-plane algorithm) corresponding to the optimal quantizers obtained above. Figure 3.4 shows these distributions, along with optimal quantizer thresholds, for different SNRs. The solid vertical lines show the locations of the input distribution points and their probabilities, while the quantizer thresholds are depicted by the dashed vertical lines. As expected, binary signaling is found to be optimal for low SNR, since it would be difficult for the receiver to distinguish between multiple input points located close to each other. The number of mass points increases as SNR is increased, with a new point emerging at 0. On increasing SNR further, we see that the non zero constellation points (and also the quantizer thresholds) move farther apart, resulting in increased capacity. When the SNR becomes enough that four input points can be disambiguated, the point at 0 disappears, and we get two new points, resulting in a 4-point constellation. The eventual convergence of this 4-point constellation to uniform PAM with mid-point quantizer thresholds (i.e., the benchmark scheme) is to be expected, since the benchmark scheme approaches the capacity bound

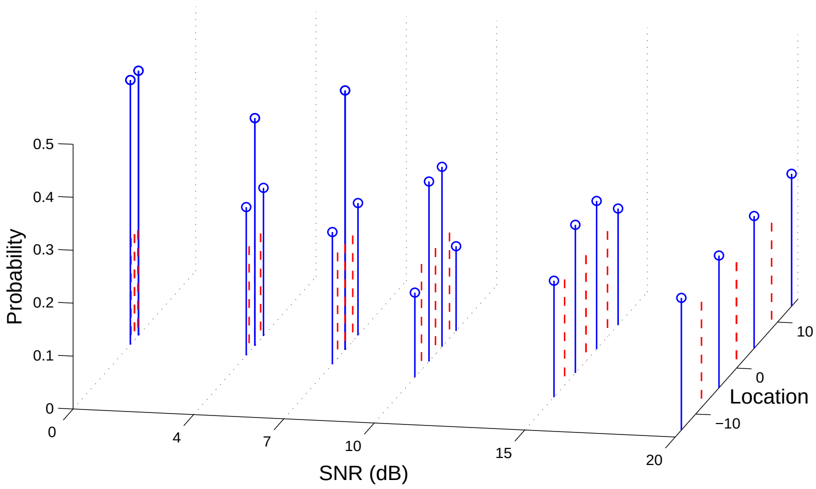

Figure 3.4: 2-bit symmetric quantization : optimal input distribution (solid vertical lines) and quantizer thresholds (dashed vertical lines) at various SNRs.

of two bits at high SNR. It is worth noting that the optimal inputs we obtained all have at most four points, even though Theorem 1 (Section 3.1.2) is looser, guaranteeing the achievability of capacity by at most five points.

3.3.2 3-Bit Quantization

For 3-bit symmetric quantization, we need to optimize over a space of 3 parameters : $\{0 < q_1 < q_2 < q_3\}$, with the quantizer thresholds being $\{0, \pm q_1, \pm q_2, \pm q_3\}$. Since brute force search is computationally complex, we investigate an alternate iterative optimization procedure for joint optimization of the input and the quantizer in this case. Specifically, we begin with an initial quantizer choice Q_1, and then iterate as follows (starting at $i = 1$)

SNR	1-bit ADC		2-bit ADC			SNR	3-bit ADC			UQ
(dB)	OPT	AQNM	OPT	BM	AQNM	(dB)	OPT	BM	AQNM	
-20	0.005	0.007	0.006	0.005	0.007	-20	0.007	0.005	0.007	0.007
-10	0.045	0.067	0.061	0.053	0.068	-10	0.067	0.056	0.069	0.069
-5	0.135	0.185	0.179	0.166	0.195	-5	0.193	0.177	0.197	0.198
0	0.369	0.424	0.455	0.440	0.479	0	0.482	0.471	0.494	0.500
3	0.602	0.610	0.693	0.687	0.736	3	0.759	0.744	0.777	0.791
5	0.769	0.733	0.889	0.869	0.931	5	0.975	0.955	1.002	1.029
7	0.903	0.843	1.098	1.064	1.133	7	1.215	1.180	1.248	1.294
10	0.991	0.972	1.473	1.409	1.417	10	1.584	1.533	1.634	1.730
12	0.992	1.032	1.703	1.655	1.579	12	1.846	1.766	1.886	2.037
15	1.000	1.091	1.930	1.921	1.765	15	2.253	2.138	2.232	2.514
17	1.000	1.115	1.987	1.987	1.853	17	2.508	2.423	2.427	2.838
20	1.000	1.136	1.999	1.999	1.938	20	2.837	2.808	2.655	3.329

Table 3.2: Performance comparison : For 1, 2, and $3-$bit ADC, the table shows the mutual information (in bits per channel use) achieved by the optimal solutions (denoted OPT), as well as the benchmark solutions (denoted BM). Also shown are the capacity estimates obtained by assuming the additive quantization noise model (AQNM). (Note that for 1-bit ADC, the benchmark solution coincides with the optimal solution, and hence is not shown separately.) The last column shows the unquantized AWGN channel's capacity.

- For the quantizer Q_i, find an optimal input. Call this input F_i.
- For the input F_i, find a locally optimal quantizer, initializing the search at Q_i. Call the resulting quantizer Q_{i+1}.
- Repeat the first two steps with $i = i + 1$.

We terminate the process when the capacity gain between consecutive iterations becomes less than a small threshold ϵ.

Although the input-output mutual information is a concave functional of the input distribution (for a fixed quantizer), it is not guaranteed to be concave jointly over the input and the quantizer. Hence, the iterative procedure is not guaranteed to provide an optimal input-quantizer pair in general. A good choice of the initial quantizer Q_1 is crucial to enhance the likelihood that it does converge to an optimal solution. We discuss this next.

High SNR Regime: For high SNRs, we know that uniform PAM with mid-point quantizer thresholds (i.e., the benchmark scheme) is nearly optimal. Hence, this quantizer is a good choice for initialization at high SNRs. The results we obtain indeed demonstrate that this initialization works well at high SNRs. This is seen by comparing the results of the iterative procedure with the results of a brute force search over the quantizer choice (similar to the 2-bit case considered earlier), as both of them provide almost identical capacity values.

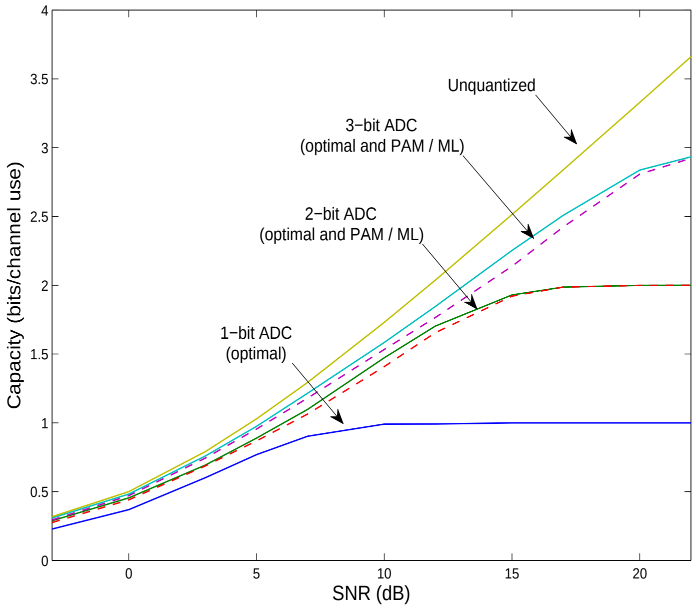

Figure 3.5: Capacity plots for different ADC precisions. For 2 and 3-bit ADC, solid curves correspond to optimal solutions, while dashed curves show the performance of the benchmark scheme (PAM input with ML quantization).

Lower SNRs: For lower SNRs, one possibility is to try different initializations Q_1. However, on trying the benchmark initialization at some lower SNRs as well, we find that the iterative procedure still provides us with near-optimal solutions (again verified by comparing with brute force optimization results).

While our results show that the iterative procedure (with benchmark initialization) has provided (near) optimal solutions at different SNRs, we leave the question of whether it will converge to an optimal solution in general as an open problem.

Comparison with the Benchmark: The efficacy of the benchmark initialization at lower SNRs suggests that the performance of the benchmark scheme should not be too far from optimal at small SNRs as well. This is indeed the case, as seen from the data values in Table 3.2 and the corresponding plots in Figure 3.5. At 0 dB SNR, for instance, the benchmark scheme achieves 98% of the capacity achievable with an optimal input-quantizer pair.

Optimal Input Distributions: Although not depicted here, we again observe (as for the 2-bit case) that the optimal inputs obtained all have at most K points ($K = 8$ in this case), while Theorem 1 guarantees the achievability of capacity by at most $K + 1$ points. Of course, Theorem 1 is applicable to any quantizer choice (and not just optimal symmetric quantizers). Thus, it is possible that there might exist a K-bin quantizer for which the capacity is indeed achieved by exactly $K + 1$ points. We leave open, therefore, the question of whether or not the result in Theorem 1 can be tightened to guarantee the achievability of capacity with at most K points for the AWGN-QO channel.

3.3.3 Comparison with Unquantized Observations

We now compare the capacity results for different quantizer precisions against the capacity with unquantized observations. Again, the plots are shown in Figure 3.5 and the data values are given in Table 3.2. We observe that at low SNR, the performance degradation due to low-precision quantization is small. For instance, at -5 dB SNR, 1-bit receiver quantization achieves 68% of the capacity achievable without any quantization, while with 2-bit quantization, we can get as much as 90% of the unquantized capacity. Even at moderately high SNRs, the loss due to low-precision quantization remains quite acceptable. For example, 2-bit quantization achieves 85% of the capacity attained using unquantized observations at 10 dB SNR, while 3-bit quantization achieves 85% of the unquantized capacity at 20 dB SNR. For the specific case of binary antipodal signaling, [45] has earlier shown that a large fraction of the capacity can be obtained by 2-bit quantization.

On the other hand, if we fix the spectral efficiency to that attained by an unquantized system at 10 dB (which is 1.73 bits/channel use), then 2-bit quantization incurs a loss of 2.30 dB (see Table 3.3). For wideband systems, this penalty in power maybe more significant compared to the 15% loss in spectral efficiency on using 2-bit quantization at 10 dB SNR. This suggests, for example, that in order to weather the impact of low-precision ADC, a moderate reduction in the spectral efficiency might be a better design choice than an increase in the transmit power.

	Spectral Efficiency (bits per channel use)				
	0.25	0.5	1.0	1.73	2.5
1-bit ADC	−2.04	1.79	–	–	–
2-bit ADC	−3.32	0.59	6.13	12.30	–
3-bit ADC	−3.67	0.23	5.19	11.04	16.90
Unquantized	−3.83	0.00	4.77	10.00	14.91

Table 3.3: SNR (in dB) required to achieve a specified spectral efficiency with different ADC precisions.

3.3.4 Additive Quantization Noise Model

It is common to model the quantization noise as independent additive noise [63, pp. 122]. Next, we compare this approximation with our exact capacity calculations. In this model $Y = X + N + N_Q$, where the quantization noise N_Q is assumed to be uniformly distributed, and independent of X, N. The signal to quantization noise ratio $\frac{P}{\mathbb{E}({N_Q}^2)}$ is assumed to be $6 \log_2 K$ dB for K-bin quantization [63, pp. 122]. As $\mathsf{SNR} \rightarrow 0$, the distribution of $N + N_Q$ approaches that of a Gaussian, and hence we expect

$$\frac{1}{2} \log \left(1 + \frac{P}{\sigma^2 + \mathbb{E}({N_Q}^2)} \right)$$

to be a good approximation of the capacity at low SNR. Table 3.2 shows that this approximation can be useful in terms of providing a quick estimate, although it can either underestimate or overestimate the actual capacity, depending on the parameters.

3.4 Conclusions

Our Shannon-theoretic investigation indicates that the use of low-precision ADC may be a feasible option for designing future high-bandwidth communication systems. At low to moderate SNR, which would be the preferred regime of operation given that the bandwidth is large, the small loss in spectral efficiency due to 2-3 bit ADC can be quite acceptable.

The observation that standard uniform PAM input, with ML receiver quantization is near-optimal (at all SNRs) is also quite encouraging, due to the ease of implementation: both the PAM input points and ML thresholds have simple analytical dependence on the SNR, eliminating the need for any complicated optimization. Of course, accurate measurement of the SNR at the receiver is still predicated on the reliable performance of the automatic gain control (AGC).

Given the encouraging nature of the these results, our next step is to go about trying to achieve them. This involves algorithm design for carrier and timing synchronization, AGC, as well as devising coding and decoding mechanisms. Given that turbo-like codes are now available for a wide variety of channels and rates, it would be safe to presume that (with some ingenuity) coding schemes that approach the capacity for our quantized channel can be devised. What is crucial, therefore, is whether we can do reliable synchronization with low-precision quantization or not. Towards that, in the next chapter, we consider the problem of carrier synchronization.

Before proceeding further, we list some open technical questions based on this chapter.

3.4.1 Open Technical Issues

There are some unresolved technical issues that we leave as open problems. While we show that at most $K+1$ points are needed to achieve capacity for K-bin output quantization of the AWGN channel, our numerical results reveal that K mass points are sufficient. Can this be proven analytically, at least for symmetric quantizers ? Are symmetric quantizers optimal ? Does our iterative procedure (with the benchmark initialization, or some other judicious initialization) for joint optimization of the input and the quantizer converge to an optimal solution in general ?

A technical assumption worth revisiting is that of Nyquist sampling (which induces the discrete-time memoryless AWGN-Quantized Output channel model considered in this work). While symbol rate Nyquist sampling is optimal for unquantized systems in which the transmit and receive filters are square root Nyquist and the channel is ideal, for quantized samples, we have obtained numerical results that show that fractionally spaced samples can actually lead to small performance gains. A detailed study quantifying such gains is important in understanding the tradeoffs between ADC speed and precision. However, we do not expect oversampling to play a significant role at low to moderate SNR, given the small degradation in our Nyquist sampled system relative to unquantized observations (for which Nyquist sampling is indeed optimal) in these regimes. Of course, oversampling in conjunction with hybrid analog/digital processing (e.g., using ideas analogous to delta-sigma quantization) could produce bigger performance gains.

Chapter 4

Carrier Synchronization with Low-Precision ADC

We now turn our attention to the problem of carrier synchronization with low-precision ADC. While, in principle, the local oscillator (LO) at the receiver can be locked to the frequency of the incoming carrier wave using a classical analog feedback loop, we stick to the modern DSP-centric design approach, in which all the processing happens at the baseband, and is mostly digital. Thus, we assume that the LO at the receiver runs at a fixed frequency, independent of the incoming carrier wave.

Denoting the frequency offset between the LO and the incoming wave by Δf, the received Nyquist-rate complex baseband sample at time n is

$$Y_n = X_n e^{j(2\pi \Delta f n T_s + \theta_o)} + W_n \ ,$$

where X_n is the transmitted symbol, W_n is complex Gaussian noise, T_s is the symbol interval, and θ_o is the initial phase offset between the LO and the incom-

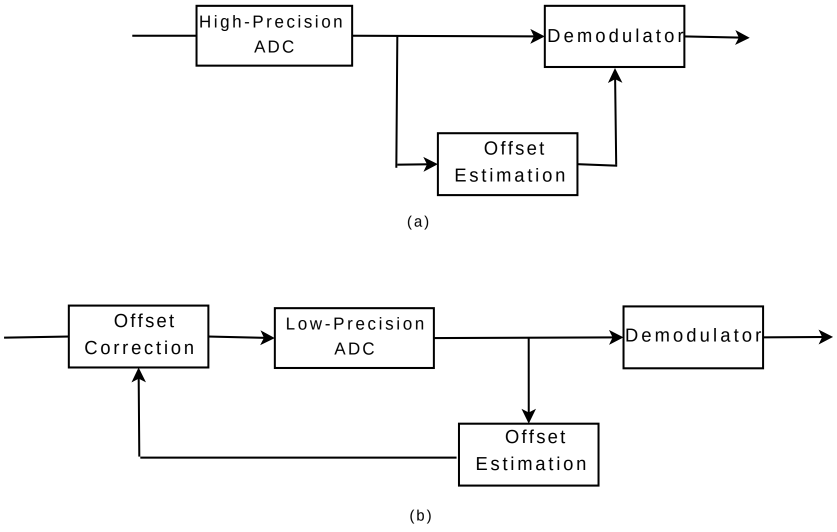

Figure 4.1: Correction for frequency offset. (a) For high-precision ADC, the correction is done in the digital domain. (b) For low-precision ADC, it may be possible to perform analog offset correction based on digital feedback.

ing wave. Both Δf and θ_o are unknown *a priori*. When the received symbols Y_n are sampled with high-precision, the standard approach to recover the transmitted symbols X_n, termed coherent demodulation, is to estimate the unknown parameters Δf and θ_o, derotate Y_n to get $Y_n e^{-j(2\pi \Delta f n T_s + \theta_o)}$ and use it to estimate X_n (Fig. 4.1(a)). The estimation of Δf and θ_o is typically based on an initial training sequence, followed by some mechanism to track the slow variations due to the drift in the frequency of the LO [64, pp. 155] However, if we now consider drastic quantization of the received samples Y_n, the feasibility of this estimate and correct approach becomes questionable. For instance, if the LO has a frequency offset of 100 ppm (parts per million) (i.e., $\Delta f = \frac{100 f_c}{10^6}$), and the bandwidth is

10 % of the carrier frequency (so that $\frac{1}{T_s} = 0.1 f_c$), then the phase rotation from one symbol to the next is ≈ 0.006 radians. With drastically quantized samples, estimating and correcting for such a small rate of phase change might be tough. While this issue is still open, one plausible way to circumvent the problem, that has indeed been explored recently in the literature as well [19], is to regress back into the analog domain and do the offset compensation before the ADC, based on feedback of the estimates generated by post-ADC DSP (Fig. 4.1(b)).

Given the seeming difficulty of explicit estimation and correction of the offsets with low-precision ADC (at least for a "mostly-digital" design), we now concentrate attention on an alternate noncoherent (or differentially coherent) approach, which exploits the fact that the phase offset $2\pi \Delta f n T_s + \theta_o$, although *a priori* unknown, can be assumed to be constant over consecutive symbols. Under this assumption, the transmitter can encode information in the phase *difference* across consecutive symbols, allowing successful recovery at the receiver even when there is no absolute phase reference. This forms the basis for the well-known modulation technique, differential phase modulation. While this approach does not require any explicit training, it does incur a loss in performance compared to a coherent receiver [64, pp. 173]. However, recent work has shown that this loss can be recovered under the assumption that the unknown phase offset is constant over a larger block of symbols (rather than just two symbols). In this chapter, we investigate the performance of this *block noncoherent* approach under low-precision output quantization.

We first provide background on the related literature.

4.1 Block Noncoherent Communication

Divsalar and Simon [65] were the first to point out the gains that could be achieved by using blocks of length greater than two. The complexity of the maximum likelihood detector used to achieve the gains however grows exponentially in the block length; we must make a joint decision on the block of input symbols, so that the cardinality of the space over which we do the optimization is exponential in the block length. Fortunately, later research showed that the detection could be performed with lower complexity; Mackenthun in [66] showed that the optimal solution could be attained with linear-logarithmic complexity in the block length, i.e., complexity of order $L \log_2 L$ for block length L, (a similar result was presented by Sweldens in [67]), Warrier and Madhow [68] provided a near optimal solution with complexity growing only linearly in the block length. From an information theoretic perspective, Peleg and Shamai obtained the capacity of the block noncoherent channel in [69], assuming, for analytical tractability, that the phase offset varies independently from one block to the next.

Note that while we are only concerned with a phase offset (induced by the asynchronous LO) in this work, the block noncoherent model extends to the setting of communication over narrowband slow fading channels as well, where the channel state, although unknown, can be assumed to be constant over a block of symbols. Indeed, this block fading model has been investigated extensively in the recent literature, ranging from capacity analysis [70], to efficient capacity approaching architectures ([71, 72], and references therein).

We now proceed to our investigation of block noncoherent communication with low-precision quantization at the receiver. There are several ways to quantize a complex-valued received signal. As illustrated in the next section, *phase quantization* can be implemented efficiently using 1-bit ADCs and analog pre-multipliers, and eliminates the need for any automatic gain control (since no amplitude information is used). Moreover, for phase constellations, we expect phase quantization to work well. Hence, we evaluate the performance under phase quantization. As with the previous literature, we assume, for analytical tractability, that the unknown phase is constant over a block of L symbols, but varies independently across blocks.

Notation: Throughout the chapter, we denote random variables by capital letters, and the specific value they take using small letters. Bold faced notation is used to denote vectors of random variables. $\mathbb{E}$ is the expectation operator.

4.2 Channel Model and Receiver Architecture

The received signal over a block of length L, after quantization is represented as

$$Z_l = \mathsf{Q}(S_l e^{j\Phi} + N_l)\ ,\ l = 0, 1, \cdots, L-1, \tag{4.1}$$

where,

- $\mathbf{S} := [S_0\ S_1\ \cdots\ S_{L-1}]$ is the transmitted vector,
- Φ is an unknown constant with uniform distribution on $[0, 2\pi)$,
- $\mathbb{N} := [N_0\ \cdots\ N_{L-1}]$ is a vector of i.i.d. complex Gaussian noise with variance $\sigma^2 = N_0/2$ in each dimension,
- $\mathsf{Q} : \mathbb{C} \to \mathcal{K} = \{0, 1, \cdots, K-1\}$ denotes a quantization function that maps each point in the complex plane to one of the K quantization indices, and
- $\mathbf{Z} := [Z_0\ Z_1\ \cdots\ Z_{L-1}]$ is the vector of quantized received symbols, so that each $Z_l \in \mathcal{K}$.

Each S_l is picked in an i.i.d. manner from a uniform M-PSK constellation denoted by the set of points $\mathcal{A} = \{e^{j\theta_0}, e^{j\theta_1}, \cdots, e^{j\theta_{M-1}}\}$, where $\theta_m = (\theta_{m-1} + \frac{2\pi}{M})$ [1], for $m = 1, 2, \cdots, M-1$.

[1] Unless stated otherwise, any arithmetic operations for phase angles are assumed to be performed modulo 2π. For the output symbols Z_l, the arithmetic is modulo K, while for the input symbols X_l (introduced immediately after in the text), it is modulo M.

We now introduce the random vector $\mathbf{X} = [X_0 \;\; X_1 \;\; \cdots \;\; X_{L-1}]$, with each X_i picked in an i.i.d. manner from a uniform distribution on the set $\{0, 1, \cdots, M-1\}$. Our channel model (4.1) can now equivalently be written as

$$Z_l = \mathsf{Q}(e^{j\theta x_l} e^{j\Phi} + N_l) \ , \ l = 0, 1, \cdots, L-1 \ , \tag{4.2}$$

with every output symbol $Z_l \in \{0, 1, \cdots, K-1\}$ as before, and every input symbol $X_l \in \{0, 1, \cdots, M-1\}$. The set of all possible input vectors is denoted by $\mathcal{X}$, while $\mathcal{Z}$ denotes the set of all possible output vectors.

We consider K-bin (or K-sector) phase quantization: our quantizer divides the interval $[0, 2\pi)$ into K equal parts, and the quantization indices go from 0 to $K-1$ in the counter-clockwise direction. Fig. 4.2(b) depicts the scenario for K=8. Thus, our quantization function is $\mathsf{Q}(c) = \lfloor \arg(c) | (\frac{2\pi}{K}) \rfloor$, where $c \in \mathbb{C}$, and $\lfloor p \rfloor$ denotes the greatest integer less than or equal to p. Such phase quantization can be implemented using 1-bit ADCs preceded by analog multipliers which provide linear combinations of the I and Q channel samples. For instance, employing 1-bit ADC on I and Q channels results in uniform 4-sector phase quantization, while uniform 8-sector quantization can be achieved simply by adding two new linear combinations, I+Q and I-Q, corresponding to a $\pi/4$ rotation of I/Q axes (no analog multipliers needed in this case), as shown in Fig. 4.2(a).

We begin our investigation by studying the inherent symmetry in the relationship between the channel input and output. This study provides us several

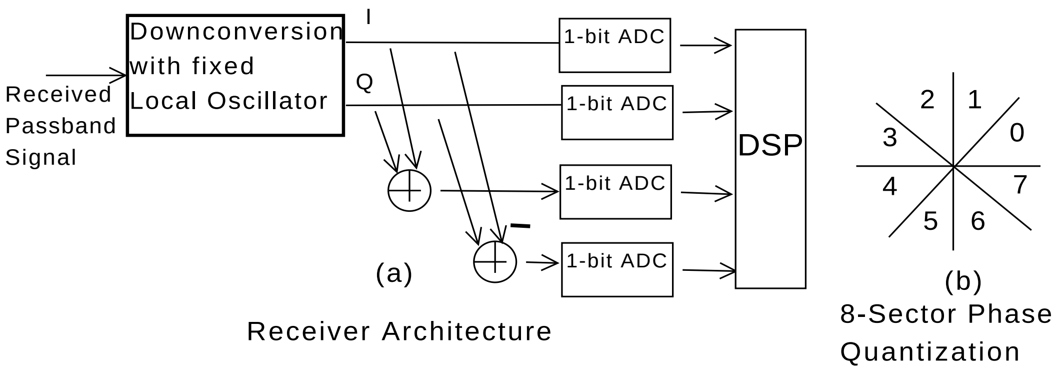

Figure 4.2: Receiver architecture for 8-sector quantization.

results that govern the structure of the output probability distribution, both conditioned on the input (i.e., $\mathsf{P}(\mathbf{Z}|\mathbf{X})$), and without conditioning (i.e., $\mathsf{P}(\mathbf{Z})$). These distributions are integral to computing the channel capacity (one of our focuses in this chapter), as well as for soft decision decoding (not considered here). While brute force computation (computing $\mathsf{P}(\mathbf{z}|\mathbf{x})$ for every $\mathbf{z} \in \mathcal{Z}$ and every $\mathbf{x} \in \mathcal{X}$) of these distributions has exponential complexity in the block length, our results allow their computation with significant reduction in the complexity.

Note: Throughout the chapter, we assume that the PSK constellation size M, and the number of quantization bins K, are such that $K = aM$ for some positive integer a. We illustrate our results with the running example of QPSK with 8-sector quantization, depicted in Fig. 4.3(a)

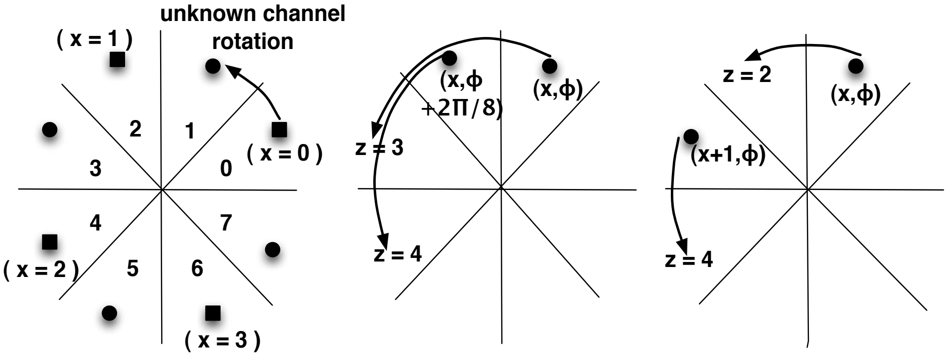

Figure 4.3: QPSK with 8-sector quantization (i.e., M=4, K=8). a) depicts how the unknown channel phase ϕ results in a rotation of the transmitted symbol (square : original constellation , circle : rotated constellation). (b) and (c) depict the circular symmetry induced in the conditional probability $\mathsf{P}(z|x,\phi)$ due to the circular symmetry of the complex Gaussian noise. (b) shows that increasing ϕ by $2\pi/K = (\pi/4)$ and z by 1 will keep the conditional probability unchanged, i.e., $\mathsf{P}(z = 3|x,\phi) = \mathsf{P}(z = 4|x,\phi + 2\pi/K)$. (c) shows that increasing x by 1 and z by 2 = (K/M) will keep the conditional probability unchanged, i.e., $\mathsf{P}(z = 2|x,\phi) = \mathsf{P}(z = 4|x + 1,\phi)$.

4.3 Input-Output Relationship

Conditioned on the channel phase Φ, $\mathsf{P}(\mathbf{Z}|\mathbf{X},\Phi)$ is a product of individual symbol probabilities $\mathsf{P}(Z_l|X_l,\Phi)$. We therefore begin by analyzing the symmetries in the latter.

4.3.1 Properties of $\mathsf{P}(Z_l|X_l,\Phi)$

We have that $\mathsf{P}(z_l|x_l,\phi)$ is the probability that $\arg(e^{j(\theta_{x_l}+\phi)} + N_l)$ belongs to the interval $[\frac{2\pi}{K}z_l \;\; \frac{2\pi}{K}(z_l + 1))$. In other words, it is the probability that the complex Gaussian noise N_l takes the point $e^{j(\theta_{x_l}+\phi)}$ on the unit circle, to another point whose phase belongs to $[\frac{2\pi}{K}z_l \;\; \frac{2\pi}{K}(z_l+1))$. Due to the circular symmetry of the complex Gaussian noise, this is the same as the probability that N_l takes the point $e^{j(\theta_{x_l}+\phi+\frac{2\pi}{K}i)}$ on the unit circle, to another point whose phase belongs to $[\frac{2\pi}{K}(z_l+i) \;\; \frac{2\pi}{K}(z_l+1+i))$, where i is an integer. We thus get our first two results.

Property A-1: $\mathsf{P}(z_l|x_l,\phi) = \mathsf{P}(z_l+i|x_l,\phi+i\frac{2\pi}{K})$.

Property A-2: $\mathsf{P}(z_l|x_l,\phi) = \mathsf{P}(z_l+ia|x_l+i,\phi)$.

Note that $\theta_{x_l+i} = \theta_{x_l} + \frac{2\pi}{M}i = \theta_{x_l} + \frac{2\pi}{K}(ia)$, which gives Property *A-2*.

Property *A*-2 simply states that if we jump from one point in the M-PSK constellation to the next, then we must jump $a = \frac{K}{M}$ quantization sectors in order to keep the conditional probability invariant. This is intuitive, since the separation between consecutive points in the input constellation is $2\pi/M$, while each quantization sector covers an angle of $2\pi/K$. For QPSK with $K = 8$, Fig. 4.3(b) and 4.3(c) depict example scenarios for the two properties.

If we put $i = -x_l$ in Property *A-2*, we get the following special case, which relates the conditioning on a general x_l to the conditioning on 0.

Property A-3: $\mathsf{P}(z_l|x_l,\phi) = \mathsf{P}(z_l-ax_l|0,\phi)$.

To motivate our final property, we consider our example of QPSK with $K = 8$. While we have 8 distinct quantization sectors, if we look at Fig. 2(a), the orientation of these 8 sectors relative to the 4 constellation points (shown as squares) can be described by dividing the sectors into 2 groups : $\{0,2,4,6\}$, and $\{1,3,5,7\}$. For instance, the positioning of the first sector ($z = 0$) w.r.t. $x = 0$ is identical to the positioning of the third sector ($z = 2$) w.r.t. $x = 1$ (and similarly $z = 4$ w.r.t $x = 2$, and $z = 6$ w.r.t $x = 3$). On the other hand, the positioning of the second sector ($z = 1$) w.r.t. $x = 0$ is identical to the positioning of the fourth sector ($z = 3$) w.r.t. $x = 1$ (and similarly $z = 5$ w.r.t $x = 2$, and $z = 7$ w.r.t $x = 3$). In terms of the conditional probabilities, this implies, for example, that we will have $\mathsf{P}(z_l = 7|x_l = 3, \phi) = \mathsf{P}(z_l = 1|x_l = 0, \phi)$, and similarly, $\mathsf{P}(z_l = 6|x_l = 3, \phi) = \mathsf{P}(z_l = 0|x_l = 0, \phi)$. In general, we can relate the conditional probability of every odd z_l with that of $z_l = 1$, and similarly of every even z_l with that of $z_l = 0$, with corresponding rotations of the symbol x_l. For general values of K and M, the number of groups equals $a = \frac{K}{M}$, and we can relate the probability of any z_l with that of z_l mod a.

Property A-4: Let $z_l = q_l a + r_l$, where q_l is the quotient on dividing z_l by a, and r_l is the remainder, i.e, $r_l = z_l \bmod a$. Then, $\mathsf{P}(z_l|x_l, \phi) = \mathsf{P}(z_l \bmod a|x_l - q_l, \phi)$.

While this result follows directly from Property A-2 by putting $i = -q_l$, it is an important special case, as it enables us to restrict attention to only the first a sectors ($Z_l \in \{0, 1, \cdots, a-1\}$), rather than having to work with all the K sectors. As detailed later, this leads to significant complexity reduction in capacity computation.

We now use these properties to present results for $\mathsf{P}(\mathbf{Z}|\mathbf{X})$.

4.3.2 Properties of $\mathsf{P}(\mathbf{Z}|\mathbf{X})$

Property B-1: Let $\mathbf{1}$ denote the row vector with all entries as 1. Then $\mathsf{P}(\mathbf{z}|\mathbf{x}) = \mathsf{P}(\mathbf{z}+i\mathbf{1}|\mathbf{x})$.

Proof: For a fixed $\mathbf{x}$, increasing each z_l by the same number i leaves the conditional probability unchanged, because the phase Φ in the channel model (4.1) is uniformly distributed in $[0, 2\pi)$. A detailed proof follows. We have

$$\begin{aligned}\mathsf{P}(\mathbf{z}|\mathbf{x}) &= \mathbb{E}_\Phi\left(\mathsf{P}(\mathbf{z}|\mathbf{x},\Phi)\right) = \mathbb{E}_\Phi\left(\prod_{l=0}^{L-1}\mathsf{P}(z_l|x_l,\Phi)\right)\\ &= \mathbb{E}_\Phi\left(\prod_{l=0}^{L-1}\mathsf{P}(z_l+i|x_l,\Phi+i\frac{2\pi}{K})\right)\\ &= \mathbb{E}_{\hat{\Phi}}\left(\prod_{l=0}^{L-1}\mathsf{P}(z_l+i|x_l,\hat{\Phi})\right)\\ &= \mathbb{E}_{\hat{\Phi}}\left(\mathsf{P}(\mathbf{z}+i\mathbf{1}|\mathbf{x},\hat{\Phi}))\right) = \mathsf{P}(\mathbf{z}+i\mathbf{1}|\mathbf{x}).\end{aligned}$$

The second equality follows by the fact that the components of $\mathbf{Z}$ are independent conditioned on $\mathbf{X}$ and Φ. Property A-1 gives the third equality. A change of variables, $\hat{\Phi} = \Phi + i\frac{2\pi}{K}$ gives the fourth equality (since Φ is uniformly distributed on $[0, 2\pi)$, so is $\hat{\Phi}$), thereby completing the proof. ■

Remark 1: For the rest of the chapter, we refer to the operation $\mathbf{z} \rightarrow \ \mathbf{z} + i\mathbf{1}$ as *constant addition.*

Our next result concerns the observation that the conditional probability remains invariant under an *identical* permutation of the components of the vectors $\mathbf{z}$ and $\mathbf{x}$.

Property B-2: Let Π denote a permutation operation, and $\Pi\mathbf{x}$ ($\Pi\mathbf{z}$) the vector obtained on permuting $\mathbf{x}$ ($\mathbf{z}$) under this operation. Then, $\mathsf{P}(\mathbf{z}|\mathbf{x}) = \mathsf{P}(\Pi\mathbf{z}|\Pi\mathbf{x})$.

Proof: As in the proof of Property 1, the idea is to condition on Φ and work with the symbol probabilities $\mathsf{P}(z_l|x_l, \Phi)$. Consider $\mathsf{P}(\mathbf{z}|\mathbf{x}, \Phi) = \prod_{l=0}^{L-1} \mathsf{P}(z_l|x_l, \Phi)$, and $\mathsf{P}(\Pi\mathbf{z}|\Pi\mathbf{x}, \Phi) = \prod_{l=0}^{L-1} \mathsf{P}((\Pi\mathbf{z})_l|(\Pi\mathbf{x})_l, \Phi)$. Since multiplication is a commutative operation, we have $\mathsf{P}(\mathbf{z}|\mathbf{x}, \Phi) = \mathsf{P}(\Pi\mathbf{z}|\Pi\mathbf{x}, \Phi)$. Taking expectation w.r.t. Φ completes the proof. ■

The next two results extend properties A-3 and A-4.

Property B-3: Define the input vector $\mathbf{x}_0 = [0 \cdots 0]$. Then, $\mathsf{P}(\mathbf{z}|\mathbf{x}) = \mathsf{P}(\mathbf{z} - a\mathbf{x}|\mathbf{x}_0)$, where $a = \frac{K}{M}$, and the subtraction is performed modulo K.

Property B-4: Let $z_l = q_l a + r_l$, where q_l is the quotient on dividing z_l by a, and r_l is the remainder, i.e, $r_l = z_l \bmod a$. Define $\mathbf{q} = [q_0, \cdots, q_{L-1}]$, and, $\mathbf{z} \bmod a = [z_0 \bmod a \quad \cdots \quad z_{L-1} \bmod a]$. Then $\mathsf{P}(\mathbf{z}|\mathbf{x}) = \mathsf{P}(\mathbf{z} \bmod a \mid \mathbf{x} - \mathbf{q})$.

Proofs: The properties follow from A-3 and A-4 respectively, by first noting that the vector probability $\mathsf{P}(\mathbf{z}|\mathbf{x}, \Phi)$ is the product of the scalar ones, and then integrating over Φ . ■

4.3.3 Properties of $\mathsf{P}(\mathbf{Z})$

We now consider the unconditional distribution $\mathsf{P}(\mathbf{z})$. The first result states that $\mathsf{P}(\mathbf{z})$ is invariant under constant addition.

Property C-1: $\mathsf{P}(\mathbf{z}) = \mathsf{P}(\mathbf{z} + i\mathbf{1})$.

Proof: Using Property B-1, this follows directly by taking expectation over $\mathbf{X}$ on both sides. ■

On similar lines, we now extend Property B-2 to show that $\mathsf{P}(\mathbf{z})$ is invariant under any permutation of $\mathbf{z}$.

Property C-2: $\mathsf{P}(\mathbf{z}) = \mathsf{P}(\Pi\mathbf{z})$.

Proof: We have $\mathsf{P}(\mathbf{z}) = \frac{1}{M^L}\sum_{\mathbf{x}\in\mathcal{X}} \mathsf{P}(\mathbf{z}|\mathbf{x})$. Using Property B-2, we get $\mathsf{P}(\mathbf{z}) = \frac{1}{M^L}\sum_{\mathbf{x}\in\mathcal{X}} \mathsf{P}\left(\Pi\mathbf{z}|\Pi\mathbf{x}\right)$. Since Π is just a permutation operation, every unique choice of $\mathbf{x} \in \mathcal{X}$ results in a unique $\Pi\mathbf{x} \in \mathcal{X}$. Hence, we can rewrite the last equation as $\mathsf{P}(\mathbf{z}) \;=\; \frac{1}{M^L}\sum_{\mathbf{x}\in\mathcal{X}} \mathsf{P}(\Pi\mathbf{z}|\mathbf{x}) \;=\; \mathsf{P}(\Pi\mathbf{z})$. ■

Our final result extends Property B-4.

Property C-3: Let $a = \frac{K}{M}$. Then $\mathsf{P}(\mathbf{z}) = \mathsf{P}(\mathbf{z} \bmod a)$.

Proof: Using the same notation as in Property B-4, we have $\mathsf{P}(\mathbf{z}|\mathbf{x}) = \mathsf{P}(\mathbf{z} \bmod a \mid \mathbf{x} - \mathbf{q})$. Noting that the transformation $\mathbf{x} \rightarrow \mathbf{x} - \mathbf{q}$ is a one-to-one mapping, the proof follows on the same lines as the proof of Property C-2. ■

Example: For QPSK with $K = 8$ and $L = 4$, $\mathsf{P}(z = [5\ 7\ 2\ 4]) = \mathsf{P}(z = [1\ 1\ 0\ 0])$.

We now apply these results for low complexity capacity computation.

4.4 Efficient Capacity Computation

We wish to compute the mutual information

$$I(\mathbf{X};\mathbf{Z}) = H(\mathbf{Z}) - H(\mathbf{Z}|\mathbf{X}).$$

We first discuss computation of the conditional entropy.

4.4.1 Conditional Entropy

We have $H(\mathbf{Z}|\mathbf{X}) = \sum_{\mathcal{X}} H(\mathbf{Z}|\mathbf{x})\mathsf{P}(\mathbf{x})$, where $H(\mathbf{Z}|\mathbf{x}) = -\sum_{\mathcal{Z}} \mathsf{P}(\mathbf{z}|\mathbf{x}) \log \mathsf{P}(\mathbf{z}|\mathbf{x})$ is the entropy of the output when the input vector $\mathbf{X}$ takes on the specific value $\mathbf{x}$. Our main result in this section is that $H(\mathbf{Z}|\mathbf{x})$ is constant $\forall\mathbf{x}$.

Property D-1: $H(\mathbf{Z}|\mathbf{x})$ is a constant.

Proof: We show that for any input vector $\mathbf{x}, H(\mathbf{Z}|\mathbf{x}) = H(\mathbf{Z}|\mathbf{x}_0)$, where $\mathbf{x}_0 = [0 \cdots 0]$ as defined before. We have

$$\begin{aligned} H(\mathbf{Z}|\mathbf{x}) &= -\sum_{\mathcal{Z}} \mathsf{P}(\mathbf{z}|\mathbf{x}) \log \mathsf{P}(\mathbf{z}|\mathbf{x}) \\ &= -\sum_{\mathcal{Z}} \mathsf{P}(\mathbf{z} - a\mathbf{x}|\mathbf{x}_0) \log \mathsf{P}(\mathbf{z} - a\mathbf{x}|\mathbf{x}_0)\ , \end{aligned} \tag{4.3}$$

where the second equality follows from Property B-3. Now, since $\mathbf{z} \rightarrow \mathbf{z} - a\mathbf{x}$ is just a subtraction operation, it is easy to see that every unique choice of $\mathbf{z} \in \mathcal{Z}$ results in a unique choice of $\mathbf{z} - a\mathbf{x} \in \mathcal{Z}$. Hence, we can rewrite (4.3) as

$$H(\mathbf{Z}|\mathbf{x}) = -\sum_{\mathcal{Z}} \mathsf{P}(\mathbf{z}|\mathbf{x}_0) \log \mathsf{P}(\mathbf{z}|\mathbf{x}_0) = H(\mathbf{Z}|\mathbf{x}_0) \tag{4.4}$$

■

Thus, $H(\mathbf{Z}|\mathbf{X}) = H(\mathbf{Z}|\mathbf{x}_0)$, but brute force computation of $H(\mathbf{Z}|\mathbf{x}_0)$ still has exponential complexity, $\mathsf{P}(\mathbf{Z}|\mathbf{x}_0)$ must be computed for each of the K^L possible output vectors $\mathbf{Z}$. However, we show that it suffices to compute $\mathsf{P}(\mathbf{Z}|\mathbf{x}_0)$ for a much smaller set of $\mathbf{Z}$ vectors.

Using Property B-2, we have $\mathsf{P}(\mathbf{z}|\mathbf{x}_0) = \mathsf{P}(\Pi\mathbf{z}|\Pi\mathbf{x}_0)$. Since $\mathbf{x}_0 = [0..0]$, any permutation of $\mathbf{x}_0$ gives back $\mathbf{x}_0$. Hence, $\mathsf{P}(\mathbf{z}|\mathbf{x}_0) = \mathsf{P}(\Pi\mathbf{z}|\mathbf{x}_0)$. Combined with Property B-1, we thus get that it suffices to compute $\mathsf{P}(\mathbf{z}|\mathbf{x}_0)$ for a set of vectors $S_{\mathbf{Z}}$ in which no vector can be obtained from another by performing the joint operations of constant addition and permutation. We do not have an exact method to get $S_{\mathbf{Z}}$, but can resort to a sub-optimal procedure, which still provides significant complexity reduction. Instead of jointly accounting for constant addition and permutation, we first account for constant addition, and then for permutation. Specifically, we first note that using Property B-1, it suffices to compute $\mathsf{P}(\mathbf{z}|\mathbf{x}_0)$ only for a set of vectors $S_{\mathbf{Z}_1}$ for which the first symbol is 0. Next, using the fact that $\mathsf{P}(\mathbf{z}|\mathbf{x}_0) = \mathsf{P}(\Pi\mathbf{z}|\mathbf{x}_0)$, within the set $S_{\mathbf{Z}_1}$, we can further restrict attention to a subset $S_{\mathbf{Z}_2}$ in which no vector can be obtained from another one by a permutation operation. Since permutations don't matter, all we are interested in is how many symbols of each type are picked, so that obtaining the set $S_{\mathbf{Z}_2}$ is simply equivalent to the well-known problem of distributing L–1 identical balls into K distinct boxes, with empty boxes allowed. The number of ways to do this is $C(K + L - 2, L - 1)$, and each of these combinations can be obtained easily using standard known procedures. For $K = 8$, and $L = \{3, 4, 5, 6, 7\}$, the cardinality of $S_{\mathbf{Z}_2}$ is $\{36, 120, 330, 792, 1716\}$, whereas the exponential figure K^L is $\{512, 4096, 32768, 2.6 \times 10^5, 2.1 \times 10^6\}$, illustrating the large reduction in complexity.

Once we have the set $S_{\mathbf{Z}_2}$, we can numerically compute the probability $\mathsf{P}(\mathbf{z}|\mathbf{x}_0)$ for every vector in $S_{\mathbf{Z}_2}$. The entropy $H(\mathbf{Z}|\mathbf{x}_0)$ can then be obtained as follows. For $\mathbf{z} \in S_{\mathbf{Z}_2}$, let $n(\mathbf{z})$ denote the number of distinct permutations that can be generated from it, while keeping the first symbol fixed. This is simply equal to $\frac{(L-1)!}{\prod_{i=0}^{K-1} r_i}$, where r_i is the number of times the symbol i occurs in $\mathbf{z}$. The conditional entropy then is $H(\mathbf{Z}|\mathbf{x}_0) = -\sum\limits_{\mathcal{Z}} \mathsf{P}(\mathbf{z}|\mathbf{x}_0) \log \mathsf{P}(\mathbf{z}|\mathbf{x}_0) = -\sum\limits_{S_{\mathbf{Z}_1}} K\mathsf{P}(\mathbf{z}|\mathbf{x}_0) \log \mathsf{P}(\mathbf{z}|\mathbf{x}_0) = -\sum\limits_{S_{\mathbf{Z}_2}} K\ n(\mathbf{z})\ \mathsf{P}(\mathbf{z}|\mathbf{x}_0) \log \mathsf{P}(\mathbf{z}|\mathbf{x}_0)$.

4.4.2 Output Entropy

The output entropy is $H(\mathbf{Z}) = -\sum_{\mathcal{Z}} \mathsf{P}(\mathbf{z}) \log \mathsf{P}(\mathbf{z})$. A brute force computation requires us to know $\mathsf{P}(\mathbf{z})\ \forall \mathbf{z} \in \mathcal{Z}$, which clearly has exponential complexity. However, using Properties C-1, C-2 and C-3, we get that it is sufficient to compute $\mathsf{P}(\mathbf{z})$ for a set of vectors $\tilde{S}_{\mathbf{Z}}$ in which no vector can be obtained from another one by performing the operations of constant addition and permutation, and also, the vector components $\in \{0, 1, \cdots, a-1\}$. This is similar to the situation we encountered earlier in the last subsection, except that the vector components there were allowed to be in $\{0, 1, \cdots, K-1\}$. To exploit this for further complexity reduction, we can begin by defining the set $\tilde{\mathcal{Z}}$ to be the set of vectors in which the vector components take values in $\{0, 1, \cdots, a-1\}$ only. Since $\mathsf{P}(\mathbf{z}) = \mathsf{P}(\mathbf{z} \bmod a)$, a moment's thought reveals that each vector in $\tilde{\mathcal{Z}}$ has the same probability as a set of $(\frac{K}{a})^L = M^L$ distinct vectors in $\mathcal{Z}$, and the sets corresponding to different vectors are disjoint. Thus $H(Z) = -\sum\limits_{\mathcal{Z}} \mathsf{P}(\mathbf{z}) \log \mathsf{P}(\mathbf{z}) = -M^L \sum\limits_{\tilde{\mathcal{Z}}} \mathsf{P}(\mathbf{z}) \log \mathsf{P}(\mathbf{z})$. To obtain $\{\mathsf{P}(\mathbf{z})\}$ for $\mathbf{z} \in \tilde{\mathcal{Z}}$, we can follow exactly the same procedure as described in the last subsection, with K being replaced by a. In particular, we need to compute $\mathsf{P}(\mathbf{z})$ only for $C(a + L - 2, L - 1)$ vectors.

Example: For QPSK with 8 sectors (so $a = 2$), the relevant vectors for block length 2 are $[0 \;\; 0]$ and $[0 \;\; 1]$.

Computation of $\mathsf{P}(\mathbf{Z})$

We now need to compute of $\mathsf{P}(\mathbf{z}) = \frac{1}{M^L} \sum_{\mathbf{x} \in \mathcal{X}} \mathsf{P}(\mathbf{z}|\mathbf{x})$ for each of the $C(a + L - 2, L - 1)$ vectors. A brute force approach is to compute $\mathsf{P}(\mathbf{z}|\mathbf{x})$ for each $\mathbf{x}$, but again, has exponential complexity. We exploit the structure in $\mathbf{z}$ to reduce the number of vectors $\mathbf{x}$ for which we need $\mathsf{P}(\mathbf{z}|\mathbf{x})$. Specifically, we have that each $z_i \in \{0, 1, \cdots, a-1\}$. Since there are only a different types of components in $\mathbf{z}$, for block length $L > a$, some of the components in $\mathbf{z}$ will be repeated. For any $\mathbf{x}$, we can then use Property B-2 to rearrange the components at those locations for which the components in $\mathbf{z}$ are identical, without changing the conditional probability. For instance, let $z_m = z_n$ for some m, n. Then, $\mathsf{P}(\mathbf{z}|\mathbf{x}) = \mathsf{P}(\mathbf{z}|\Pi\mathbf{x})$, where $\Pi\mathbf{x}$ is obtained from $\mathbf{x}$ by rearranging the components at locations m and n. To sum up, we can restrict attention to a set of vectors $S_{\mathbf{X}}$ in which no vector can be obtained from another one by permutations between those locations for which the elements in $\mathbf{z}$ are identical.

To obtain the set $S_{\mathbf{X}}$, we divide the L locations into a groups, and permutations are allowed only between locations belonging to the same group. The problem then breaks down into a sub-problems. Specifically, let the number of locations in the groups be $n_0, n_1, \cdots, n_{a-1}$, then we need to distribute n_i identical balls into M distinct boxes, for each i. The required number of combinations is the product of the individual solutions. While for large a, the reduction in complexity may not be huge, for small values of a (which is the paradigm of interest in this work), the savings will be significant. For instance, for QPSK with $L = 8$, and $a = 2$, the worst case (which happens when $n_0 = n_1$) number of combinations is 1225, compared to the exponential figure of $M^L = 65536$. Once the set S_X has been obtained, we can get $\mathsf{P}(\mathbf{z}) = \frac{1}{M^L} \sum_{\mathbf{x} \in S_X} q(x) \mathsf{P}(\mathbf{z}|\mathbf{x})$. Here, $q(x) = \prod_{i=0}^{a-1} \frac{(n_i)!}{\prod_{j=0}^{M-1} (r_{i,j}(x))!}$

, where $r_{i,j}(x)$ is the number of times the input symbol j occurs in the locations belonging to group i, for the vector x.

Our numerical results for capacity computation are provided in Section 4.6. In the next section, we focus on efficient block noncoherent demodulation. This enables us to evaluate the uncoded error rates for our phase-quantized channel model.

4.5 Block Noncoherent Demodulation

We consider the generalized likelihood ratio test (GLRT) for block noncoherent demodulation. This entails a joint maximum likelihood estimation of the unknown block of input symbols and the unknown channel phase. Specifically, given the received vector $\mathbf{z}$, the GLRT estimate for $\mathbf{x}$ is given by

$$\hat{\mathbf{x}}(\mathbf{z}) = \operatorname*{argmax}_{\mathbf{x} \in \mathcal{X}} \max_{\phi \in [0,2\pi)} \mathsf{P}(\mathbf{z}|\mathbf{x}, \phi) \ . \tag{4.5}$$

Brute force computation of the solution to (4.5) has prohibitive complexity, since the cardinality of the input space $\mathcal{X}$ grows exponentially with the block length. For unquantized observations, it is known that the solution can rather be obtained with linear-logarithmic complexity [67]. The key idea used to obtain this complexity reduction works for quantized observations as well, and we illustrate it next.

First, we make some observations resulting due to the symmetry of our channel model. As before, we let $a = \frac{K}{M}$, and $\mathbf{q} = [q_0 \cdots q_{L-1}]$, where q_l is the quotient obtained on dividing z_l by a. Using Property A-4, we get

$$\hat{\mathbf{x}}(\mathbf{z}) = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \max_{\phi \in [0,2\pi)} \mathsf{P}(\mathbf{z} \bmod a | \mathbf{x} - \mathbf{q}, \phi) \ , \tag{4.6}$$

which in turn gives

$$\hat{\mathbf{x}}(\mathbf{z}) = \hat{\mathbf{x}}(\mathbf{z} \mod a) + \mathbf{q}(\mathbf{z}) \ , \tag{4.7}$$

where we have explicitly noted that $\mathbf{q}$ is a function of $\mathbf{z}$. This result is useful in the sense that the solution for a received vector $\mathbf{z}$ can be easily obtained if the solution for $\mathbf{z} \bmod a$ is known, since computing $\mathbf{q}(\mathbf{z})$ is a trivial task. Hence, we restrict attention to computing the GLRT solution only for those $\mathbf{z}$ for which the vector components $\in \{0, 1, \cdots, a-1\}$. Also observe that $\mathsf{P}(\mathbf{z}|\mathbf{x}, \phi) = \mathsf{P}(\mathbf{z}|\mathbf{x} + i, \phi - i\frac{2\pi}{M})$. This implies that the demodulator can not distinguish between two input vectors that are related by the operation of constant addition. This is well known (for unquantized observations), and is the basis for using techniques such as differential modulation.

To obtain a low complexity solution, the key is to interchange the order of maximization in (4.5). Consider

$$\max_{\phi} \max_{\mathbf{x} \in \mathcal{X}} \ \mathsf{P}(\mathbf{z}|\mathbf{x}, \phi) \ . \tag{4.8}$$

For a fixed ϕ, the inner maximization over $\mathbf{x}$ is straightforward since it can done in a coherent manner, i.e., on a symbol by symbol basis. For $\phi = 0$, let the coherent solution be denoted by $\mathbf{c}(0) = [c_0(0) \cdots c_{L-1}(0)]$. (We dropped the dependence on $\mathbf{z}$ to simplify notation). Note that this means $c_l(0) = \underset{x_l \in \{0, \cdots, M-1\}}{\operatorname{argmax}} \mathsf{P}(z_l | x_l, \phi = 0)$. As ϕ is increased, the coherent solution $\mathbf{c}$ will change. However, this will happen only when any of the individual solutions c_l changes. The crucial observation now is that as ϕ is varied over 0 to $\frac{2\pi}{M}$, each of the individual solutions $c_l(\phi)$ changes only once. In other words, for each l, there is a *crossover angle* α_l, such that

$$\begin{aligned} c_l(\phi) &= c_l(0) \ , \quad \text{if} \ \ 0 \leq \phi \leq \alpha_l \\ &= c_l(0) + 1 \ , \quad \text{if} \ \ \alpha_l < \phi < \frac{2\pi}{M} \ . \end{aligned} \tag{4.9}$$

The exact crossover angles can be obtained simply as a function of z_l, K, M and the locations of the input constellation points. Now, since we only consider those $\mathbf{z}$ vectors for which every component $\in \{0, \cdots, a-1\}$, there can be at most a distinct crossover angles. Hence, when ϕ is varied between $[0, \frac{2\pi}{M})$, the number of distinct coherent solutions to the inner maximization in (4.8) is at most a, and these solutions can be obtained simply by sorting the crossover angles in an ascending order. For each of these (at most) a input vectors, we can now numerically compute the metric $\max\limits_{\phi \in [0, 2\pi)} \mathsf{P}(\mathbf{z} | \mathbf{x}, \phi)$, and pick the one with the largest metric as the GLRT solution. This numerical computation can be done, for example, by fine discretization of the interval $[0, 2\pi)$, and computing $\mathsf{P}(\mathbf{z} | \mathbf{x}, \phi)$ for every ϕ in this discrete set. The number of computations (multiplications) required to obtain $\max_{\phi \in [0,2\pi)} \mathsf{P}(\mathbf{z}|\mathbf{x}, \phi)$ then scales linearly in the block length L.

Note that we restricted attention to $\phi \in [0, \frac{2\pi}{M})$ only while performing the inner maximization in (4.8). This is because if we go on beyond $\frac{2\pi}{M}$, any new solution we get, say $\mathbf{c}_1$ will be related to one of the existing solutions, say $\mathbf{c}_2$, by the operation of constant addition, so that the noncoherent demodulator can not distinguish between $\mathbf{c_1}$ and $\mathbf{c_2}$.

4.6 Numerical Results

We now present results for QPSK with 8-sector and 12-sector phase quantization, for different block lengths L. We begin with the symbol error rate (SER) plots for block demodulation. Fig. 4.4 (left plot) shows the results for 8-sector quantization. Looking at the topmost curve, which corresponds to $L = 2$, we find that the performance is disastrous. As the SNR is increased, the SER falls off extremely slowly. A close analysis of the block demodulator reveals that the reason behind this is an ambiguity in the demodulator decision rule: for certain outputs $\mathbf{z}$, irrespective of the SNR, the demodulator always returns two equally likely solutions for the input $\mathbf{x}$. While we do not provide a complete analysis of this ambiguous behavior, an example scenario is shown in Fig. 4.5 to give insight. If

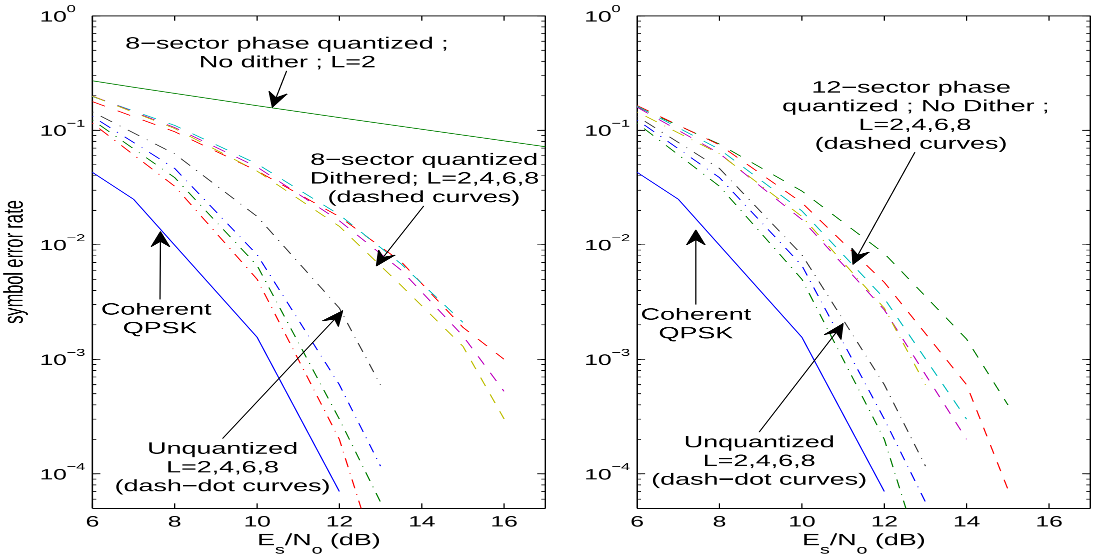

Figure 4.4: Symbol error rate performance for QPSK with 8-sector phase quantization (left figure) and 12-sector phase quantization (right figure), for block lengths varying from 2 to 8. Also shown for comparison are the curves for coherent QPSK, and noncoherent unquantized QPSK.

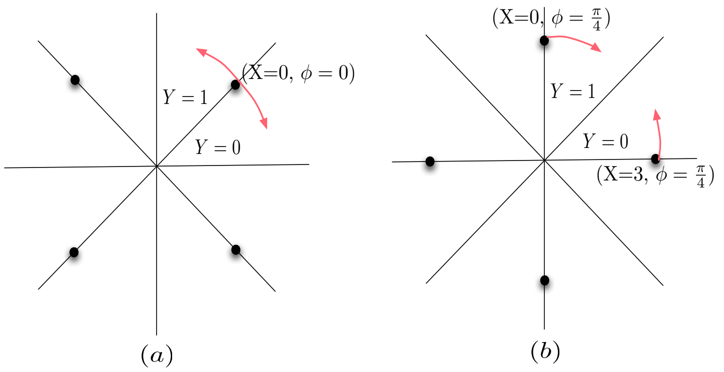

Figure 4.5: Ambiguity in the block noncoherent demodulator. If the received vector is $\mathbf{Z} = [1\ 0]$, then $(\mathbf{X} = [0\ 0], \phi = 0)$, and, $(\mathbf{X} = [0\ 3], \phi = \frac{\pi}{4})$ are both equally likely solutions.

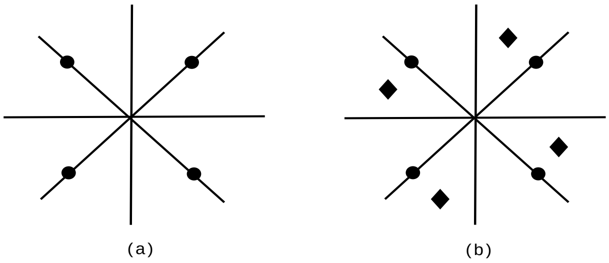

Figure 4.6: (a) Standard PSK : the same constellation (the one shown) is used for both symbols in the block. (b) Dithered-PSK : the constellations used for the two symbols are not identical, but the second one is a dithered version of the first one.

the quantized output vector is $\mathbf{z} = [1\ 0]$, then we find that $\mathsf{P}(\mathbf{z}|\mathbf{x}, \phi)$ is maximized by two equally likely pairs, $(\mathbf{x}_1, \phi_1) = ([0\ 0], 0)$, and $(\mathbf{x}_2, \phi_2) = ([0\ 3], \pi/4)$, so that the block demodulator, which does joint maximum likelihood estimation over the input and the unknown phase, becomes ambiguous. In other words, the symmetry inherent in the channel model, which on the one hand helped us reduce the complexity of capacity computations, is also making it impossible to distinguish between the effect of the unknown phase offset and the phase modulation on the received signal, resulting in poor performance. While we showed the performance plot for $L = 2$ only, we find that the ambiguity persists for larger block lengths also.

Possible ways to break the undesirable symmetries could be to use non-uniform phase quantization, or to employ dithering. Here we investigate the role of the latter. We can either dither the QPSK constellation points at the transmitter, or use analog pre-multipliers to dither the phase quantization boundaries at the receiver. We use a transmit dither scheme in which we rotate the QPSK constellation by an angle of $\frac{1}{L}(\frac{2\pi}{K})$ from one symbol to the next. Fig. 4.6 shows this scheme for block length L=2 and K=8. The constellation used for the second symbol (shown by the diamond shape) is dithered from the constellation used for the first symbol by an angle of $\pi/8$. With this choice of transmit constellations, we find that the ambiguity in the block demodulator is removed, and hence the performance is expected to improve. The results in Fig. 4.4 (left plot) indeed show a significant performance improvement compared to the no-dithering case, although increasing the block length does not provide much gain. At SER of 10^{-3}, 8-sector quantization with $L = 8$ results in a loss of about 4 dB compared to unquantized observations.

On the other hand, if we consider the performance with 12-sector quantization, it is observed that the block demodulator performs well, and dithering is not required. This suggests that 12-sector quantization does not result in any undesirable symmetries in the channel model. Fig. 4.4 (right plot) shows the performance for different block lengths. At SER of 10^{-3}, and $L = 8$, the loss compared to the unquantized observations is reduced to about 2 dB.

Next we show the plots for channel capacity. For all our results, we normalized the mutual information $I(\mathbf{X};\mathbf{Z})$ by L-1 to obtain the per symbol capacity, since

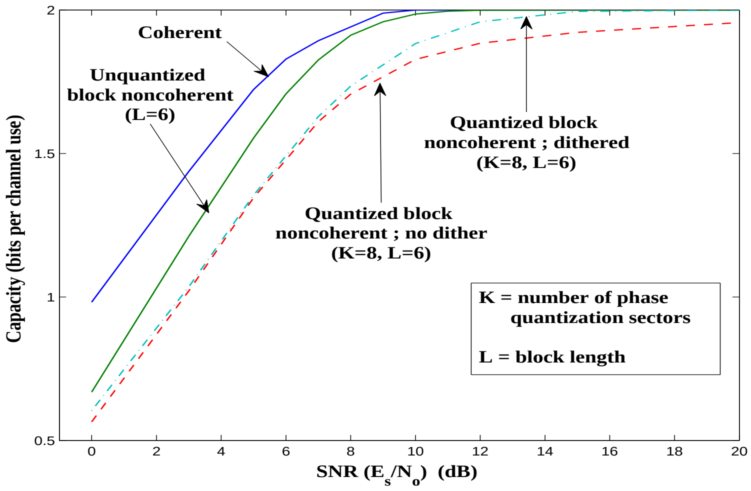

Figure 4.7: Performance comparison for QPSK with block length $L = 6$: plots depict the capacity of the block noncoherent channel without quantization, and with 8-sector quantization (with and without dithering). Also shown is the capacity for coherent QPSK.

in practice the successive blocks can be overlapped by one symbol due to slow phase variation from one block to the next. Fig. 4.7 shows the results for 8-sector quantization. (To avoid clutter, we show the results for $L = 6$ only.) Also shown for reference are the capacity values for the coherent case, and for the block noncoherent case without any quantization. Despite the disastrous performance of the uncoded scheme witnessed earlier, we see that, in terms of the channel capacity, 8-sector quantization scheme recovers more than 80-85% of the capacity obtained with unquantized observations, for $\mathsf{SNR} >$ 2-3 dB . However, the capacity approaches 2 bits/per channel use extremely slowly. Since $H(\mathbf{X})$ is constant, this implies that $H(\mathbf{X}|\mathbf{Z})$ falls off very slowly as $\mathsf{SNR} \to \infty$, which is consistent with

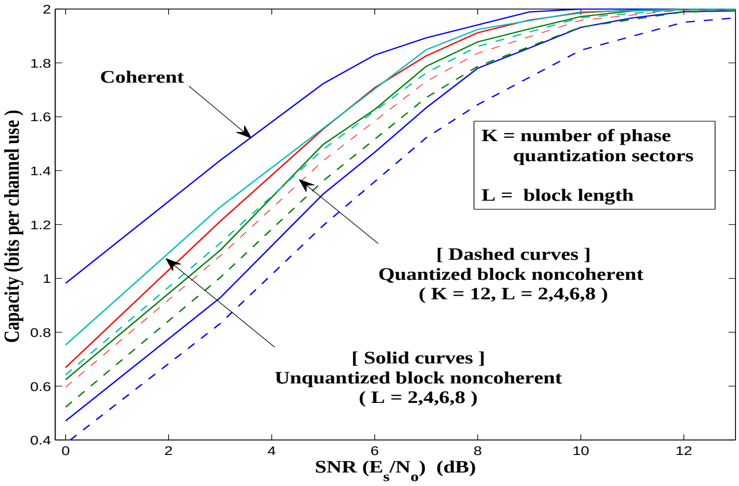

Figure 4.8: Performance comparison for QPSK : plots depict the capacity of the coherent channel, unquantized block noncoherent channel (different block lengths), and the 12-sector quantized block noncoherent channel (different block lengths).

the earlier observation that there is significant ambiguity in $\mathbf{X}$, given $\mathbf{Z}$, even at high SNR. The performance improvement obtained by using a dithered-QPSK input is also shown in Fig. 4.7. It is seen that the slow increase of capacity towards 2 bits/channel use has been eliminated. [2]

While the simple transmit dither scheme considered here has improved the performance (in terms of both the SER, as well as channel capacity), we hasten to add that there is no optimality associated with it. A more detailed investigation of different dithering schemes and their potential gains is therefore an important topic for future research.

[2] Since the low-complexity procedure outlined in Section 4.4 does not work once we dither, we used Monte Carlo simulations to compute the capacity with dithering.

In Fig. 4.8, we plot the capacity curves for QPSK with 12-sector quantization, for block length L=2,4,6,8. Also shown for reference are the coherent and unquantized block noncoherent performance curves. For identical block lengths, the loss in capacity (at a fixed $\mathsf{SNR} > 2\text{-}3$ dB) compared to the unquantized case is less than 5-10 %, while the loss in power efficiency (for fixed capacity) varies between 0.5-2 dB, and as before, dithering is not required.

4.7 Open Issues

There are several open issues to be addressed. Given the performance improvement obtained by using the simple dithering scheme considered here, a more detailed investigation of different dithering schemes is required. Another possibility to consider would be non uniform phase quantization. While we have restricted attention to PSK inputs in this work, it is important to evaluate performance with QAM alphabets as well, in which case we need to consider amplitude quantization. Note that, amplitude quantization can, in principle help improve performance with PSK inputs as well, especially if the SNR is low, and the block lengths are small.

As with prior work in the literature, we assumed that the phase across the different blocks varies independently. While this allows analytical tractability, the continuous variation of the phase from one block to the next can be used to enhance performance, especially when we are constrained to using low-precision samples. How best to leverage this memory might be worth investigating.

Chapter 5

Conclusions

The work in this thesis marks the first steps towards a comprehensive investigation, and a theory, of communication system design with low-precision ADC at the receiver. The results obtained from our Shannon-theoretic investigations in Chapter 3 indicate that the choice of low-precision ADC is consistent with the overall design goals for future high-bandwidth systems. The availability of a large amount of bandwidth encourages us to use power-efficient communication using small constellations (for two of the emerging high-bandwidth systems, this is essential as well: regulatory restrictions prohibit large transmit powers in UWB, while at mmwave frequencies, it is difficult to generate large transmit powers with integrated circuits in low-cost silicon processes), so that the symbol rate, and hence the sampling rate, for a given bit rate must be high. This forces us towards using ADCs with lower precision. Fortunately, this turns out to be consistent with the use of small constellations in the first place for power-efficient design.

Thus, if we plan on operating at low to moderate SNR, the small reduction in spectral efficiency due to low-precision ADC is acceptable in such systems, given that bandwidth is plentiful.

For the problem of carrier synchronization with low-precision ADC, we have investigated the feasibility of the block noncoherent approach, that amounts to joint detection of the transmitted symbols and the unknown carrier phase. While our results indicate that this approach could be a feasible option, they also lead to two important observations: first, low-precision quantization might lead to unexpected and ambiguous receiver operation when dealing with unknown parameters, and second, mechanisms such as dithering might be essential to attain good performance in the face of such ambiguities.

5.1 Directions for Future Work

While this thesis has laid some of the basic groundwork, there is clearly a lot of research to be done before low-precision ADC can make its way into real systems. Immediate problems to investigate, are aplenty: further exploration is needed to tackle carrier asynchronism, for both the implicit noncoherent approach, as well as for possible explicit estimation and correction approaches; robust and elegant solutions are also required for the problems of timing synchronization and automatic gain control, which we have not investigated in this work.

Another topic of interest, which needs exploration, is the use of low-precision ADC for fading and dispersive channel environments. Since low-precision ADC may not provide enough dynamic range to work with the signal received over such channels, it would be fruitful to investigate feedback-based transmit precoding strategies, wherein the receiver estimates the channel impulse response (possibly using sophisticated dithering techniques [27]) and feeds it back to the transmitter for precompensation. Such an architecture could be suitable for slowly varying indoor wireless personal area network channels, and especially for applications where the transmitter is more powerful than the receiver (e.g., laptop transmitting to a handheld). Indeed, if the channel impairments can be ignored after precoding, then we are back to the AWGN model, so that a receiver employing low-precision ADC can be expected to work well. Possible techniques to accomplish transmit precoding could include time reversal [17], or nonlinear techniques such as Tomlinson-Harashima precoding [73, 74]. Such techniques must, however, be assessed keeping in mind the small dynamic range available in a receiver employing low-precision ADC.

A complementary approach to using low-precision ADC, which can overcome the ADC bottleneck at high speeds, is to use a time-interleaved (TI)-ADC architecture, in which several low-speed, high-precision ADCs operate in parallel to synthesize a high-speed high-precision ADC. The problem to address here is effective compensation of the mismatches between the different sub-ADCs, such as gain and timing mismatches, which if left uncompensated can lead to error floors in the performance [75]. Much further research and experimentation is needed to determine the relative merits of the low-precision and TI-ADC approaches, as well as to investigate combinations thereof, for different application scenarios.

Bibliography

[1] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, April 1999.

[2] J. Huang and S. P. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2336–2351, July 2005.

[3] R. Hiremane, "From Moores law to Intel innovation-prediction to reality," *Technology@Intel Magazine*, April 2005.

[4] IEEE 802.15 WPAN Millimeter Wave Alternative PHY Task Group 3c, http://www.ieee802.org/15/pub/TG3c.html.

[5] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Transactions on Communications*, vol. 57, no. 12, Dec. 2009.

[6] J. Singh and U. Madhow, "On block noncoherent communication with low-precision phase quantization at the receiver," in *IEEE International Symposium on Information Theory, Seoul, Korea*, 2009.

[7] J. Singh, P. Sandeep, and U. Madhow, "Multi-gigabit communication: The ADC bottleneck," in *IEEE International Conference on Ultra-Wideband, Vancouver, Canada, invited paper*, 2009.

[8] R. V. Plassche, *CMOS Integrated Analog-to-Digital and Digital-to-Analog Converters.* Kluwer Academic Publishers, 2nd Edition.

[9] IEEE 802.15 WPAN High Rate Alternative PHY Task Group 3a, http://www.ieee802.org/15/pub/TG3a.html.

[10] S. Hoyos, B. M. Sadler, and G. R. Arce, "Monobit digital receivers for ultra-wideband," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, July 2005.

[11] R. Blazquez, F. S. Lee, D. Wentzloff, P. Newaskar, J. Powell, and A. Chandrakasan, "Digital architecture for an ultra-wideband radio receiver," in *Proc. 19th International Conference on VLSI Design*, 2003.

[12] I. S.-C. Lu, N. Weste, and S. Parameswaran, "ADC precision requirement for digital ultra-wideband receivers with sublinear front-ends: a power and performance perspective," in *Proc. 19th International Conference on VLSI Design (VLSID'06)*, 2006.

[13] O. Dabeer and U. Madhow, "Detection and interference suppression for ultra-wideband signaling with analog processing and one bit A/D," in *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers*, 2003.

[14] S. Hoyos and B. M. Sadler, "Ultra-wideband analog-to-digital conversion via signal expansion," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 5, Sep. 2005.

[15] ——, "Frequency-domain implementation of the transmitted-reference ultra-wideband receiver," *IEEE Transactions On Microwave Theory And Techniques*, vol. 54, no. 4, Apr. 2006.

[16] W. Namgoong, "A channelized digital ultrawideband receiver," *IEEE Transactions On Wireless Communications*, vol. 2, no. 3, May 2003.

[17] T. Strohmer, M. Emami, J. Hansen, G. Papanicolaou, and A. Paulraj, "Application of time-reversal with MMSE equalizer to UWB communications," in *IEEE GLOBECOM*, vol. 5, 2004, pp. 3123–3127.

[18] C. Zhou and R. Qiu, "Spatial focusing of time-reversed UWB electromagnetic waves in a hallway environment," in *Proc. of the Thirty-Eighth Southeastern Symposium on System Theory*, 2006, pp. 102–106.

[19] D. A. Sobel, "A baseband mixed-signal receiver front-end for 1Gbps wireless communications at 60GHz," *PhD Thesis, UC Berkeley*, 2008.

[20] E. Masry, "The reconstruction of analog signals from the sign of their noisy samples," *IEEE Transactions on Information Theory*, vol. 27, no. 6, pp. 735–745, Nov. 1981.

[21] Z. Cvetkovic and I. Daubechies, “Single-bit oversampled A/D conversion with exponential accuracy in the bit-rate,” in *Proceedings DCC*, Utah, USA, March 2000, pp. 343–352.

[22] P. Ishwar, A. Kumar, and K. Ramachandran, “Distributed sampling for dense sensor networks: A bit-conservation principal,” in *Proceedings of Information Processing in Sensor Networks*, Palo Alto, USA, April 2003, pp. 17–31.

[23] A. Host-Madsen and P. Handel, “Effects of sampling and quantization on single-tone frequency estimation,” *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 650–662, Mar. 2000.

[24] T. Andersson, M. Skoglund, and P. Handel, “Frequency estimation by 1-bit quantization and table look-up processing,” in *Proceedings of European Signal Processing Conference*, September 2000.

[25] D. Rousseau, G. V. Anand, and F. Chapeau-Blondeau, “Nonlinear estimation from quantized signals: Quantizer optimization and stochastic resonance,” in *Third International Symposium on Physics in Signal and Image Processing*, Grenoble, France, Jan. 2003, pp. 89–92.

[26] O. Dabeer and A. Karnik, “Consistent signal parameter estimation with 1-bit dithered quantization,” in *Proc. 14th European Signal Processing Conference*, Italy, September 2006.

[27] ——, “Signal parameter estimation with 1-bit dithered quantization,” *IEEE Transactions on Information Theory*, vol. 52, no. 12, Dec. 2006.

[28] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul and Oct 1948.

[29] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon limit error-correcting coding and decoding,” in *IEEE International Conference on Communications, Geneva*, 1993, pp. 1064–1070.

[30] C. Berrou and A. Glavieux, “Near optimum error correcting coding and decoding: Turbo-codes,” *IEEE Transactions on Communications*, vol. 44, no. 10, 1996.

[31] R. Gallager, “Low density parity check codes,” *IRE Transactions on Information Theory*, vol. 8, pp. 21–28, 1962.

[32] D. J. C. MacKay, “Good error-correcting codes based on very sparse matrices,” *IEEE Transactions on Information Theory*, vol. 45, pp. 299–431, 1999.

[33] R. E. Blahut, “Computation of channel capacity and rate-distortion functions,” *IEEE Transactions on Information Theory*, vol. 18, pp. 460–473, 1972.

[34] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memoryless channels,” *IEEE Transactions on Information Theory*, vol. 18, pp. 14–20, Jan. 1972.

[35] J. G. Smith, “The information capacity of amplitude and variance-constrained scalar Gaussian channels,” *Information and Control*, vol. 18, pp. 203–219, 1971.

[36] S. Shamai and I. Bar-David, “The capacity of average and peak-power-limited quadrature Gaussian channels,” *IEEE Transactions on Information Theory*, vol. 41, pp. 1060–1071, 1995.

[37] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, “The capacity of discrete-time memoryless Rayleigh fading channels,” *IEEE Transactions on Information Theory*, vol. 47, no. 4, May 2001.

[38] M. C. Gursoy, H. V. Poor, and S. Verdu, “The noncoherent Rician fading channel Part I : Structure of the capacity-achieving input,” *IEEE Transactions Wireless Communication*, vol. 4, pp. 2193–2206, 2005.

[39] T. H. Chan, S. Hranilovic, and F. R. Kschischang, “Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs,” *IEEE Transactions on Information Theory*, vol. 51, pp. 2073–2088, 2005.

[40] M. Katz and S. Shamai, “On the capacity-achieving distribution of the discrete-time noncoherent and partially coherent AWGN channels,” *IEEE Transactions on Information Theory*, vol. 50, pp. 2257–2270, 2004.

[41] A. Tchamkerten, “On the discreteness of capacity-achieving distributions,” *IEEE Transactions on Information Theory*, pp. 2273–2278, 2004.

[42] R. G. Gallager, *Information Theory and Reliable Communication.* John Wiley And Sons, Inc., 1968.

[43] H. Witsenhausen, “Some aspects of convexity useful in Information theory,” *IEEE Transactions on Information Theory*, vol. 26, no. 3, May 1980.

[44] L. Dubins, “On extreme points of convex sets,” *Journal of Mathematical Analysis and Applications*, vol. 5, pp. 237–244, May 1962.

[45] N. Phamdo and F. Alajaji, “Soft-decision demodulation design for COVQ over white, colored, and ISI Gaussian channels,” *IEEE Transactions on Communications*, vol. 48, no. 9, pp. 1499–1506, 2000.

[46] F. Behnamfar, F. Alajaji, and T. Linder, “Channel-optimized quantization with soft-decision demodulation for space-time orthogonal block-coded channels,” *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3935–3946, 2006.

[47] J. A. Nossek and M. T. Ivrlac, “Capacity and coding for quantized MIMO systems,” in *International Conference on Wireless Comm. and Mobile Computing, Vancouver, Canada.*, 2006.

[48] L. N. Lee, “On optimal soft decision demodulation,” *IEEE Transactions on Information Theory*, vol. 22, pp. 437–444, 1976.

[49] J. Salz and E. Zehavi, “Decoding under integer metrics constraints,” *IEEE Transactions on Communications*, vol. 43, pp. 307–317, 1995.

[50] O. Dabeer, J. Singh, and U. Madhow, “On the limits of communication performance with one-bit analog-to-digital conversion,” in *IEEE Workshop on Signal Processing Advances in Wireless Communication, Cannes, France*, 2006.

[51] J. Singh, O. Dabeer, and U. Madhow, “Communication limits with low-precision analog-to-digital conversion at the receiver,” in *IEEE International Conference on Communications, Glasgow, Scotland*, 2007.

[52] ——, “Capacity of the discrete-time AWGN channel under output quantization,” in *IEEE International Symposium on Information Theory, Toronto, Canada*, 2008.

[53] Y. Wu, L. M. Davis, and R. Calderback, “On the capacity of the discrete-time channel with uniform output quantization,” in *IEEE International Symposium on Information Theory, Seoul, Korea*, 2009.

[54] S. Krone and G. Fettweis, “Fundamental limits to communications with analog-to-digital conversion at the receiver,” in *IEEE International Workshop on Signal Processing Advances in Wireless Communications, Perugia, Italy*, 2009.

[55] A. Mezghani and J. A. Nossek, “Analysis of rayleigh-fading channels with 1-bit quantized output,” in *IEEE International Symposium on Information Theory, Toronto, Canada*, 2008.

[56] A. Mezghani, M. S. Khoufi, and J. A. Nossek, “Spatial MIMO decision feedback equalizer operating on quantized data,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[57] A. Seyedi, “Three phase analog-to-digital conversion for high-rate short-range communication,” in *IEEE International Workshop on Signal Processing Advances in Wireless Communications, Perugia, Italy*, 2009.

[58] T. Cover and J. Thomas, *Elements of Information Theory.* Wiley Series in Telecommunications.

[59] I. Csiszar and J. Korner, *Information Theory : Coding Theorems For Discrete Memoryless Systems.* Academic Press, 1981.

[60] M. Chiang and S. Boyd, “Geometric programming duals of channel capacity and rate distortion,” *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 245–257, February 2004.

[61] A. Lapidoth and S. M. Moser, “Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, Oct. 2003.

[62] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[63] J. G. Proakis, *Digital Communications.* McGraw Hill, 4th Edition.

[64] U. Madhow, *Fundamentals of Digital Communication.* Cambridge University Press, 2008.

[65] D. Divsalar and M. Simon, “Multiple-symbol differential detection of MPSK,” *IEEE Transactions on Communications*, vol. 38, no. 3, pp. 300–308, Mar. 1990.

[66] K. M. Mackenthun, “A fast algorithm for multiple-symbol differential detection of MPSK,” *IEEE Transactions on Communications*, vol. 42, pp. 1471–1474, 1994.

[67] W. Sweldens, “Fast block noncoherent decoding,” *IEEE Communications Letters*, vol. 5, pp. 132–134, 2001.

[68] D. Warrier and U. Madhow, “Spectrally efficient noncoherent communication,” *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 651–668, Mar. 2002.

[69] M. Peleg and S. Shamai, “On the capacity of the blockwise incoherent MPSK channel,” *IEEE Transactions on Communications*, vol. 46, pp. 603–609, Mar. 1998.

[70] T. Marzetta and B. Hochwald, “Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading,” *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

[71] R.-R. Chen, R. Koetter, U. Madhow, and D. Agrawal, “Joint noncoherent demodulation and decoding for the block fading channel: a practical framework for approaching Shannon capacity,” *IEEE Transactions on Communications*, Oct. 2003.

[72] N. Jacobsen and U. Madhow, “Coded noncoherent communication with amplitude/phase modulation: from Shannon theory to practical turbo architectures,” *IEEE Transactions on Communications*, vol. 56, no. 12, pp. 2040–2049, Dec. 2008.

[73] M. Tomlinson, “New automatic equalizer employing modulo arithmetic,” *Electronic Letters*, vol. 7, pp. 138–139, March 1971.

[74] H. Harashima and H. Miyakawa, “Matched transmission technique for channels with intersymbol interference,” *IEEE Transactions Communications*, vol. 20, pp. 774–780, August 1972.

[75] P. Sandeep, U. Madhow, M. Seo, and M. Rodwell, “Joint channel and mismatch correction for OFDM reception with time-interleaved ADCs: towards mostly digital multigigabit transceiver architectures,” in *IEEE Global Telecommunications Conference, New Orleans*, 2008.

[76] D. G. Luenberger, *Optimization by Vector Space Methods.* John Wiley And Sons, Inc., 1969.

[77] M. Fozunbal, S. W. McLaughlin, and R. W. Schafer, “Capacity analysis for continuous-alphabet channels with side information, part I: A general framework,” *IEEE Transactions on Information Theory*, vol. 51, no. 9.

[78] P. Billingsley, *Convergence of Probability Measures.* Wiley Series in Probability and Mathematical Statistics, 1968.

Appendices

Appendix A

A.1 Achievability of Capacity

Theorem 3 *[76] Let $\mathcal{V}$ be a real normed linear vector space, and $\mathcal{V}^*$ be its normed dual space. A weak* continuous real-valued functional f evaluated on a weak* compact subset $\mathcal{F}$ of $\mathcal{V}^*$ achieves its maximum on $\mathcal{F}$.*

Proof: See [76, p. 128, Thm 2]. ■

The use of this optimization theorem to establish the existence of a capacity achieving input distribution is standard (see [37, 77] for details). To use the theorem for our channel model (3.1), we need to show that the set $\mathcal{F}$ of all average power constrained distribution functions is weak* compact, and the mutual information functional I is weak* continuous over $\mathcal{F}$, so that I achieves its maximum on $\mathcal{F}$ [1]. The weak* compactness of $\mathcal{F}$ has been shown in [37]. (The authors in [77] later generalized this result, to show the weak* compactness of a larger class of sets of distribution functions). To prove continuity, we need to show that

$$F_n \xrightarrow{weak^*} F \Longrightarrow I(F_n) \longrightarrow I(F)$$

The finite cardinality of the output for our problem trivially ensures this. Specifically,

$$\begin{aligned} I(F) &= H_Y(F) - H_{Y|X}(F) \\ &= -\sum_{i=1}^{K} R(y_i; F) \log R(y_i; F) + \int dF(x) \sum_{i=1}^{K} W_i(x) \log W_i(x) \end{aligned}$$

where,

$$R(y_i; F) = \int_{-\infty}^{\infty} W_i(x) dF(x).$$

[1] The notion of weak* convergence here is actually the same as the standard weak convergence defined in probability theory [78].

The continuous and bounded nature of $W_i(x)$ ensures that $R(y_i; F)$ is continuous (by the definition of weak* topology). Moreover, the function $\sum_{i=1}^{K} W_i(x) \log W_i(x)$ is also continuous and bounded, implying that $H_{Y|X}(F)$ is also continuous (again by the definition of weak* topology). The continuity of $I(F)$ thus follows.

A.2 KKT Condition

The KKT condition holds if the mutual information is weak* continuous and weak differentiable. The weak* continuity of mutual information for our problem has already been shown above, and we show the weak differentiability next.

Weak Differentiability of Mutual Information

The weak derivative of I at a point $F_0 \in \mathcal{F}$ is defined as ([35, 37])

$$I'_{F_0}(F) = \lim_{\theta \to 0} \frac{I((1-\theta)F_0 + \theta F) - I(F_0)}{\theta} \qquad \forall F \in \mathcal{F} \tag{A.1}$$

Let us define the divergence function

$$d(x; F) = \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F)}$$

and also let, $F_\theta = (1-\theta)F_0 + \theta F$.

Then,

$$I(F_\theta) - I(F_0) = \theta \int dF(x) d(x; F_\theta) - \theta \int dF_0(x) d(x; F_\theta) + \int dF_0(x) \sum_{i=1}^{K} \log \frac{R(y_i; F_0)}{R(y_i; F_\theta)}$$

Putting $R(y_i; F_\theta) = (1-\theta)R(y_i; F_0) + \theta R(y_i; F)$, we get

$$I'_{F_0}(F) = \lim_{\theta \to 0} \frac{I(F_\theta) - I(F_0)}{\theta} = \int dF(x) d(x; F_0) - I(F_0) \qquad \forall\ F_0, F \in \mathcal{F}$$

The weak derivative defined above exists for our case because both terms in the difference are finite (due to the discrete nature (with finite cardinality K) of the output Y).

A.3 Proof of Proposition 2

We extend Witsenhausen's result in [43] to incorporate an average power constraint on the input. Our approach is the same as taken by Witsenhausen.

Proof: Let $\mathcal{S}$ be the set of all average power constrained distributions with support in the interval $[A_1, A_2]$. The required capacity, by definition, is $C = \sup_{\mathcal{S}} I(X;Y)$, where $I(X;Y)$ denotes the mutual information between X and Y. The achievability of the capacity is guaranteed by Theorem 3 in Appendix A.1. The result [77, Lemma 3.1] ensures the weak* compactness of the set S, while weak* continuity of $I(X;Y)$ is easily proven given the assumption that the transition functions $W_i(x)$ are continuous. Let S^* be a capacity achieving input distribution.

The key idea that we employ is a theorem by Dubins [44], which characterizes extreme points of the intersection of a convex set with hyperplanes. We first give some necessary definitions, and then state the theorem.

Definitions :

- Let $\mathcal{E}$ be a vector space over the field of real numbers, and $\mathcal{M}$ be a convex subset of $\mathcal{E}$. $\mathcal{M}$ is said to be *linearly bounded* (respectively, *linearly closed*) if every line intersects $\mathcal{M}$ in a bounded (respectively closed) subset of the line.

- Let $f : \mathcal{E} \to \mathbb{R}$ be a linear functional (not identically zero). The set $\{x \in \mathcal{E} : f(x) = c\}$ defines a hyperplane, for any real c.

Dubins' Theorem : Let $\mathcal{M}$ be a linearly closed and linearly bounded convex set and $\mathcal{U}$ be the intersection of $\mathcal{M}$ with n hyperplanes, then every extreme point of $\mathcal{U}$ is a convex combination of at most $n+1$ extreme points of $\mathcal{M}$.

To apply Dubins' theorem to our problem, we begin by defining $C[A_1, A_2]$: the real normed linear space of all continuous functions on the interval $[A_1, A_2]$, with sup-norm. The dual of $C[A_1, A_2]$ is the space of functions of bounded variations [76, Sec 5.5], and it includes the (convex) set of all distribution functions with support in $[A_1, A_2]$. We take $\mathcal{E}$ to be the dual of $C[A_1, A_2]$, and $\mathcal{M}$ to be the subset of $\mathcal{E}$ consisting of all distribution functions with support in $[A_1, A_2]$. Note that the optimal input distribution $S^* \in \mathcal{M}$.

Let the probability vector of the output Y, when the input is S^*, be $R^* = \{{p_1}^*, {p_2}^*, \ldots, {p_K}^*\}$. Also, let the average power of the input under the distribution S^* be P_0, where $P_0 \leq P$.

Now, consider the following subset $\mathcal{U}$ of $\mathcal{M}$

$$\mathcal{U} = \{M \in \mathcal{M} | R(y; M) = R^* \text{ and } E(X^2) = P_0\}. \tag{A.2}$$

The set $\mathcal{U}$ is the intersection of the set $\mathcal{M}$ with the following K hyperplanes

$$H_i : \int_{A_1}^{A_2} W_i(x) dM(x) = p_i{}^* \quad 1 \leq i \leq K-1 \tag{A.3}$$

and,

$$H_K : \int_{A_1}^{A_2} x^2 dM(x) = P_0 \tag{A.4}$$

where $W_i(x)$ are the transition probability functions. Note that there are only $K-1$ hyperplanes in (A.3) since the probabilities must sum to 1, thus making the requirement on $p_K{}^*$ redundant.

We know that the set $\mathcal{M}$ is compact in the weak* topology [77, Lemma 3.1]. Also, each of the hyperplanes $H_i, 1 \leq i \leq K-1$, is a closed set since the functions $W_i(x)$ are continuous. The hyperplane H_K is closed as well, since x^2 is a continuous function. Therefore, the set $\mathcal{U}$, being the intersection of a weak* compact set with K closed sets, is weak* compact. It is easy to see that $\mathcal{U}$ is a convex set as well. On the set $\mathcal{U}$, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= -\sum_{i=1}^{K} p_i{}^* \log p_i{}^* + \int_{A_1}^{A_2} dM(x) \sum_{i=1}^{K} W_i(x) \log W_i(x). \end{aligned}$$

As a function of the distribution $M(\cdot)$, we get

$$I(X;Y) = \text{ constant } + \text{ linear } ,$$

and the linear part is weak* continuous since $\sum_{i=1}^{K} W_i(x) \log W_i(x)$ is in $C[A_1, A_2]$.

It follows that the (continuous and linear) functional $I(X;Y)$ attains its maximum over the (compact and convex) set $\mathcal{U}$ at an extreme point of $\mathcal{U}$. However, since $S^* \in \mathcal{U}$, any maxima over $\mathcal{U}$ is a maxima over $\mathcal{S}$ as well. Hence, the required capacity is achieved at an extreme point of $\mathcal{U}$.

We now apply Dubins' theorem to characterize the extreme points of $\mathcal{U}$. Since $\mathcal{U}$ is the intersection of $\mathcal{M}$ with K hyperplanes, every extreme point of $\mathcal{U}$ is a convex combination of at most $K+1$ extreme points of $\mathcal{M}$. The extreme points of $\mathcal{M}$ however are distributions concentrated at single points within the interval $[A_1, A_2]$. Therefore, we get that the required capacity is achievable by a discrete distribution with at most $K+1$ points of support. ■

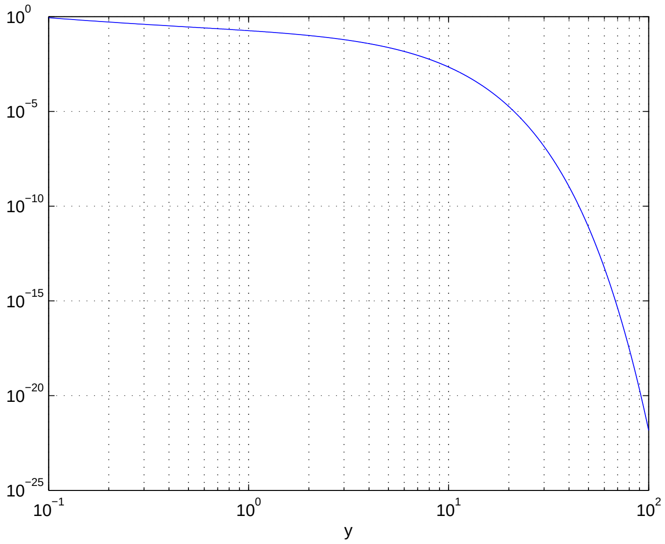

Figure A.1: The second derivative of $h(Q(\sqrt{y}))$ is positive for small values of y.

A.4 Convexity of the Function $h(Q(\sqrt{y}))$

To show convexity, we verify that the second derivative of the function $h(Q(\sqrt{y}))$ is positive everywhere. For $y > 2$, we do this analytically, while for $0 \leq y \leq 2$, the positivity of the second derivative is demonstrated numerically in Figure A.1.

Let $u(y) = h(Q(\sqrt{y}))$. Then,

$$u'(y) = \frac{-e^{-y/2}}{2\sqrt{2\pi y}\ln 2} \ln\left(\frac{1-Q(\sqrt{y})}{Q(\sqrt{y})}\right)$$

Note that $\frac{1-Q(\sqrt{y})}{Q(\sqrt{y})} \geq 1, \forall y \geq 0$. Therefore, to show that the second derivative $u''(y)$ is positive, it suffices to show that the function $v(y) = e^{-y/2}\ln\left[\frac{1-Q(\sqrt{y})}{Q(\sqrt{y})}\right]$ is a decreasing function of y. Taking the derivative of $v(y)$, we get

$$v'(y) = \frac{-e^{-y/2}}{2}\left[\ln\left(\frac{1-Q(\sqrt{y})}{Q(\sqrt{y})}\right) - \frac{e^{-y/2}}{\sqrt{2\pi y}\, Q(\sqrt{y})(1-Q(\sqrt{y}))}\right]$$

To show that $v(y)$ is decreasing, it suffices to show that

$$\ln\left(\frac{1-Q(\sqrt{y})}{Q(\sqrt{y})}\right) \geq \frac{e^{-y/2}}{\sqrt{2\pi y}\, Q(\sqrt{y})(1-Q(\sqrt{y}))} \tag{A.5}$$

Using the fact [64, pp. 78] that $Q(y) \geq (1 - \frac{1}{y^2})\frac{e^{-y^2/2}}{y\sqrt{2\pi}}$, we get that if $y > 1$, then the following condition is sufficient for (A.5) to be true

$$\ln\left(\frac{1 - Q(\sqrt{y})}{Q(\sqrt{y})}\right) \geq \frac{1}{(1 - \frac{1}{y})(1 - Q(\sqrt{y}))} \tag{A.6}$$

or, equivalently

$$(1 - \frac{1}{y})(1 - Q(\sqrt{y}))\ln\left(\frac{1 - Q(\sqrt{y})}{Q(\sqrt{y})}\right) \geq 1 \tag{A.7}$$

The left hand side of (A.7) is a monotone increasing function of y. For $y = 2$, it equals 1.133. Thus (A.7) holds $\forall y > 2$, and hence the second derivative of $h(Q(\sqrt{y}))$ must be positive for $y > 2$.