On the Limits of Communication with Low-Precision Analog-to-Digital Conversion at the Receiver

Jaspreet Singh, Onkar Dabeer, and Upamanyu Madhow,

Abstract

As communication systems scale up in speed and bandwidth, the cost and power consumption of high-precision (e.g., 8-12 bits) analog-to-digital conversion (ADC) becomes the limiting factor in modern transceiver architectures based on digital signal processing. In this work, we explore the impact of lowering the precision of the ADC on the performance of the communication link. Specifically, we evaluate the communication limits imposed by low-precision ADC (e.g., 1-3 bits) for transmission over the real discrete-time Additive White Gaussian Noise (AWGN) channel, under an *average power* constraint on the input. For an ADC with K quantization bins (i.e., a precision of $\log_2 K$ bits), we show that the input distribution need not have any more than K+1 mass points to achieve the channel capacity. For 2-bin (1-bit) symmetric quantization, this result is tightened to show that binary antipodal signaling is optimum for any signal-to-noise ratio (SNR). For multi-bit quantization, a dual formulation of the channel capacity problem is used to obtain tight upper bounds on the capacity. The cutting-plane algorithm is employed to compute the capacity numerically, and the results obtained are used to make the following encouraging observations : (a) up to a moderately high SNR of 20 dB, 2-3 bit quantization results in only 10-20% reduction of spectral efficiency compared to unquantized observations, (b) standard equiprobable pulse amplitude modulated input with quantizer thresholds set to implement maximum likelihood hard decisions is asymptotically optimum at high SNR, and works well at low to moderate SNRs as well.

Index Terms

Jaspreet Singh and Upamanyu Madhow are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, 93106, USA. (e-mail: {jsingh, madhow}@ece.ucsb.edu). Their research was supported by the National Science Foundation under grants CCF-0729222 and CNS-0832154.

Onkar Dabeer is with the School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, 400005, India. (e-mail: onkar@tcs.tifr.res.in). His research was supported by the Homi Bhabha Fellowship.

Channel Capacity, Optimum Input Distribution, AWGN Channel, Analog to Digital Converter, Quantization.

I. INTRODUCTION

Digital signal processing (DSP) forms the core of modern digital communication receiver implementations, with the analog baseband signal being converted to digital form using analog-to-digital converters (ADCs) which typically have fairly high (e.g., 8-12 bits) precision. Operations such as synchronization, equalization and demodulation are then performed in the digital domain, greatly enhancing the flexibility available to the designer. The continuing exponential advances in digital electronics, often summarized by Moore's "law" [1], imply that integrated circuit implementations of such DSP-centric architectures can be expected to continue scaling up in speed and down in cost. However, as the bandwidth of a communication system increases, accurate conversion of the analog received signal into digital form requires high-precision, high-speed ADC, which is costly and power-hungry [2]. One possible approach for designing such high-speed systems is to drastically reduce the number of bits of ADC precision (e.g., to 1-3 bits) as sampling rates scale up. Such a design choice has significant implications for all aspects of receiver design, including carrier and timing synchronization, equalization, demodulation and decoding. However, before embarking on a comprehensive rethinking of communication system design, it is important to understand the fundamental limits on communication performance imposed by low-precision ADC. In this paper, we take a first step in this direction, investigating Shannon-theoretic performance limits for the following idealized model : linear modulation over a real baseband Additive White Gaussian Noise (AWGN) channel, with symbol rate Nyquist samples quantized by a low-precision ADC at the receiver. This induces a discrete-time memoryless *AWGN-Quantized Output* channel, which is depicted in Figure 1. Under an *average power* constraint on the input, we obtain the following results:

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 AWGN 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.

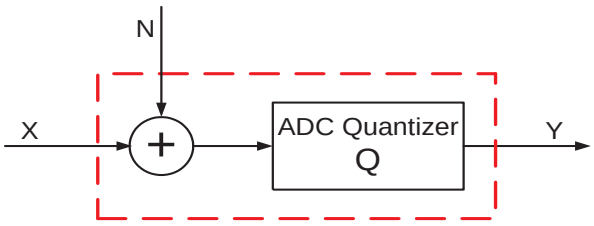

Fig. 1. $Y = \mathsf{Q}(X + N)$: The *AWGN-Quantized Ouput* channel induced by the output quantizer Q.

2) For 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).
3) 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 [3].
4) 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.
5) From our numerical results, we infer that the use of low-precision ADC incurs a relatively small loss in spectral efficiency compared to unquantized observations. For example, at 0 dB SNR, a receiver with 2-bit ADC achieves 95% of the spectral efficiency attained with unquantized observations. Even at a moderately high SNR of 20 dB, a receiver with 3-bit ADC achieves 85% of the spectral efficiency 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 a "sensible" choice of standard equiprobable pulse amplitude modulated (PAM) input with ADC thresholds set to implement maximum likelihood (ML) hard decisions achieves performance which is quite close to that obtained by numerical optimization of the quantizer and input distribution.

Related Work: For a discrete memoryless channel (DMC), Gallager first showed that the number of input points with nonzero probability mass need not exceed the cardinality of the output [4, p. 96, Corollary 3]. In our setting, the input alphabet is not a priori discrete, and there is a power constraint, so that the result in [4] does not apply. Our key result on the achievability of the capacity by a discrete input is actually inspired by Witsenhausen's result in [5], where

Dubins' theorem [6] was used to show that the capacity of a discrete-time memoryless channel with output cardinality K, under a *peak power* constraint is achievable by a discrete input with at most K points. The key to our proof is to show that under output quantization, an average power constraint automatically induces a peak power constraint, after which we can use Dubins' theorem in a manner similar to the development in [5] to show that $K+1$ mass points suffice to achieve the average power constrained capacity.

Prior work on the effect of reduced ADC precision on channel capacity with *fixed input distribution* includes [7], [8], [9]. However, other than our own results reported earlier in [10], [11], [12], we are not aware of an information-theoretic investigation with low-precision ADC 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 (see, e.g. [13], [14]).

Given the encouraging results here, it becomes important to explore the impact of low-precision ADC on receiver tasks such as synchronization and equalization, which we have ignored in our idealized model (essentially assuming that these tasks have somehow already been accomplished). Related work on estimation using low-precision samples which may be relevant for this purpose includes the use of dither for signal reconstruction [15], [16], [17], frequency estimation using 1-bit ADC [18], [19], choice of quantization threshold for signal amplitude estimation [20], and signal parameter estimation using 1-bit dithered quantization [21], [22].

Organization of the Paper: The rest of the paper is organized as follows. In Section II, we describe the channel model and present results concerning the structure of the optimal input distributions. In Section III, we discuss capacity computation, including duality-based upper bounds on capacity. Numerical results are provided in Section IV, followed by the conclusions in Section V.

II. Channel Model and Structure of Optimal Input Distributions

We consider linear modulation over a real AWGN channel, with symbol rate Nyquist samples quantized by a K-bin quantizer Q at the receiver. This induces the following discrete-time memoryless *AWGN-Quantized Output* (AWGN-QO) channel

$$Y = \mathsf{Q}\left(X + N\right) \ . \tag{1}$$

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

$$\begin{aligned} W_i(x) &= \mathsf{P}(Y = y_i | X = x) \\ &= Q\left(\frac{q_{i-1} - x}{\sigma}\right) - Q\left(\frac{q_i - x}{\sigma}\right), \qquad 1 \le i \le K, \end{aligned} \tag{2}$$

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

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp(-t^2/2) dt \ . \tag{3}$$

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

$$R(y_i; F) = \int_{-\infty}^{\infty} W_i(x) dF(x), \qquad 1 \le i \le K, \tag{4}$$

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

$$I(F) = \int_{-\infty}^{\infty} \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F)} dF(x) \ .^{1} \tag{5}$$

Under an average power constraint P, we wish to find the capacity of the channel (1), given by

$$C = \sup_{F \in \mathcal{F}} I(F), \tag{6}$$

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 $\mathbb{R}$.

Structural Properties 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 (1) must have bounded support. [2]

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

[2]That there *exists* a capacity achieving distribution follows by standard function analytic arguments [23]. For details, see Appendix A.

A. An Implicit Peak Power Constraint

The KKT optimality condition for an average power constrained channel has been derived in [25]. The mild technical conditions required for it to hold are verified for our channel model in Appendix B. The condition states that an input distribution F^* achieves the capacity C in (6) *if and only if* there exists $\gamma \geq 0$ such that

$$\sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F^*)} + \gamma(P - x^2) \leq C \tag{7}$$

for all x, with equality if x is in the support [3] of F^*, where the transition probability function $W_i(x)$, and the output probability $R(y_i; F^*)$ are as specified in (2) and (4), respectively.

The summation on the left-hand side (LHS) of (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,

$$d(x; F) = \sum_{i=1}^{K} W_i(x) \log \frac{W_i(x)}{R(y_i; F)} \ . \tag{8}$$

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

Lemma 1. *For the AWGN-QO channel* (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 $x > A_0, d(x; F) < -\log R(y_K; F)$.[4]

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

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

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

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* (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 (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 (7) for any $x > {A_2}^*$.

From Lemma 1, we get $\lim_{x\to\infty} d(x;F^*) = -\log R(y_K;F^*) =: L$. 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 (7) reduces to

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

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

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

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. □

B. Achievability of Capacity by a Discrete Input

In [5], 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 [6], 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 Appendix C), 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 C. □

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* (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 (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. □

C. Symmetric Inputs for Symmetric Quantization

For our capacity computations ahead, we assume that the quantizer Q employed in (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 (1), although a proof of this conjecture has eluded us. However, once we assume that the quantizer in (1) is symmetric, we can restrict attention to only symmetric inputs without loss of optimality, as stated in the following Lemma.[5]

Lemma 2. *If the quantizer in* (1) *is symmetric, then, without loss of optimality, we can consider only symmetric inputs for the capacity computation in* (6)*.*

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. □

III. Capacity Computation

In this section, we consider capacity computation for the AWGN-QO channel. We first provide an explicit capacity formula for the extreme scenario of 1-bit symmetric quantization, and then discuss numerical computations for multi-bit quantization.

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

A. 1-bit Symmetric Quantization : Binary Antipodal Signaling is Optimal

With 1-bit symmetric quantization, the channel is

$$Y = \text{sign}(X + N). \tag{9}$$

Theorem 1 (Section II-B) 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 (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,

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

and $Q(\cdot)$ is the complementary Gaussian distribution function shown in (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 $\mathbb{E}[X^2] \le P$. 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

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

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 Appendix D, we show that the function $h(Q(\sqrt{z}))$ is convex in z. Jensen's inequality [26] 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

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

□

B. Multi-Bit Quantization

We now consider K-bin symmetric quantization for $K > 2$. Every choice of the quantizer results in a unique channel model (1). In this section, we discuss capacity computation assuming a fixed quantizer only. Optimization over the quantizer choice is performed in Section IV.

1) **Capacity computation using 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 [3, 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 II-A) 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.

2) **Duality-based upper bound on channel capacity:** 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

$$C \leq U(R) = \min_{\gamma \geq 0} \sup_{x \in \mathcal{X}} [D(W(\cdot|x)||R(\cdot)) + \gamma(P - x^2)] \ , \tag{10}$$

where γ is a Lagrange parameter, and $D(W(\cdot|x)||R(\cdot))$ is the divergence between the transition and output distributions. While [27] provides this bound for a discrete channel, its extension to continuous alphabet channels has been established in [28]. For a more detailed perspective on duality-based upper bounds, see [29].

For an arbitrary choice of $R(y)$, the bound (10) might be quite loose. Therefore, to obtain a tight upper bound, we may need to evaluate (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 (10) for different choices of α.

It remains to specify how to compute the upper bound (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 [36, pp. 81]. The convexity of $m(\gamma)$ guarantees that $fminimax$ provides us the global minimum over γ.

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 quantization, 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 2, and the mutual information achieved by them is given in Table I [6]. As predicted by Theorem 1 (Section II-B), the support set of the input distribution (at each SNR) has cardinality ≤ 5.

For upper bound computations, we evaluate (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 $(0, 0.5)$, 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 I. 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.

[6] If, in the input distribution generated by the cutting-plane algorithm, we observed that two adjacent support set points (on our fine discrete grid) possessed non-zero probability masses, we merged them into a single point (while ensuring that the power constraint is satisfied).

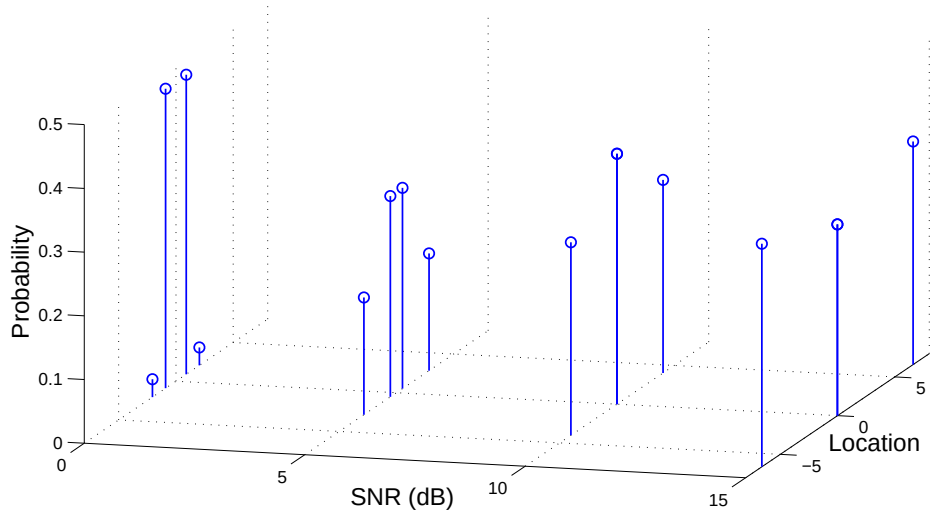

Fig. 2. Probability mass function of the optimal input generated by the cutting-plane algorithm [3] at various SNRs, for the 2-bit symmetric quantizer with thresholds $\{-2, 0, 2\}$. (Noise variance $\sigma^2 = 1$.)

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 I

DUALITY-BASED UPPER BOUNDS ON CHANNEL CAPACITY, COMPARED WITH THE MUTUAL INFORMATION (MI) ACHIEVED BY THE DISTRIBUTIONS GENERATED USING THE CUTTING-PLANE ALGORITHM.

It is insightful to verify that the preceding near-optimal input distributions satisfy the KKT condition (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 plots, as a function of x, the LHS of (7) for this input distribution. (The value of γ used in the plot was obtained by equating the LHS of (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 (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.

IV. QUANTIZER OPTIMIZATION AND NUMERICAL RESULTS

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

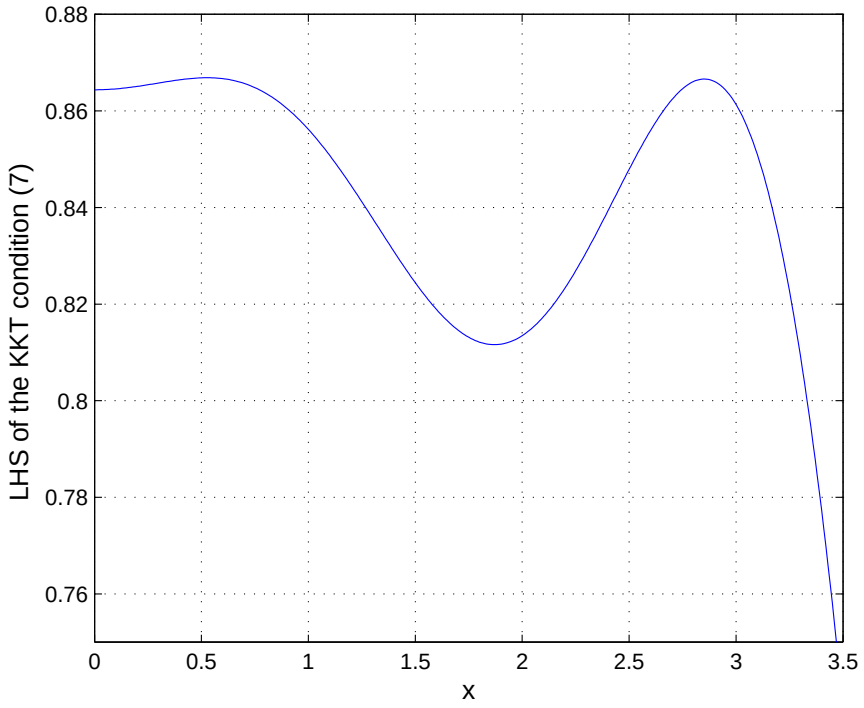

Fig. 3. The left-hand side of the KKT condition (7) for the input distribution generated by the cutting-plane algorithm ($\mathsf{SNR} = 5$ dB). The KKT condition is seen to be satisfied (up to the numerical precision of our computations).

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

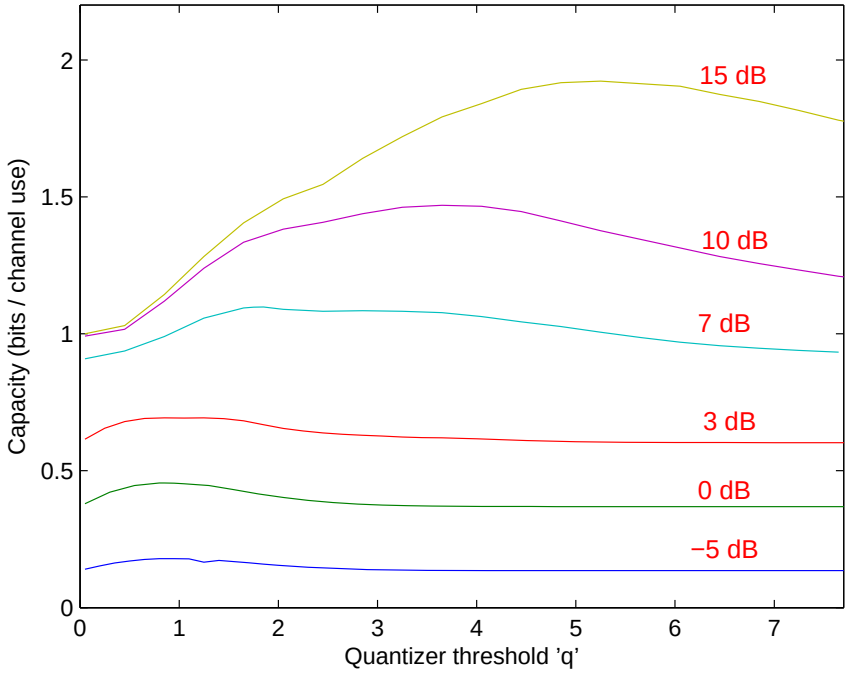

Fig. 4. 2-bit symmetric quantization : channel capacity (in bits per channel use) as a function of the quantizer threshold q. (noise variance $\sigma^2 = 1$.)

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.

A. 2-Bit Symmetric 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 4, 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 4). 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 III, 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 6. 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 5 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

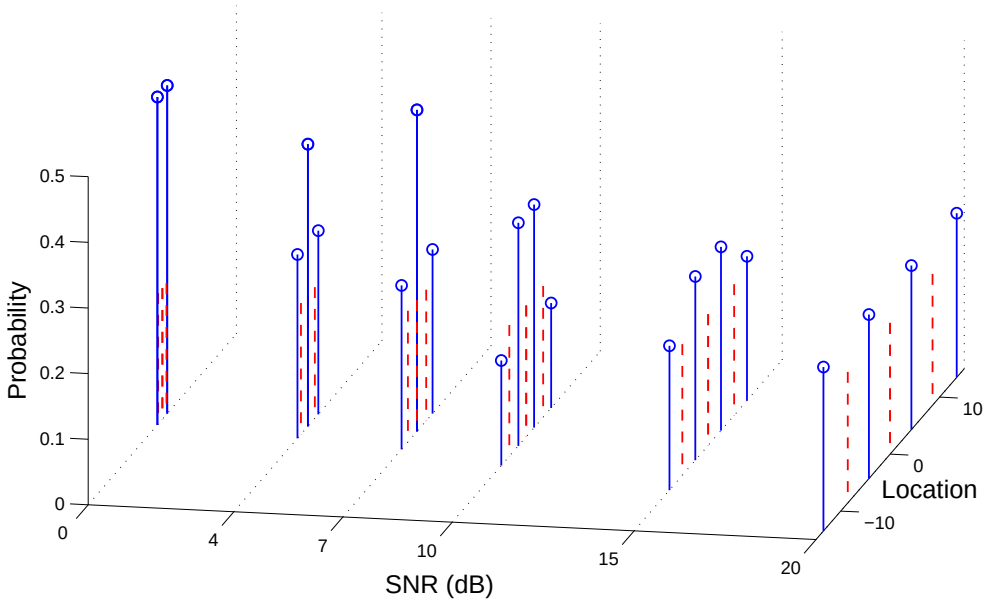

Fig. 5. 2-bit symmetric quantization : optimal input distribution (solid vertical lines) and quantizer thresholds (dashed vertical lines) at various SNRs.

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 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 II-B) is looser, guaranteeing the achievability of capacity by at most five points.

B. 3-bit Symmetric 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 $i = 1$)

- 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 $i = i + 1$.

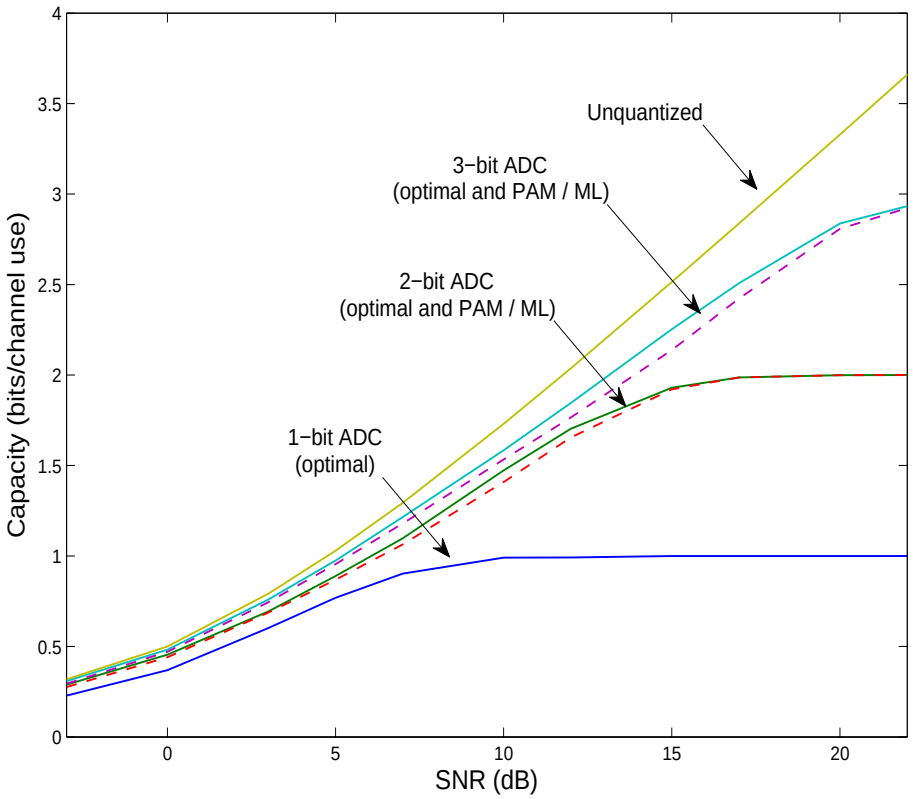

Fig. 6. Capacity plots for different ADC precisions. For 2 and 3-bit ADC, solid curves correspond to optimal solutions, while dashed curves show the performance of the benchmark scheme (PAM input with ML quantization).

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.

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 III and the corresponding plots in Figure 6. 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.

C. 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 6 and the data values are given in Table III. 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, [7] has earlier shown that a large fraction of the capacity can be obtained by 2-bit quantization.

	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 II

SNR (IN DB) REQUIRED TO ACHIEVE A SPECIFIED SPECTRAL EFFICIENCY WITH DIFFERENT ADC PRECISIONS.

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 II). 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.

D. Additive Quantization Noise Model (AQNM)

It is common to model the quantization noise as independent additive noise [30, 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 [30, pp. 122]. As $\mathsf{SNR} \to 0$, 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 III 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.

V. 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. The availability of a

SNR	1-bit ADC		2-bit ADC			SNR	3-bit ADC			Unquantized
(in dB)	Optimal	AQNM	Optimal	Benchmark	AQNM	(in dB)	Optimal	Benchmark	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 III

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, AS WELL AS THE BENCHMARK SOLUTIONS. 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.)

large amount of bandwidth encourages power-efficient communication using small constellations, 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, but fortunately, this is 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 the bandwidth is plentiful.

There are several 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, but this falls outside the scope of the present model.

While our focus in this paper was on non-spread systems, it is known that low-precision ADC is often employed in spread spectrum systems for low cost implementations [31]. In our prior examination of Shannon limits for direct sequence spread spectrum systems with 1-bit ADC [10], we demonstrated that binary signaling was suboptimal, but did not completely characterize an optimal input distribution. The approach in the present paper implies that, for a spreading gain G, a discrete input distribution with at most $G+2$ points can achieve capacity (although in practice, much smaller constellations would probably work well).

Finally, we would like to emphasize that the Shannon-theoretic perspective provided in this paper is but a first step towards the design of communication systems with low-precision ADC. Major technical challenges include the design of ADC-constrained methods for receiver tasks such as carrier and timing synchronization, channel estimation and equalization, demodulation and decoding.

APPENDIX A
ACHIEVABILITY OF CAPACITY

Theorem 3. *[33] 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 [33, p. 128, Thm 2]. □

The use of this optimization theorem to establish the existence of a capacity achieving input distribution is standard (see [25], [23] for details). To use the theorem for our channel model (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 $\mathcal{F}$ [7]. The weak* compactness of $\mathcal{F}$ has been shown in [25]. (The authors in [23] 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,

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

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.

APPENDIX B
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.

[7]The notion of weak* convergence here is actually the same as the standard weak convergence defined in probability theory [32].

Weak Differentiability of Mutual Information: The weak derivative of I at a point $F_0 \in \mathcal{F}$ is defined as ([24], [25])

$$I'_{F_0}(F) = \lim_{\theta \to 0} \frac{I((1-\theta)F_0 + \theta F) - I(F_0)}{\theta} \qquad \forall F \in \mathcal{F} \tag{11}$$

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).

APPENDIX C
PROOF OF PROPOSITION 2

We extend Witsenhausen's result in [5] 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. The result [23, 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 [6], 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 [33, 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 $\mathcal{M}$

$$\mathcal{U} = \{M \in \mathcal{M} | R(y; M) = R^* \text{ and } E(X^2) = P_0\}. \tag{12}$$

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

$$H_i : \int_{A_1}^{A_2} W_i(x) dM(x) = {p_i}^* \quad 1 \leq i \leq K-1 \tag{13}$$

and,

$$H_K : \int_{A_1}^{A_2} x^2 dM(x) = P_0 \tag{14}$$

where $W_i(x)$ are the transition probability functions. Note that there are only $K-1$ hyperplanes in (13) 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 [23, 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. □

APPENDIX D
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 7.

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

$$\ln\left(\frac{1-Q(\sqrt{y})}{Q(\sqrt{y})}\right) \geq \frac{e^{-y/2}}{\sqrt{2\pi y}\ Q(\sqrt{y})(1-Q(\sqrt{y}))} \tag{15}$$

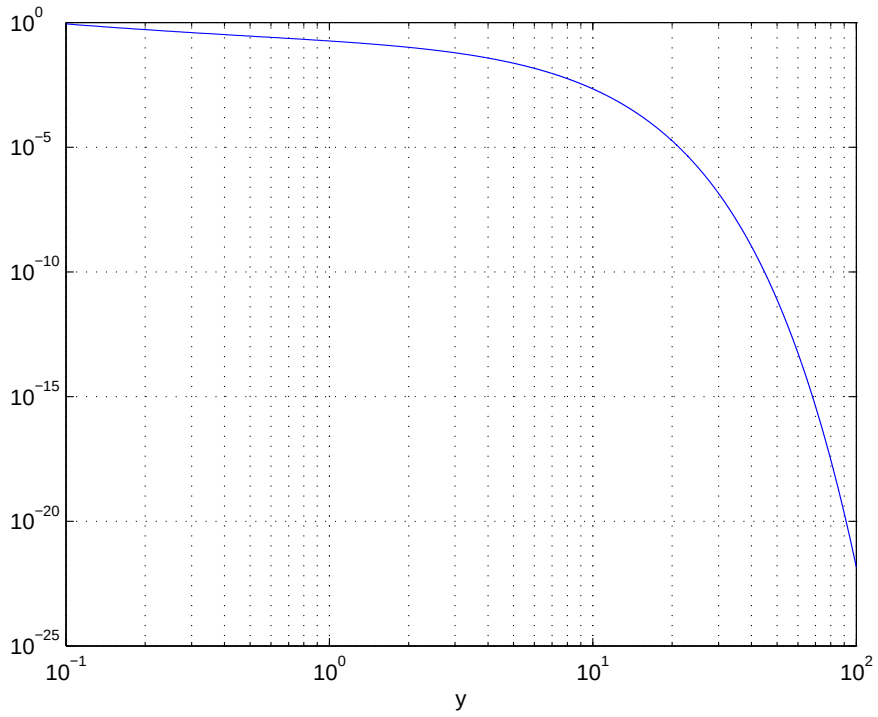

Fig. 7. The second derivative of $h(Q(\sqrt{y}))$ is positive for small values of y.

Using the fact [35, 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 (15) to be true

$$\ln\left(\frac{1 - Q(\sqrt{y})}{Q(\sqrt{y})}\right) \geq \frac{1}{(1 - \frac{1}{y})(1 - Q(\sqrt{y}))} \tag{16}$$

or, equivalently

$$(1 - \frac{1}{y})(1 - Q(\sqrt{y}))\ln\left(\frac{1 - Q(\sqrt{y})}{Q(\sqrt{y})}\right) \geq 1 \tag{17}$$

The left hand side of (17) is a monotone increasing function of y. For $y = 2$, it equals 1.133. Thus (17) holds $\forall y > 2$, and hence the second derivative of $h(Q(\sqrt{y}))$ must be positive for $y > 2$.

ACKNOWLEDGMENTS

The authors thank an anonymous reviewer for suggesting the analysis of the additive quantization noise model in Section IV-D. The first author also thanks Prof. Janos Englander (Department of Statistics and Applied Probability, UCSB), Prof. Shiv Chandrasekaran (Department of Electrical and Computer Engineering, UCSB) and Mr. Vinay Melkote (Signal Compression Laboratory, UCSB) for several useful discussions.

REFERENCES

[1] R. Hiremane, *From Moore's Law to Intel Innovation-Prediction to Reality*, Technology@Intel Magazine, Apr. 2005.

[2] R. Walden, *Analog-to-Digital Converter Survey and Analysis*, IEEE J. Select. Areas Comm., Vol. 17, No. 4, pp. 539-550, Apr. 1999.

[3] J. Huang and S. P. Meyn, *Characterization and Computation of Optimal Distributions for Channel Coding*, IEEE Trans. Info. Theory, Vol. 51, No. 7, pp. 2336-2351, Jul. 2005.

[4] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, New York, 1968.

[5] H. S. Witsenhausen, *Some Aspects of Convexity Useful in Information Theory*, IEEE Trans. Info. Theory, Vol. 26, No. 3, pp. 265-271, May 1980.

[6] L. E. Dubins, *On Extreme Points of Convex Sets*, J. Math. Anal. Appl., Vol. 5, pp. 237-244, May 1962.

[7] N. Phamdo and F. Alajaji, *Soft-Decision Demodulation Design for COVQ over White, Colored, and ISI Gaussian Channels*, IEEE Trans. Comm., Vol. 48, No. 9, pp. 1499-1506, Sep. 2000.

[8] F. Behnamfar, F. Alajaji and T. Linder, *Channel-Optimized Quantization With Soft-Decision Demodulation for Space-Time Orthogonal Block-Coded Channels*, IEEE Trans. Sig. Proc., Vol. 54, No. 10, pp. 3935-3946, Oct. 2006.

[9] J. A. Nossek and M. T. Ivrlac, *Capacity and Coding for Quantized MIMO Systems*, In Proc. Intl. Conf. on Wireless Comm. and Mobile Computing (IWCMC), 2006, Vancouver, Canada.

[10] O. Dabeer, J. Singh and U. Madhow, *On the Limits of Communication Performance with One-Bit Analog-To-Digital Conversion*, In Proc. IEEE Workshop on Signal Proc. Advances in Wireless Comm. (SPAWC), 2006, Cannes, France.

[11] J. Singh, O. Dabeer and U. Madhow, *Communication Limits with Low Precision Analog-To-Digital Conversion at the Receiver*, In Proc. IEEE Intl. Conf. on Communications (ICC), 2007, Glasgow, Scotland.

[12] J. Singh, O. Dabeer and U. Madhow, *Capacity of the Discrete-Time AWGN Channel Under Output Quantization*, In Proc. IEEE Intl. Symp. on Info. Theory (ISIT), 2008, Toronto, Canada.

[13] L. N. Lee, *On Optimal Soft Decision Demodulation*, IEEE Trans. Info. Theory, Vol. 22, pp. 437-444, Jul. 1976.

[14] J. Salz and E. Zehavi, *Decoding under Integer Metrics Constraints*, IEEE Trans. Comm., Vol. 43, No. 3, pp. 307-317, Mar. 1995.

[15] E. Masry, *The Reconstruction of Analog Signals from the Sign of Their Noisy Samples*, IEEE Trans. on Info. Theory, Vol. 27, pp. 735-745, Nov. 1981.

[16] Z. Cvetkovic and I. Daubechies, *Single-bit Oversampled A/D Conversion with Exponential Accuracy in the Bit-Rate*, In Proc. Data Compression Conference (DCC), 2000, Utah, USA.

[17] P. Ishwar, A. Kumar and K. Ramachandran, *Distributed Sampling for Dense Sensor Networks: A Bit-Conservation Principal*, In Proc. Info. Proc. in Sensor Networks (IPSN), 2003, Palo Alto, USA.

[18] A. Host-Madsen and P. Handel, *Effects of Sampling and Quantization on Single-Tone Frequency Estimation*, IEEE Trans. on Sig. Proc., Vol. 48, pp. 650-662, Mar. 2000.

[19] T. Andersson, M. Skoglund, and P. Handel, *Frequency Estimation by 1-bit Quantization and Table Look-Up Processing*, In Proc. Euro. Sig. Proc. Conf. (EUSIPCO), 2000, Tampere, Finland.

[20] D. Rousseau, G. V. Anand, and F. Chapeau-Blondeau, *Nonlinear Estimation from Quantized Signals: Quantizer Optimization and Stochastic Resonance*, In Third Intl. Symp. on Physics in Signal and Image Proc. (PSIP), 2003, Grenoble, France.

[21] O. Dabeer and A. Karnik, *Signal Parameter Estimation Using 1-bit Dithered Quantization*, IEEE Trans. Info. Theory, Vol. 52, No. 12, pp. 5389-5405, Dec. 2006.

[22] O. Dabeer and E. Masry, *Multivariate Signal Parameter Estimation Under Dependent Noise From 1-bit Dithered Quantized Data*, IEEE Trans. Info. Theory, Vol. 54, No. 4, pp. 1637-1654, Apr. 2008.

[23] M. Fozunbal, S. W. McLaughlin and R. W. Schafer, *Capacity Analysis for Continuous-Alphabet Channels With Side Information, Part I: A General Framework*, IEEE Trans. Info. Theory , Vol. 51, No. 9, pp. 3075-3085, Sep. 2005.

[24] J. G. Smith, *On the Information Capacity of Peak and Average Power Constrained Gaussian Channels*, Ph.D. Dissertation, Univ. of California, Berkeley, Dec. 1969.

[25] I. C. Abou-Faycal, M. D. Trott and S. Shamai, *The Capacity of Discrete-Time Memoryless Rayleigh Fading Channels*, IEEE Trans. Info. Theory, Vol. 47, No. 4, pp. 1290-1301, May 2001.

[26] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, New York, 2nd Edition.

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

[28] M. Chiang and S. Boyd, *Geometric Programming Duals of Channel Capacity and Rate Distortion*, IEEE Trans. Info. Theory, Vol. 50, No. 2, pp. 245-258, Feb. 2004.

[29] A. Lapidoth and S. M. Moser, *Capacity Bounds via Duality with Applications to Multiple-Antenna Systems on Flat Fading Channels*, IEEE Trans. Info. Theory, Vol. 49, No. 10, pp. 2426-2467, Oct. 2003

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

[31] M. S. Braasch, A. J. van Dierendonck, *GPS Receiver Architectures and Measurements*, Proc. of the IEEE, Vol. 87, No. 1, pp. 48-64, Jan. 1999.

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

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

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

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

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