Capacity of the Discrete-Time AWGN Channel Under Output Quantization

Jaspreet Singh, Onkar Dabeer and Upamanyu Madhow*

Abstract— **We investigate the limits of communication over the discrete-time Additive White Gaussian Noise (AWGN) channel, when the channel output is quantized using a small number of bits. We first provide a proof of our recent conjecture on the optimality of a discrete input distribution in this scenario. Specifically, we show that for any given output quantizer choice with K quantization bins (i.e., a precision of $\log_2 K$ bits), the input distribution, under an *average power* constraint, need not have any more than $K + 1$ mass points to achieve the channel capacity. The cutting-plane algorithm is employed to compute this capacity and to generate optimum input distributions. Numerical optimization over the choice of the quantizer is then performed (for 2-bit and 3-bit symmetric quantization), and the results we obtain show that the loss due to low-precision output quantization, which is small at low signal-to-noise ratio (SNR) as expected, can be quite acceptable even for moderate to high SNR values. For example, at SNRs up to 20 dB, 2-3 bit quantization achieves 80-90% of the capacity achievable using infinite-precision quantization.**

I. INTRODUCTION

Analog-to-digital conversion (ADC) is an integral part of modern communication receiver architectures based on digital signal processing (DSP). Typically, ADCs with 6-12 bits of precision are employed at the receiver to convert the received analog baseband signal into digital form for further processing. However, as the communication systems scale up in speed and bandwidth (for e.g., systems operating in the ultrawide band or the mm-wave band), the cost and power consumption of such high precision ADC becomes prohibitive [1]. A DSP-centric architecture nonetheless remains attractive, due to the continuing exponential advances in digital electronics (Moore's law). It is of interest, therefore, to understand whether DSP-centric design is compatible with the use of low-precision ADC.

In this paper, we continue our investigation of the Shannon-theoretic communication limits imposed by the use of low-precision ADC for ideal Nyquist sampled linear modulation in AWGN. The discrete-time memoryless AWGN-*Quantized Output* (AWGN-QO) channel model thus induced is shown in Fig. 1. In our prior work for this channel model, we have shown that for the extreme scenario of 1-bit symmetric quantization, binary antipodal signaling achieves the channel
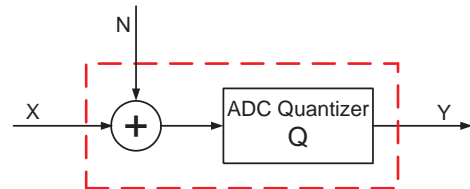
Fig. 1. The *AWGN-Quantized Ouput* Channel : $Y = Q(X + N)$.

capacity for any signal-to-noise ratio (SNR) [2]. For multi-bit quantization [3], we provided a duality-based approach to bound the capacity from above, and employed the cutting-plane algorithm to generate input distributions that nearly achieved these upper bounds. Based on our results, we conjectured that a discrete input with cardinality not exceeding the number of quantization bins achieves the capacity of the *average power* constrained AWGN-QO channel. In this work, we prove that a discrete input is indeed optimal, although our result only guarantees its cardinality to be *at most* $K + 1$, where K is the number of quantization bins. Our proof is inspired by Witsenhausen's result in [4], where Dubins' theorem [5] was used to show that the capacity of a discrete-time memoryless channel with output cardinality K, under only 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 use Dubins' theorem as done by Witsenhausen. Although not applicable to our setting, it is worth noting that for a Discrete Memoryless Channel, Gallager first showed that the number of inputs with nonzero probability mass need not exceed the number of outputs [6, p. 96, Corollary 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 do this numerically for 2-bit and 3-bit symmetric quantization, and use our numerical results to make the following encouraging observations: (a) Low-precision ADC incurs a relatively small loss in spectral efficiency compared to unquantized observations. While this is expected for low SNRs, we find that even at moderately high SNRs of up to 20 dB, 2-3 bit ADC still achieves 80-90% of the spectral efficiency attained using unquantized observations. These results indicate the feasibility of system design using low-precision ADC for high bandwidth systems. (b) Standard uniform Pulse Amplitude Modulated (PAM) input with quantizer thresholds

set to implement maximum likelihood (ML) hard decisions achieves nearly the same performance as that attained by an optimal input and quantizer pair. This is useful from a system designer's point of view, since the ML quantizer thresholds have a simple analytical dependence on SNR, which is an easily measurable quantity.

The rest of the paper is organized as follows. The quantized output AWGN channel model is given in the next section. In Section III, we show that a discrete input achieves the capacity of this channel. Quantizer optimization results are presented in Section IV, followed by the conclusions in Section V.

II. CHANNEL MODEL

We consider linear modulation over a real AWGN channel, and assume that the Nyquist criterion for no intersymbol interference is satisfied [7, pp. 50]. Symbol rate sampling of the receiver's matched filter output using a finite-precision ADC therefore results in the following discrete-time memoryless AWGN-*Quantized Output* (AWGN-QO) channel (Fig. 1)

$$Y = Q(X + N) . \qquad (1)$$

Here $X \in \mathbb{R}$ is the channel input with distribution $F(x)$ and N is $\mathcal{N}(0, \sigma^2)$. The quantizer Q maps the real valued input $X + N$ to one of the K bins, producing a discrete channel output $Y \in \{y_1, \cdots, y_K\}$. We only consider quantizers for which each bin is an interval of the real line. The quantizer Q with K bins can therefore be 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 resulting transition probability functions are given by

$$W_i(x) = P(Y = y_i | X = x) = Q\left(\frac{q_{i-1} - x}{\sigma}\right) - Q\left(\frac{q_i - x}{\sigma}\right),$$
$$(2)$$

where $Q(x)$ denotes the complementary Gaussian distribution function $\frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-t^2/2) dt$.

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

$$I(F) = \int_{-\infty}^\infty \sum_{i=1}^K W_i(x) \log \frac{W_i(x)}{R(y_i; F)} dF(x) , \qquad (3)$$

where $\{R(y_i; F), 1 \leq i \leq K\}$ is the Probability Mass Function (PMF) of the output when the input is F. Under an average power constraint P (i.e., $\mathbb{E}[X^2] \leq P$), we wish to compute the capacity of the channel (1), which is given by

$$C = \sup_{F \in \mathcal{F}} I(F), \qquad (4)$$

where $\mathcal{F}$ is the set of all average power constrained distributions on $\mathbb{R}$.

III. DISCRETE INPUT ACHIEVES CAPACITY

We first use the Karush-Kuhn-Tucker (KKT) optimality condition to show that an average power constraint for the AWGN-QO channel automatically induces a constraint on the peak power, in the sense that an optimal input distribution must have a bounded support set. This fact is then exploited to show the optimality of a discrete input.

A. An Implicit peak power Constraint

The following KKT condition can be derived for the AWGN-QO channel, using convex optimization principles in a manner similar to that in [8], [9]. The input distribution F is optimal if and only if there exists a $\gamma \geq 0$ such that

$$\sum_{i=1}^K W_i(x) \log \frac{W_i(x)}{R(y_i; F)} + \gamma(P - x^2) \leq I(F) , \qquad (5)$$

for all x, with equality if x is in the support of F.

The first term on the left hand side of the KKT condition (5) is the divergence (or the relative entropy) between the transition and the output PMFs. For convenience, let us denote it by $d(x; F)$. The following result concerning the behavior of $d(x; F)$ has been proved in [10].

Lemma 1: For the AWGN-QO channel (1) with input distribution F, 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 $\forall x > A_0$, $d(x; F) < -\log R(y_K; F)$.
   Proof: See [10].

We now use Lemma 1 to prove the main result of this subsection.

Proposition 1: A capacity-achieving input distribution for the average power constrained AWGN-QO channel (1) must have bounded support.
Proof: Assume that the input distribution F^* achieves[1] the capacity in (4) (i.e., $I(F^*) = C$), with $\gamma^* \geq 0$ being a corresponding optimal Lagrange parameter in the KKT condition. In other words, with $\gamma = \gamma^*$, and, $F = F^*$, (5) must be satisfied with an equality at every point in the support of F^*. We exploit this necessary condition next 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 (5) for any $x > A_2^*$.

Using Lemma 1, we first let $\lim_{x \to \infty} d(x; F^*) = -\log(R(y_K; F^*)) = L$, and also assume that there exists a finite constant A_0 such that $\forall x > A_0$, $d(x; F^*) < L$. We consider two possible cases.

- Case 1: $\gamma^* > 0$.
  If $C > L + \gamma^* P$, then pick $A_2^* = A_0$.
  Else pick $A_2^* \geq \max\{A_0, \sqrt{(L + \gamma^* P - C)/\gamma^*}\}$.
  In either situation, for $x > A_2^*$, we get $d(x; F^*) < L$, and, $\gamma^* x^2 > L + \gamma^* P - C$.
  This gives

  $$d(x; F^*) + \gamma^*(P - x^2) < L + \gamma^* P - (L + \gamma^* P - C) = C.$$

- Case 2: $\gamma^* = 0$.
  Putting $\gamma^* = 0$ in the KKT condition (5), we get

  $$d(x; F^*) = \sum_{i=1}^K W_i(x) \log \frac{W_i(x)}{R(y_i; F^*)} \leq C , \quad \forall x.$$

---

[1]That the capacity is achievable can be shown using standard results from optimization theory. For lack of space here, we refer the reader to [10] for details.

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

Picking $A_2{}^* = A_0$, we therefore have that for $x > A_2{}^*$
$$d(x; F^*) + \gamma^*(P - x^2) = d(x; F^*) < L.$$
$$\implies d(x; F^*) + \gamma^*(P - x^2) < C.$$

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 a bounded support. ∎

B. Achievability of Capacity by a Discrete Input

To show the optimality of a discrete input for our problem, we use the following theorem which we have proved in [10]. The theorem holds for channels with a finite output alphabet, under the condition that the input is constrained in both peak power and average power.

Theorem 1: Consider a stationary discrete-time memoryless channel with a continuous input X taking values in the bounded interval $[A_1, A_2]$, and a discrete output $Y \in \{y_1, y_2, \cdots, y_K\}$. Let the transition probability function $W_i(x) = \mathsf{P}(Y = y_i | X = x)$ be continuous in x, for each i in $\{1, .., K\}$. The capacity of this channel, under an average power constraint on the input, is achievable by a discrete input with at most $K + 1$ points.
Proof: See [10]. ∎

Our proof in [10] uses Dubins' theorem [5], and is an extension of Witsenhausen's result in [4], wherein he showed that a distribution with only K points would be sufficient to achieve the capacity if the average power of the input was not constrained.

The implicit peak power constraint derived in Section III-A allows us to use Theorem 1 to get the following result.

Proposition 2: 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: Using notation from the last subsection, let F^* be an optimal distribution for (4), with the support of F^* being contained in the bounded interval $[A_1{}^*, A_2{}^*]$. Define $\mathcal{F}_1$ to be the set of all average power constrained distributions whose support is contained in $[A_1{}^*, A_2{}^*]$. Note that $F^* \in \mathcal{F}_1 \subset \mathcal{F}$, where $\mathcal{F}$ is the set of all average power constrained distributions on $\mathbb{R}$. Consider the maximization of the mutual information $I(X; Y)$ over the set $\mathcal{F}_1$
$$C_1 = \max_{F \in \mathcal{F}_1} I(F). \tag{6}$$

Since the transition probability functions in (2) are continuous in x, Theorem 1 implies that a discrete distribution with at most $K + 1$ mass points achieves the maximum C_1 in (6). Denote such a distribution by F_1. However, since F^* achieves the maximum C in (4) and $F^* \in \mathcal{F}_1$, it must also achieve the maximum in (6). This implies that $C_1 = C$, and that F_1 is optimal for (4), thus completing the proof. ∎

C. Capacity Computation

We have already addressed the issue of computing the capacity (4) in our prior work. Specifically, in [2], we have shown analytically that for the extreme scenario of 1-bit symmetric quantization, binary antipodal signaling achieves the capacity (at any SNR). Multi-bit quantization has been considered in [3], [10], where we show that the cutting-plane algorithm [11] can be employed for computing the capacity and obtaining optimal input distributions.

IV. OPTIMIZATION OVER QUANTIZER

Until now, we have addressed the problem of capacity computation given a fixed quantizer. In this section, we consider the issue of quantizer optimization, while restricting attention to *symmetric quantizers only*. Given the symmetric nature of the AWGN noise and the power constraint, it seems intuitively plausible that restriction to symmetric quantizers should not be sub-optimal from the point of view of optimizing over the quantizer choice in (1), although a proof of this conjecture has eluded us.

A Simple Benchmark: While an optimal quantizer (with a corresponding optimal input) provides the absolute communication limits for our model, from a system designer's perspective, it would also be useful to evaluate the performance degradation if we use some standard input constellations and quantizer choices. We take the following input and quantizer pair as our *benchmark strategy* : for K-bin quantization, consider equispaced uniform K-PAM (Pulse Amplitude Modulated) input distribution, with quantizer thresholds as the mid-points of the input mass point locations (i.e., ML hard decisions). With the K-point uniform input, we have the entropy $H(X) = \log_2 K$ bits for any SNR. Also, it is easy to see that as SNR $\to \infty$, $H(X|Y) \to 0$ for the benchmark input-quantizer pair. Therefore, our benchmark scheme is near-optimal if we operate in the high SNR regime. The main issue to investigate ahead, therefore is: at low to moderate SNRs, how much gain does an optimal quantizer choice provide over the benchmark.

In all the results that follow, we take the noise variance $\sigma^2 = 1$. However, 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 an identical performance.

NUMERICAL RESULTS

A. 2-bit Symmetric Quantization

A 2-bit symmetric quantizer is characterized by a single parameter q, with $\{-q, 0, q\}$ being the quantizer thresholds. Hence we use a brute force search over q to optimize the quantizer. In Fig. 2, we plot the variation of the channel capacity (computed using the cutting-plane algorithm) as a function of the parameter q at various SNRs. We observe that for any SNR, there is an optimal choice of q that maximizes the capacity. At high SNRs, the optimal q is seen to increase monotonically with SNR, which is not surprising since the
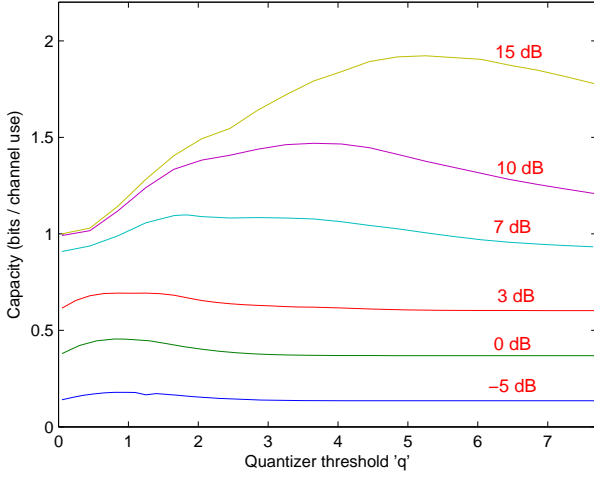
Fig. 2. 2-bit symmetric quantization : channel capacity versus the quantizer threshold q (noise variance $\sigma^2 = 1$).

SNR(dB)	-10	-5	0	7	15
1-bit optimal	0.0449	0.1353	0.3689	0.9020	0.9974
2-bit optimal	0.0613	0.1792	0.4552	1.0981	1.9304
2-bit benchmark	0.0527	0.1658	0.4401	1.0639	1.9211

TABLE I

MUTUAL INFORMATION (IN BITS/CHANNEL USE) AT DIFFERENT SNRS.

benchmark quantizer's q scales as $\sqrt{\text{SNR}}$ and is known to be near-optimal at high SNRs.

Comparison with the benchmark: In Table I, we compare the performance of the optimal solution obtained as above with the benchmark scheme. The capacity with 1-bit quantization is also shown for reference. While being near-optimal at moderate to high SNRs, the benchmark scheme is seen to perform fairly well at low SNRs also. For instance, at -10 dB SNR, it achieves 86% of the capacity achieved with an optimal 2-bit quantizer and input pair. From a practical standpoint, these results imply that the benchmark scheme, which requires negligible computational effort (due to its well-defined dependence on SNR), can be employed even at small SNRs while incurring an acceptable loss of performance.
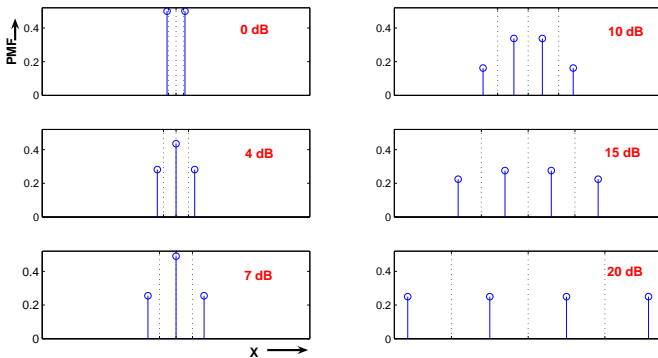

Fig. 3. 2-bit symmetric quantization : optimal input distribution and quantizer at various SNRs (the dashed vertical lines depict the locations of the quantizer thresholds).

SNR(dB)	-10	0	5	10	20
3-bit optimal	0.0667	0.4817	0.9753	1.5844	2.8367
3-bit benchmark	0.0557	0.4707	0.9547	1.5332	2.8084

TABLE II

MUTUAL INFORMATION (IN BITS/CHANNEL USE) AT DIFFERENT SNRS.

Optimal Input Distributions: The optimal input distributions (given by the cutting-plane algorithm) corresponding to the optimal quantizers obtained above are depicted in Fig. 3, for different SNR values. The locations of the optimal quantizer thresholds are also shown (by the dashed vertical lines). Binary signaling is found to be optimal at low SNRs, and the number of mass points increases (first to 3 and then to 4) with increasing SNR. Further increase in SNR eventually leads to the uniform 4-PAM input, thus approaching the capacity bound of 2 bits. It is worth noting that all the optimal inputs we obtained have 4 or less mass points, whereas Proposition 2 is looser as it guarantees the achievability of capacity using at most 5 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 $\{\pm q_1, \pm q_2, \pm q_3\}$. Instead of brute force search, we use an alternate optimization procedure for joint optimization of the input and the quantizer in this case. Due to lack of space, we refer the reader to [10] for details, and proceed directly to the numerical results. (Table II)

Comparison with the benchmark: As for 2-bit quantization considered earlier, we find that the benchmark scheme performs quite well at low SNRs with 3-bit quantization also. At -10 dB SNR, for instance, the benchmark scheme achieves 83% of the capacity achievable with an optimal quantizer choice. Table II gives the comparison for different SNRs.

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 Proposition 2 guarantees the achievability of capacity by at most $K+1$ points. Of course, Proposition 2 is applicable to any quantizer choice (and not just optimal symmetric quantizers that we consider in this section), it still leaves us with the question whether it can be tightened to guarantee achievability of capacity with at most K points.

C. Comparison with Unquantized Observations

We now compare the capacity results obtained above with the case when the receiver ADC has infinite precision. Table III provides these results, and the corresponding plots are shown in Fig. 4. We observe that at low SNRs, low-precision quantization is a very feasible option. For instance, at -5 dB SNR, even 1-bit receiver quantization achieves 68% of the capacity achievable with infinite-precision. 2-bit quantization at the same SNR provides as much as 90% of the infinite-precision capacity. Such high figures are understandable, since if noise dominates the message signal, increasing the quantizer

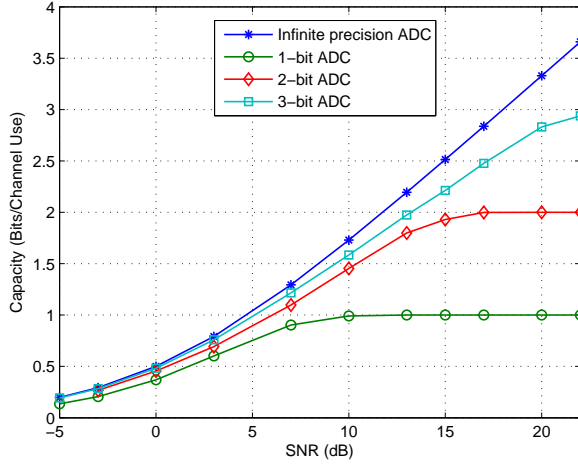| SNR(dB)	−5	0	5	10	15
1-bit ADC	0.1353	0.3689	0.7684	0.9908	0.9999
2-bit ADC	0.1792	0.4552	0.8889	1.4731	1.9304
3-bit ADC	0.1926	0.4817	0.9753	1.5844	2.2538
Unquantized	0.1982	0.5000	1.0286	1.7297	2.5138

TABLE III

CAPACITY (IN BITS/CHANNEL USE) AT VARIOUS SNRs.

Fig. 4. Capacity with 1-bit, 2-bit, 3-bit, and infinite-precision ADC.

V. CONCLUSIONS

Our Shannon-theoretic investigation indicates the feasibility of low-precision ADC for designing future high-bandwidth communication systems such as those operating in UWB and mm-wave band. The small reduction in spectral efficiency due to low-precision ADC is acceptable in such systems, given that the available bandwidth is plentiful. Current research is therefore focussed on developing ADC-constrained algorithms to perform receiver tasks such as carrier and timing synchronization, channel estimation and equalization.

An unresolved technical issue concerns the number of mass points required to achieve capacity. While we have shown that the capacity for the AWGN channel with K-bin output quantization is achievable by a discrete input distribution with at most $K + 1$ points, numerical computation of optimal inputs reveals that K mass points are sufficient. Can this be proven analytically, at least for symmetric quantizers? Are symmetric quantizers optimal? Another problem for future investigation is whether our result regarding the optimality of a discrete input can be generalized to other channel models. Under what conditions is the capacity of an average power constrained channel with output cardinality K achievable by a discrete input with at most $K + 1$ points?

precision beyond a point does not help much in distinguishing between different signal levels. However, we surprisingly find that even if we consider moderate to high SNRs, the loss due to low-precision sampling is still very acceptable. At 10 dB SNR, for example, the corresponding ratio for 2-bit quantization is still a very high 85%, while at 20 dB, 3-bit quantization is enough to achieve 85% of the infinite-precision capacity. Similar encouraging results have been reported earlier in [12], [13] also. However, the input alphabet in these works was taken as binary to begin with, in which case the good performance with low-precision output quantization is perhaps less surprising.

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), we find that 2-bit quantization incurs a loss of 2.30 dB (see Table IV). From a practical viewpoint, this penalty in power is more significant compared to the 15% loss in spectral efficiency on using 2-bit quantization at 10 dB SNR. This suggests, for example, that the impact of low-precision ADC should be weathered by a moderate reduction in the spectral efficiency, rather than by increasing the transmit power.

REFERENCES

[1] R. Walden, *Analog-to-Digital Converter Survey and Analysis*, IEEE J. Select. Areas Comm., 17(4):539–550, Apr. 1999.

[2] O. Dabeer, J. Singh and U. Madhow, *On the Limits of Communication Performance with One-Bit Analog-To-Digital Conversion*, In Proc. SPAWC'2006, Cannes, France.

[3] J. Singh, O. Dabeer and U. Madhow, *Communication Limits with Low-Precision Analog-To-Digital Conversion at the Receiver*, In Proc. ICC'2007, Glasgow, Scotland.

[4] H.S. Witsenhausen, *Some Aspects of Convexity Useful in Information Theory*, IEEE Tran. Info. Theory, 26(3):265–271, May 1980.

[5] L.E. Dubins, *On Extreme Points of Convex Sets*, J. Math. Anal. Appl., 5:237–244, May 1962.

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

[7] U. Madhow, *Fundamentals of Digital Communications*, Cambridge University Press, 2008.

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

[9] I. C. Abou-Faycal, M. D. Trott and S. Shamai, *The Capacity of Discrete-Time Memoryless Rayleigh Fading Channels*, IEEE Tran. Info. Theory, 47(4):1290–1301, May 2001.

[10] J. Singh, O. Dabeer and U. Madhow, *Transceiver Design with Low-Precision Analog-to-Digital Conversion : An Information-Theoretic Perspective*, Submitted to IEEE Tran. Info. Theory, Mar 2008. Available online at $http : //arxiv.org/PS\_cache/arxiv/pdf/0804/0804.1172v1.pdf$

[11] J. Huang and S. P. Meyn, *Characterization and Computation of Optimal Distributions for Channel Coding*, IEEE Tran. Info. Theory, 51(7):2336–2351, Jul. 2005.

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

[13] J. A. Nossek and M. T. Ivrlac, *Capacity and Coding for Quantized MIMO Systems*, In Proc. IWCMC, 2006.

	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 IV

SNR (IN dB) REQUIRED FOR A GIVEN SPECTRAL EFFICIENCY.