An Information-Theoretic Framework for Receiver Quantization in Communication

Jing Zhou, *Member, IEEE*, Shuqin Pang, and Wenyi Zhang, *Senior Member, IEEE*

Abstract

We investigate information-theoretic limits and design of communication under receiver quantization. Unlike most existing studies that focus on low-resolution quantization, this work is more focused on the impact of weak nonlinear distortion due to resolution reduction from high to low. We consider a standard transceiver architecture, which includes independent and identically distributed (i.i.d.) complex Gaussian codebook at the transmitter, and a symmetric quantizer cascaded with a nearest neighbor decoder at the receiver. Employing the generalized mutual information (GMI), an achievable rate under general quantization rules is obtained in an analytical form, which shows that the rate loss due to quantization is $\log\left(1+\gamma\mathsf{SNR}\right)$, where SNR is the signal-to-noise ratio at the receiver front-end, and γ is determined by thresholds and levels of the quantizer. Based on this result, the performance under uniform receiver quantization is analyzed comprehensively. We show that the front-end gain control, which determines the loading factor (normalized one-sided quantization range) of quantization, has an increasing impact on performance as the resolution decreases. In particular, we prove that the unique loading factor that minimizes the mean square error (MSE) of the uniform quantizer also maximizes the GMI, and the corresponding irreducible rate loss is given by $\log\left(1+\mathsf{mmse}\cdot\mathsf{SNR}\right)$, where mmse is the minimum MSE normalized by the variance of quantizer input, and is equal to the minimum of γ. A geometrical interpretation for the optimal uniform quantization at the receiver is further established. Moreover, by asymptotic analysis, we characterize the impact of biased gain control, including how small rate losses decay to zero and achievable rate approximations under large bias. From asymptotic expressions of the optimal loading factor and mmse, approximations and several "per-bit rules" for performance are also provided. Finally we discuss more types of receiver quantization and show that the consistency between achievable rate maximization and MSE minimization does not hold in general.

Index Terms

Achievable rate, analog-to-digital converter, Gaussian channel, generalized mutual information, nearest neighbor decoding rule, mean square error, MMSE, transceiver design, uniform quantization.

CONTENTS

This work was supported in part by the National Natural Science Foundation of China through Grant 62231022 and in part by Henan Key Laboratory of Visible Light Communications through Grant HKLVLC2023-B03. The material in this paper will be presented in part at the IEEE International Symposium on Information Theory (ISIT), Ann Arbor, MI, USA, June 2025 [1]. (*Corresponding author: Wenyi Zhang.*)

Jing Zhou is with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing 312000, China (e-mail: jzhou@usx.edu.cn).

Shuqin Pang and Wenyi Zhang are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: shuqinpa@mail.ustc.edu.cn, wenyizha@ustc.edu.cn).

I. Introduction

THE analog-to-digital conversion (ADC), including sampling and quantization, is essential for any digital receiver. The power dissipation of state-of-the-art ADCs increases four times as the resolution increases by one bit, while every doubling of the sampling rate leads to a one-bit loss of resolution [2], [3]. The impact of ADC on the performance has received increasing attention along with the recent evolution of wireless communications, in which several challenges are faced, such as the increasing processing speed due to the utilization of larger bandwidth in mmWave and higher frequencies, the increasing scale of hardware due to the use of massive multiple-input-multiple-output (MIMO), and the critical need for low cost energy-efficient devices in emerging scenarios, e.g., massive machine-type communications (mMTC).

A majority of the studies on ADC at communication receivers have focused on the performance and design under low-resolution output quantization, and one-bit quantization has been of particular interest due to its negligible power dissipation and simplicity of implementation, even without requiring automatic gain control (AGC). In such studies the end-to-end channel is highly nonlinear, typically incurring a substantial performance loss, and necessitating a rethinking of the transceiver design.

On the other side, the transceiver architecture used in present wireless systems is built without considering the effect of output quantization. It is thus necessary to ask, under such conventional transceiver architecture, how much is the loss caused by output quantization with *moderate to high* resolution? In other words, if a small loss in achievable rate is acceptable, how fine need the quantization be? Analytical results on these problems appear to be lacking. Moreover, limited resolution of quantization leads to new problems, e.g., sensitivity of performance to the error of gain control, residual interference in multiuser systems, and so on. These largely unexplored problems prompt us to revisit the topic of receiver quantization in communication in this work.

A. Related Work on Quantization at Communication Receivers

We begin from the impact of output quantization in the (discrete-time) additive white Gaussian noise (AWGN) channel, which is a benchmark model in communication theory. The performance gain of using more output quantization levels in *coded* transmission (i.e., soft-decision decoding) was observed very early in [4] via an information theoretic approach. In the classic textbook of Wozencraft and Jacobs [5, Chap. 6.2], cutoff rate analysis showed that, for equiprobable uniformly spaced pulse amplitude modulation (PAM) input, the degradation due to output quantization is approximately 2 dB when the alphabet size equals to the number of quantization levels, and the degradation vanishes when the quantization becomes increasingly fine. Particularly, in the low-signal-to-noise-ratio (low-SNR) limit, hard-decision decoding (one-bit quantization that observes

the sign of output) leads to a power loss of $\pi/2$ (approximately 2 dB) [4], [5]; see also [6, Chap. 2.11 and 3.4].[1] For K-level output quantization, it was proved that a discrete input of at most $K+1$ mass points suffices to achieve the constrained capacity [8].

In vector (MIMO) channels, the quantization loss in achievable rate can be very small at low-to-moderate SNR even if 1~3-bit output quantization is used. This fact was shown in information theoretic studies [9]–[11] by numerical examples for full-rank channels with multiplexing gains 2~4. Although the constrained capacity is still unknown even in the one-bit case, recent information theoretic studies have provided various results on MIMO systems with coarse output quantization [12]–[19], which show that proper transceiver design is critical in realizing efficient communication in such systems. In particular, for one-bit output quantization, it was shown that the high-SNR capacity grows linearly with the rank of channel [13], while the low-SNR asymptotic power loss of $\pi/2$ still exists in the case of vector channel [12].

In light of these positive theoretical results, performance and design of wireless systems with low-resolution quantization have been extensively studied in recent years; see, e.g., [20]–[26]. The most common approach therein, however, is not information-theoretic (since exact evaluation of mutual information can be difficult). Instead, achievable rate estimation based on the additive quantization noise model (AQNM) has been widely used, which comes from Bussgang-like decomposition [27]–[29]. Results in [12]–[14], [20] suggested that such estimation approximates the mutual information well at low SNR, but becomes inaccurate at high SNR.

Although mutual information is a fundamental performance measure, for communications under transceiver nonlinearity it has limited operational meaning, in the sense that the decoder that achieves the predicted rate can be too complex to implement, while that rate is not necessarily achievable by a standard transceiver architecture designed without considering nonlinearity because the decoder is typically *mismatched* to the nonlinear channel. In [30], [31], a more meaningful performance measure that takes decoding rule into account, namely the generalized mutual information (GMI) [32], has been adopted, yielding analytical expressions of the achievable rate under output quantization and *nearest neighbor decoding rule*. Under a given (possibly mismatched) decoder, the GMI determines the highest rate below which the average probability of error, averaged over a given i.i.d. codebook ensemble, converges to zero as the block length N grows without bound, and it is thus a lower bound on mismatch capacity [32], [33]. The GMI has been applied in various scenarios for performance evaluation, including the bit-interleaved coded modulation (BICM) [34], fading channels [35]–[37], and nonlinear fiber-optic channels [38]. In fact, the rate estimation based on the AQNM (or Bussgang-like decomposition) is consistent with the GMI for scalar channel under Gaussian input and nearest neighbor decoding [30]; see [39] for more discussions. However, GMI analyses also show that the AQNM-based estimation is not accurate in general. For example, it may overestimate the achievable rate in multiantenna systems [40].

B. Related Work in Quantization Theory

The rich theory of quantization was surveyed comprehensively in [41], in which two well established asymptotic theories were emphasized. The first is Shannon's information theoretic approach (rate distortion theory [42]), which places quantization in the framework of lossy source coding and focuses on the high-dimension regime, thereby shedding light on vector quantization. The second is the asymptotic quantization theory, which sheds light on quantizer design in the high-resolution regime. The asymptotic quantization theory is more relevant to receiver quantization in communication which typically does not employ coding or vector quantization. Although not applicable directly to receiver quantization, some classical results in asymptotic quantization theory, especially those for uniform scalar quantization, are reviewed here for comparison.

A basic result known in [43] and [44] (rigorously proved in [45]) states that, for a high-resolution uniform quantizer with step size ℓ, the mean square error (MSE) can be approximated by $\ell^2/12$. This yields the "6-dB-per-bit rule" that each additional bit in resolution reduces the MSE by 6.02 dB. The rule reflects the impact of step size that causes *granular* distortion; but it ignores *overload* distortion due to finite quantization range. The interplay between these two types of distortion determines how the optimal quantization range scales with the resolution. The scaling law has been characterized in the seminal work [46] for several types of input densities. Take the Gaussian source as example. In [46] it has been shown that, for the optimal $2K$-level uniform quantization that minimizes the MSE, 1) the *loading factor*[2] scales like $2\sqrt{\ln(2K)}$ (cf. the conventional "four-sigma" rule of thumb [44], [47], [48]), and 2) the granular distortion dominates and the overload distortion is asymptotically negligible. Further properties of the uniform quantization have been analyzed in [49]–[51]. For quantization at communication receivers, we need parallel results to characterize the optimal loading factor, which is essential for the design of AGC. We note that, although the importance of AGC design in the presence of output quantization has been recognized for a long time [6], [29], it was only investigated by numerical results in several works; see, e.g., [9], [10], [22], [52], [53].

Similar to the AQNM, there is also an additive noise model for source quantization [41], [44], [47], which approximates the quantization error as an independent white noise term added to the quantizer input (but does not include a scaling factor like that in the AQNM), though the "noise" is in fact a deterministic function of the input. The quantization error can be

[1]Interestingly, the loss can be fully recovered if we replace the hard-decision decoder (a.k.a. sign quantizer) by a carefully designed asymmetric one-bit quantizer (a.k.a. threshold quantizer) and employ asymmetric input [7].

[2]The loading factor of a quantizer is the one-sided width of its quantization range normalized by the standard deviation of the input [47].

white (uncorrelated between samples) when the source is i.i.d., and it is approximately uncorrelated with the input when the resolution is sufficiently high. So the model may give a useful approximation under certain circumstances (see, e.g., [54]).

C. Our Work

1) Problem: In this paper, we consider communications in the presence of output quantization, and focus on the effect of resolution reduction on the performance of a standard transceiver architecture designed without considering that effect. Rather than considering only low-resolution output quantization (1~3 bits) as in most existing studies, we consider the entire region of resolution, especially the transition from high resolution (typically 8~12 bits or more so that performance loss is negligible) to low resolution. Since in the transition only weak to moderate nonlinearity is introduced, it is natural to keep the transceiver architecture unchanged rather than rebuild it. The considered standard transceiver architecture includes independent and identically distributed (i.i.d.) complex Gaussian codebook at the transmitter, and a uniform output quantizer cascaded with a (weighted) nearest neighbor decoder at the receiver, where the loading factor of the quantizer can be adjusted by gain control. See Sec. II for details. We choose such an architecture in view of the following facts.

- When the impact of quantization is negligible, this architecture is capacity-achieving in several important channel models, such as the AWGN channel and the flat-fading channel with Gaussian noise and channel state information at the receiver [55]. It is also a very robust architecture in general noisy channels [56]. This architecture has been adopted in various performance evaluation problems, e.g., [35], [36], [56] for linear channels with fading and [30], [38], [57] for nonlinear channels.
- The nearest neighbor decoding (minimum Euclidean distance decoding) can be implemented efficiently and has been widely employed as a standard decoding rule in communication systems. The uniform quantizer is also a standard component of practical receivers as well as a common assumption in performance analysis [9]–[11], [14], [15], [25], [53], [58].
- The achievable rate of complex Gaussian input approximates that of regular high-order modulation schemes such as quadrature amplitude modulation (QAM). In the AWGN channel, the high-SNR gap between their achievable rates is 1.53 dB, which can be further reduced by constellation shaping [59].

2) Method: The achievable rate results in this paper are derived based on the GMI, which, as discussed in Sec. I-A, is a convenient performance measure for the problem of information transmission under transceiver nonlinearity. Moreover, as a performance measure under a *mismatched* decoder (since the nearest neighbor decoding rule becomes suboptimal in the presence of receiver quantization), the GMI possesses optimality in the sense that it is the maximum achievable rate of the i.i.d. random code ensemble, thereby indicating the performance of a "typical" codebook [32], [35]. Specifically, for Gaussian codebook and nearest neighbor decoding, the GMI has a simple expression which can be evaluated by the correlation between the channel input and output. Our asymptotic analyses also rely on methods and results in asymptotic (high-resolution) quantization theory. A notable tool originating from numerical analysis is the Euler-Maclaurin summation formula [60], which was initially introduced to source quantization theory in [49].

3) Summary of Contribution: We provide information-theoretic results for the transceiver architecture including i.i.d. complex Gaussian codebook and nearest neighbor decoding, in the presence of complex Gaussian noise and symmetric receiver quantization, especially uniform quantization. Our main contributions, including exact expressions, asymptotic formulas, and numerical results, are summarized as follows.

- In Sec. III, for the considered transceiver architecture, we show that the GMI under a given SNR can be expressed by

$$I_{\mathrm{GMI}} = C - \log(1 + \gamma \mathsf{SNR}), \tag{1}$$

  where C is the channel capacity when the resolution of quantization is unlimited, and the parameter γ, which does not depend on the SNR, is determined by thresholds and levels of the quantizer in an analytical form.
- In Sec. IV, for uniform quantization (with equispaced thresholds and mid-rise levels), we show that optimizing the loading factor by gain control before quantization (thereby minimizing γ in (1)) is increasingly important as the resolution decreases, thus imposing a critical challenge to the AGC. Interestingly, the problems of MSE minimization and achievable rate maximization are proven to be consistent, in the sense that there is a unique loading factor $L = L^*$ satisfying

$$L^* = \arg\max_{L} I_{\mathrm{GMI}}(L) = \arg\min_{L} \mathsf{mse}(L), \tag{2}$$

  where $I_{\mathrm{GMI}}(L)$ is the achievable rate as a function of the loading factor. This fact, combined with existing result, implies that the optimal loading factor L^* scales like $2\sqrt{b \ln 2}$ as the resolution b (in bits) increases. We further prove that the minimum of γ is exactly the minimum mean square error (MMSE) normalized by the variance of quantizer input (denoted by mmse), so that the *irreducible loss* in achievable rate due to uniform quantization is determined by

$$C - I_{\mathrm{GMI}}(L^*) = \log\left(1 + \mathsf{mmse} \cdot \mathsf{SNR}\right). \tag{3}$$

  For uniform quantization with the optimal loading factor, we establish a geometrical interpretation of how the MMSE and the additive Gaussian noise jointly determine the achievable rate.

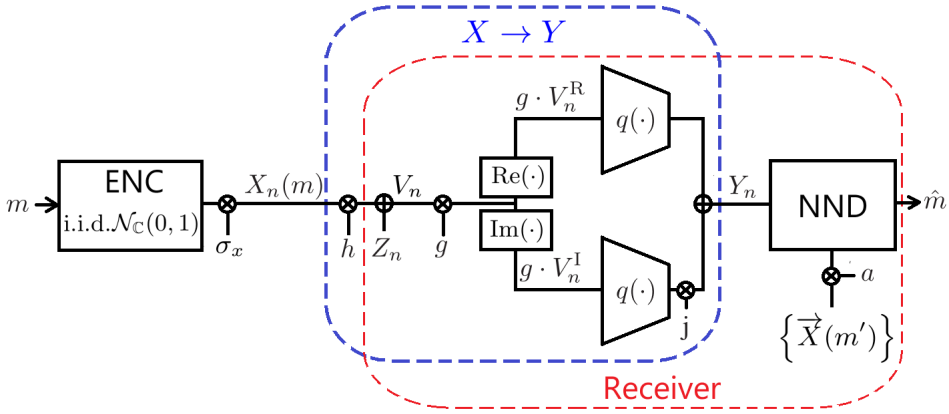


Fig. 1. Transceiver architecture.

- The uniform receiver quantization is further studied by asymptotic analysis. In Sec. V, the impact of biased gain control is characterized by asymptotic behaviors and approximations of $I_{\mathrm{GMI}}(L)$. Approximations of $I_{\mathrm{GMI}}(L)$ for overload region ($L < L^*$) and underload region ($L > L^*$) of the loading factor are proposed, respectively. Specifically, in the high-resolution regime, we characterize the *loading loss* in achievable rate, showing that i) the loading loss due to overload distortion decays exponentially as the loading factor increases, and ii) the loading loss due to granular distortion decays quadratically as the step size decreases. In Sec. VI, focusing on the optimal uniform quantization, we provide a new approximation of L^*, an approximation of $I_{\mathrm{GMI}}(L^*)$, and several per-bit rules for performance metrics such as saturation rates and irreducible rate loss;
- In Sec. VII, for general quantization rules, we illustrate that the consistency between achievable rate maximization and MSE minimization does not necessarily hold. We also discuss the possible gain of introducing more types of quantization for communication receiver.

Notation: We write $f(k) = o(g(k))$ to denote the asymptotic relationship $\lim_{k\to\infty} \frac{f(k)}{g(k)} = 0$. Specifically, we use $o_k(1)$ to denote a function $f(k)$ satisfying $\lim_{k\to\infty} f(k) = 0$, and may omit the subscript if there is no danger of confusion. We use $\phi(t)$ to denote the function $(2\pi)^{-1/2}\exp(-t^2/2)$, which is the probability density function (PDF) of the standard normal distribution, and use $Q(u)$ to denote the Q-function, i.e., $Q(u) := \int_u^\infty \phi(t)\mathrm{d}t$. We further define $\phi(\infty) = 0$ and $Q(\infty) = 0$. The complex conjugate of A is denoted by $\overline{A}$. The Euclidean norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|$. We use $X \perp Y$ to indicate the independence of two random variables X and Y.

II. Preliminaries

A. A Standard Transceiver Architecture

The transceiver architecture we consider is shown in Fig. 1. For a code rate R bits/channel use (c.u.), a message is selected uniformly randomly from the index set $\mathcal{M} = \{1, 2, ..., \lceil 2^{NR} \rceil\}$. If a message m is selected, then the encoder maps it to a codeword $[X_1(m), ..., X_N(m)]$ of block length N, which is generated according to a product complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(0, \sigma_x^2 \mathbf{I}_N)$. During transmission, each transmitted symbol is scaled by a channel gain $h \in \mathbb{C}$ which remains constant over the transmission duration of a codeword. The scaled symbols are corrupted by i.i.d. complex Gaussian noise at the receiver front-end before quantization. Then the channel output after quantization is given by

$$Y_n = q\left(g \cdot V_n^{\mathrm{R}}\right) + \mathrm{j} \cdot q\left(g \cdot V_n^{\mathrm{I}}\right), \tag{4}$$

where $n = 1, ..., N$, $q(\cdot)$ denotes the quantizer which introduces nonlinear distortion, $g \in \mathbb{R}^+$ is a gain-control factor, and

$$V_n^{\mathrm{R}} = \mathrm{Re}(hX_n(m) + Z_n), \tag{5}$$
$$V_n^{\mathrm{I}} = \mathrm{Im}(hX_n(m) + Z_n) \tag{6}$$

are real part and imaginary part of the received signal, respectively, where the noise $Z_n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent of X_n. In this model, we let the real and imaginary parts of the received signal be quantized by the same rule with the same gain-control factor.

Note (channel discretization): In practical systems, the described transceiver architecture also includes Nyquist-type pulse shaping at the transmitter and matched filtering combined with symbol-rate sampling at the receiver. It has a discrete-time memoryless model at symbol-level as (4), if the channel does not introduce memory. Note that oversampling (more accurately,

sampling faster than the symbol rate) may improve achievable rate in the presence of output quantization, especially in the case of one-bit quantization (see [61], [62] and [30], [63]–[65]). However, such performance improvement relies on several conditions, including significant nonlinearity (leading to frequency dispersion that can be utilized by oversampling), non-standard transceiver architecture (like new waveform design at the transmitter [64], [65]), and sufficiently high SNR. Receivers based on oversampling are beyond the scope of this paper since our focus is the effect of low-to-moderate nonlinearity on a standard transceiver architecture.

1) Nearest Neighbor Decoding Rule: The decoder selects a message according to the (scaled) nearest neighbor decoding rule [35], [56] as

$$\hat{m} = \arg \min_{m \in \mathcal{M}} \sum_{n=1}^{N} |Y_n - aX_n(m)|^2 . \tag{7}$$

That is, it selects a message corresponding to the codeword (scaled by a parameter a) with the minimum Euclidean distance to the received vector $[Y_1, ..., Y_N]$.[3]

2) Symmetric Quantizer and Uniform Quantizer: Let the quantizer in (4) be symmetric with $2K$ representation points (levels) $\{\pm y_1, ..., \pm y_K\}$ and normalized thresholds $\{0, \pm l_1, ..., \pm l_{K-1}\}$, where $y_k \geq 0$ and $l_k \geq 0$ for $k = 1, ..., K$. Then its resolution (bit-width) is $b = \log_2 2K$ bits, which typically satisfies $b \in \mathbb{Z}^+$. Let $V \in \mathbb{R}$ be the input to be quantized with standard deviation σ_v. For both quantizers in (4) we have $\sigma_v = \sqrt{(|h|^2\sigma_x^2 + \sigma^2)/2}$. Then for the input V, the output of the quantizer is

$$q(gV) = y_k \cdot \mathrm{sgn}(V), \text{ if } l_{k-1}\sigma_v \leq g|V| < l_k\sigma_v, \tag{8}$$

where the thresholds satisfy $l_0 = 0 < l_1 < ... < l_{K-1} < l_K = \infty$. Apparently, the quantizer output is a nonlinear function of its input, and the thus introduced nonlinearity degrades performance. In the presence of gain control, we may turn our attention to an equivalent quantization rule for a normalized input V/σ_v with adjustable thresholds $\{\ell_k = l_k/g,\ k = 1, ..., K-1\}$, where g can be adjusted to optimize the performance. A special case is the uniform quantizer, which has equispaced thresholds

$$\ell_k = k\ell,\ k = 1, ..., K-1, \tag{9}$$

and mid-rise levels

$$y_k = \left(k - \frac{1}{2}\right)\ell,\ k = 1, ..., K, \tag{10}$$

where ℓ is the step size. Thus, we define its quantization range or support as $[-K\ell, K\ell]$. Then the loading factor or support limit of the uniform quantizer is $L = K\ell$.

B. An Achievable Rate Formula from Generalized Mutual Information

Following the notation of [33], consider a memoryless channel $X \to Y$ with general alphabets $\mathcal{X}$ and $\mathcal{Y}$, input probability distribution $P_X(x)$, transition probability $P_{Y|X}(y|x)$, and decoding metric $d(X, Y)$. The decoder selects a message according to

$$\hat{m} = \arg \min_{m \in \mathcal{M}} \sum_{n=1}^{N} d(X_n(m), Y_n). \tag{11}$$

As a lower bound on mismatch capacity, the GMI can be given by its dual expression as [32], [33]

$$I_{\mathrm{GMI}} = \sup_{s \geq 0} \mathrm{E}\left[\log \frac{e^{-sd(X,Y)}}{\mathrm{E}\left[e^{-sd(X',Y)}|Y\right]}\right], \tag{12}$$

where $(X, Y, X') \sim P_X(x)P_{Y|X}(y|x)P_X(x')$. The GMI gives the maximum rate below which the probability of decoding error, averaged over the i.i.d. random codebook emsemble, converges to zero as the coding block length grows without bound. For communications with transceiver nonlinearity, if the input distribution $P_X(x)$ is complex Gaussian with variance σ_x^2 and the decoding rule (11) is specified by (7), then from (12) we obtain [30]

$$I_{\mathrm{GMI}} = \sup_{s \geq 0} \left(\log(1 + s|a|^2\sigma_x^2) - s\mathrm{E}\left[|Y - aX|^2\right] + \frac{s\mathrm{E}\left[|Y|^2\right]}{1 + s|a|^2\sigma_x^2} \right). \tag{13}$$

Maximizing the GMI by optimizing the scaling factor a yields the following result [30, Appendix C], which provides a general approach for achievable rate analysis in the presence of transceiver nonlinearity with known transition probability.

[3]Some more general forms of nearest neighbor decoding rule can be found in, e.g., [35], [36], [56]. A more recent work including a detailed review and some generalizations is [39].

Proposition 1 [30]: *For a memoryless SISO channel $X \to Y$ with transition probability $p_{Y|X}(y|x)$ and nearest neighbor decoding rule (7), where $X, Y \in \mathbb{C}$ and $\mathrm{Var}(X) = \sigma_x^2$, the maximum GMI under i.i.d. complex Gaussian codebook is given by*

$$I_{\text{GMI}} = \log \frac{1}{1-\Delta}, \tag{14}$$

where

$$\Delta = \frac{\left|\mathrm{E}\left[X\overline{Y}\right]\right|^2}{\sigma_x^2 \mathrm{E}\left[|Y|^2\right]}. \tag{15}$$

To achieve the maximum GMI given in (14), the scaling factor in (7) should be set as

$$a = \alpha := \frac{\mathrm{E}\left[\overline{X}Y\right]}{\sigma_x^2}. \tag{16}$$

Besides the proof in [30] based on direct evaluation and optimization of the dual expression of the GMI, here we provide a sketch of an alternative proof.

Proof Sketch: It has been noted in [35] and [66] that, for an additive *uncorrelated* noise channel $Y = S + U$ (i.e., the noise U satisfies $\mathrm{E}\left[S\overline{U}\right] = 0$, but is not necessarily independent of the input S), if $S \sim \mathcal{N}_{\mathbb{C}}(0, \mathrm{E}\left[|S|^2\right])$, then

$$I(X;Y) \geq \log\left(1 + \frac{\mathrm{E}\left[|S|^2\right]}{\mathrm{E}\left[|U|^2\right]}\right). \tag{17}$$

In [56], under Gaussian codebook and nearest neighbor decoding rule, the achievability of the RHS of (17) and a random coding converse for it are established by a geometric argument,[4] where the Gaussian codebook can either be an i.i.d. Gaussian codebook or an equienergy one. In the former case the rate (17) is the GMI. For the scalar channel $X \to Y$, $X \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_x^2)$, by a Bussgang-like decomposition, we can always write $Y = \alpha X + D$, where α is given in (16), and $D = Y - \alpha X$ satisfies $\mathrm{E}\left[X\overline{D}\right] = 0$. That is, we eliminate the correlation between the scaled input αX and the corresponding distortion D by a carefully chosen scaling factor. It is straightforward to show that (14) can be obtained by (17) when $S = \alpha X$ and $U = D$. ■

Let σ_y be the standard deviation of Y. We note that when $\mathrm{E}[Y] = 0$ (e.g., when the quantizer is symmetric, the channel output in (4) has zero mean) we have $\Delta = |\rho_{XY}|^2$, where

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} \tag{18}$$

is the *Pearson correlation coefficient* between X and Y. Also note that the equality in (17) holds if and only if $X \perp D$ and $D \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ (e.g., when there is no quantization in (4)), so that the channel $X \to Y$ reduces to the AWGN channel, the nearest neighbor decoding rule is optimal, and the GMI equals to the mutual information (also the channel capacity) $\log\left(1 + \frac{\sigma_x^2}{\sigma^2}\right)$.

III. Achievable Rate Under Receiver Quantization: Exact and Asymptotic Results

Based on Proposition 1, we establish the following result which provides an analytical expression for the achievable rate of the transceiver architecture considered in this paper.

Theorem 2: *For the channel (4) where $q(\cdot)$ is the symmetric quantizer described in Sec. II-A2, the achievable rate under i.i.d. complex Gaussian codebook and nearest neighbor decoding rule (7) is*

$$I_{\text{GMI}} = \log(1 + \mathsf{SNR}) - \log(1 + \gamma \mathsf{SNR}), \tag{19}$$

where $\mathsf{SNR} = |h|^2\sigma_x^2/\sigma^2$ is the SNR at the receiver front-end, and γ is a parameter determined by the quantizer as

$$\gamma = 1 - \frac{\mathcal{A}^2}{\mathcal{B}}, \tag{20}$$

in which

$$\mathcal{A} = \sqrt{2\pi} \sum_{k=1}^{K} y_k \left(\phi(\ell_{k-1}) - \phi(\ell_k)\right), \tag{21}$$

and

$$\mathcal{B} = \pi \sum_{k=1}^{K} y_k^2 \left(Q(\ell_{k-1}) - Q(\ell_k)\right). \tag{22}$$

[4]The proof in [56] was intended for independent noise. However, as noted in [35], it goes through verbatim for uncorrelated noise.

Proof: Applying Proposition 1, it is sufficient to show that

$$\Delta = \frac{\mathsf{SNR}}{1+\mathsf{SNR}} \frac{\mathcal{A}^2}{\mathcal{B}}, \tag{23}$$

which can be obtained by showing that

$$\mathrm{E}\left[X\overline{Y}\right] = 2\mathrm{E}\left[\mathrm{Re}(X)\cdot\mathrm{Re}(Y) + \mathrm{j}\cdot\mathrm{Im}(X)\cdot\mathrm{Re}(Y)\right] \tag{24a}$$

$$= \frac{2\mathrm{Re}(h)\sigma_x^2}{\sqrt{\pi(|h|^2\sigma_x^2+\sigma^2)}}\mathcal{A} - \frac{2\mathrm{j}\cdot\mathrm{Im}(h)\sigma_x^2}{\sqrt{\pi(|h|^2\sigma_x^2+\sigma^2)}}\mathcal{A} \tag{24b}$$

$$= \frac{2\overline{h}\sigma_x^2}{\sqrt{\pi(|h|^2\sigma_x^2+\sigma^2)}}\mathcal{A} \tag{24c}$$

and

$$\mathrm{E}\left[|Y|^2\right] = 2\mathrm{E}\left[\mathrm{Re}(Y)^2\right] = \frac{4}{\pi}\mathcal{B}. \tag{25}$$

The identities (24c) and (25) can be obtained by lengthy but straightforward evaluations of expectations (cf. [30, Appendix D]). ■

Remark: In [30, Sec. V], a parallel result of Theorem 2 for *real-valued* channel with symmetric output quantization was given by deriving its effective SNR, while Theorem 2 shows that the GMI in the complex-valued case has exactly the same expression except that the pre-log factor is doubled.

We note that, if the impact of quantization is omitted, then the model (4) reduces to a linear Gaussian channel $Y_n = hX_n(m) + Z_n$, and the achievable rate of the considered transceiver architecture is $\log(1+\mathsf{SNR})$, which is also the channel capacity C when the power of the input is σ_x^2. The expression of I_{GMI} in Theorem 2 shows explicitly that the rate loss due to quantization is

$$C - I_{\mathrm{GMI}} = \log\left(1+\gamma\mathsf{SNR}\right), \tag{26}$$

which has the same expression as C except for an SNR reduction of at least 4.4 dB (see (28) given later). For a given SNR, the reduction is determined by the parameter γ, which reflects the impact of nonlinearity and does not depend on the SNR.

Remark: Dithering, an intentional randomization technique in source quantization [41], may also be beneficial in receiver quantization; see, e.g., [31], which showed that *nonsubtractive* Gaussian dithering improves GMI under certain circumstances.[5] However, in our setting, nonsubtractive Gaussian dithering is always harmful, because it is equivalent to reducing the SNR, while the GMI is a monotonically increasing function of the SNR since its derivative (in nats) satisfies

$$\frac{\mathrm{d}I_{\mathrm{GMI}}}{\mathrm{d}\mathsf{SNR}} = \frac{1-\gamma}{(1+\mathsf{SNR})(1+\gamma\mathsf{SNR})} > 0. \tag{27}$$

From Theorem 2, we immediately obtain the following corollary, which shows that γ dominates the asymptotic behavior of performance, especially the low-SNR slope of achievable rate, and the high-SNR saturation rate.

Corollary 3: *The achievable rate I_{GMI} given in (19) has the following properties.*

- *The parameter γ satisfies*

$$0 < \gamma \le 1 - \frac{2}{\pi}, \tag{28}$$

where the equality holds when $q(\cdot)$ is a one-bit quantizer, corresponding to an achievable rate

$$I_{\mathrm{GMI}}^{\text{1-bit}} = \log\frac{1+\mathsf{SNR}}{1+\frac{\pi-2}{\pi}\mathsf{SNR}}, \tag{29}$$

which converges to $\log_2\frac{\pi}{\pi-2} = 1.4604$ bits/c.u. at high SNR (see also footnote 6). As the quantization becomes increasingly fine, we have $\gamma \to 0$ (from above), $\mathsf{SNR}_{\mathrm{e}} \to \mathsf{SNR}$ (from below), and

$$C - I_{\mathrm{GMI}} = \mathsf{SNR}\cdot\gamma - \frac{\mathsf{SNR}^2}{2}\gamma^2 + o(\gamma^2) \ \text{ nats/c.u.} \tag{30}$$

- *High- and low-SNR asymptotics: As $\mathsf{SNR} \to \infty$, we have $\mathsf{SNR}_{\mathrm{e}} \to \frac{1-\gamma}{\gamma}$ and*

$$I_{\mathrm{GMI}} = \log\frac{1}{\gamma} - \frac{1-\gamma}{\gamma}\frac{1}{\mathsf{SNR}} + o\left(\frac{1}{\mathsf{SNR}}\right). \tag{31}$$

[5] In nonsubtractive dithering, one adds a dither signal W_{d} independent of the quantizer input V before quantization, yielding an output $q(V+W_{\mathrm{d}})$. Subtractive dithering has an additional step that subtracts the dither signal after quantization and finally obtains $q(V+W_{\mathrm{d}}) - W_{\mathrm{d}}$. See [41] for more discussion.

So the saturation rate is given by

$$\bar{I}_{\text{GMI}} = \log \frac{1}{\gamma}. \tag{32}$$

As $\mathsf{SNR} \to 0$*, we have*

$$I_{\text{GMI}} = (1-\gamma)\mathsf{SNR} - \frac{1-\gamma^2}{2}\mathsf{SNR}^2 + o\left(\mathsf{SNR}^2\right) \quad \text{nats/c.u.} \tag{33}$$

An alternative expression of (19) is given in terms of effective SNR (or signal to noise-and-distortion ratio) as

$$I_{\text{GMI}} = \log\left(1 + \mathsf{SNR}_{\text{e}}\right), \tag{34}$$

where

$$\mathsf{SNR}_{\text{e}} = \frac{(1-\gamma)|h|^2\sigma_x^2}{\gamma|h|^2\sigma_x^2 + \sigma^2} = \frac{(1-\gamma)}{\gamma\mathsf{SNR}+1}\mathsf{SNR}. \tag{35}$$

From this expression, we see that the effect of quantization is to transfer a fraction γ of the power to the denominator of the effective SNR. At low SNR, the power loss can be evaluated by $\mathsf{SNR}/\mathsf{SNR}_{\text{e}}$, which asymptotically converges to $1/(1-\gamma) \le \pi/2$, where equality holds when one-bit quantization is used.[6]

From Theorem 2, it is simple to check that, for one-bit quantization, the gain control factor g has no impact on the achievable rate. In general, the achievable rate has the following property.

Proposition 4: The achievable rate I_{GMI} given in (19), as a function of the gain control factor g, satisfies

$$\lim_{g\to 0} I_{\text{GMI}}(g) = \lim_{g\to\infty} I_{\text{GMI}}(g) = I_{\text{GMI}}^{\text{1-bit}}. \tag{36}$$

Proof: See Appendix A. ■

This result can be interpreted intuitively as follows. Note that as $g \to \infty$ we have

$$\Pr\left(\min\left\{|V_n^{\text{R}}|, |V_n^{\text{I}}|\right\} \ge \ell_{K-1}\sigma_v\right) \to 1, \tag{37}$$

and as $g \to 0$ we have

$$\Pr\left(\max\left\{|V_n^{\text{R}}|, |V_n^{\text{I}}|\right\} < \ell_1\sigma_v\right) \to 1. \tag{38}$$

Thus, in both limits, the effective resolution of the quantizer reduces to one, so that the achievable rate converges to $I_{\text{GMI}}^{\text{1-bit}}$.

Remark: We note that $I_{\text{GMI}}^{\text{1-bit}}$ is not the infimum of $I_{\text{GMI}}(g)$ in general. An example will be given in Sec. IV-D, Fig. 17(f).

Intuitively, a smaller MSE is preferred in quantization. However, for communications the main performance metric is the achievable rate. Thus the relationship of MSE minimization and achievable rate maximization is of interest.

Due to symmetry, we may define the normalized MSE of the quantizer $q(\cdot)$ in the channel (4) in terms of the in-phase component of Y_n as

$$\mathsf{mse} := \mathrm{E}\left[\left(q(g\cdot V^{\text{R}}) - \frac{V^{\text{R}}}{\sigma_v}\right)^2\right], \tag{39}$$

where we omit the index of V^{R} and Y since we assume i.i.d. input, and we normalize the input since we use normalized levels of quantization. The following result shows that the normalized MSE can be expressed by $\mathcal{A}$ and $\mathcal{B}$, and is lower bounded by γ.

Proposition 5:

$$\mathsf{mse} = 1 - \frac{2}{\pi}\left(\sqrt{2\pi}\mathcal{A} - \mathcal{B}\right) \ge \gamma, \tag{40}$$

where equality holds if and only if

$$\frac{\mathcal{A}}{\mathcal{B}} = \sqrt{\frac{2}{\pi}}. \tag{41}$$

.

Proof: See Appendix B. ■

We define the normalized MMSE as

$$\mathsf{mmse} := \min_g \mathsf{mse}. \tag{42}$$

[6]The channel capacity under one-bit receiver quantization, given by $1 - H_2(Q(\mathsf{SNR}))$ [6], [8], has the same low-SNR asymptotic behavior. Note that the GMI is derived under Gaussian input and nearest neighbor decoder, and the capacity is achieved by antipodal signaling (the decoding rule is unrestricted). The difference becomes evident at high-SNR, where the capacity converges to 2 bits/c.u., while the GMI converges to 1.4604 bits/c.u.

From (40), the normalized MSE is minimized only if

$$\sqrt{2\pi}\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}g} = \frac{\mathrm{d}\mathcal{B}}{\mathrm{d}g}. \tag{43}$$

The different expressions of mse and γ show that, in general, the gain control factor that minimizes the MSE does not necessarily maximizes the achievable rate. However, for uniform quantization, the next section will show that the two optimization problems are consistent; i.e., there is a unique gain control factor (and correspondingly, a unique loading factor) that solves both problems simultaneously.

IV. Optimal Uniform Quantization at the Receiver

A. Achievable Rate Evaluation under Uniform Quantization

Proposition 6: *If $q(\cdot)$ is the uniform quantizer described in Sec. II-A2, then*

$$\mathcal{A} = \sqrt{2\pi}\sum_{k=0}^{K-1} \ell \cdot \phi(k\ell) - \frac{\ell}{2}, \tag{44}$$

and

$$\mathcal{B} = \pi \sum_{k=0}^{K-1} 2k\ell^2 Q(k\ell) + \frac{1}{8}\pi\ell^2. \tag{45}$$

Proof: See Appendix C. ∎

Specifically, for a uniform quantizer with a given resolution, (44) and (45) show that γ is determined solely by the step size ℓ or equivalently by the loading factor L, which can be optimized by adjusting the gain-control factor g in (4) according to the channel gain h.

Remark: The quantities $\mathcal{A}$ and $\mathcal{B}$ are important in our analysis. For uniform quantization, we note that the summation $\sum_{k=0}^{K-1} \ell \cdot \exp\frac{-k^2\ell^2}{2}$ in $\mathcal{A}$ is exactly the left Riemann sum of $\exp\frac{-t^2}{2}$ over $[0, K\ell]$ with a regular partition, and similarly, the summation $\pi\sum_{k=0}^{K-1} 2k\ell^2 Q(k\ell)$ in $\mathcal{B}$ is exactly the left Riemann sum of $2\pi t Q(t)$ over $[0, K\ell]$ with a regular partition. Therefore, for a fixed loading factor L we have the following high-resolution limits:

$$\lim_{K\to\infty} \mathcal{A} = \sqrt{2\pi}\lim_{K\to\infty}\left(\sum_{k=0}^{K-1}\frac{L}{K}\phi\left(\frac{kL}{K}\right) - \frac{L}{2K}\right) \tag{46a}$$

$$= \int_0^L \exp\frac{-t^2}{2}\mathrm{d}t \tag{46b}$$

$$= \sqrt{2\pi}\left(\frac{1}{2} - Q(L)\right), \tag{46c}$$

$$\lim_{K\to\infty} \mathcal{B} = 2\pi\lim_{K\to\infty}\left(\sum_{k=0}^{K-1}\frac{L}{K}\frac{kL}{K}Q\left(\frac{kL}{K}\right) + \frac{L^2}{16K^2}\right) \tag{47a}$$

$$= 2\pi\int_0^L tQ(t)\mathrm{d}t \tag{47b}$$

$$= \frac{\pi}{2} - 2\pi\int_L^\infty tQ(t)\mathrm{d}t. \tag{47c}$$

In Fig. 2 and Fig. 3 we show numerical evaluations of I_{GMI} in Theorem 2, which has been maximized over $L > 0$ for each resolution and is then denoted by I^*_{GMI}. The corresponding unique loading factor is denoted by L^*. Its uniqueness will be proved later. In Fig. 4 and Fig. 5, we show how the achievable rate I_{GMI} varies with the loading factor L (some more numerical results and details therein will be interpreted in subsequent sections). As the resolution increases, the increasing of L^* is clear. A "waterfall" near $L = 0$ can be observed in all figures, implying that an underestimate of the optimal loading factor always causes serious rate loss. To interpret this phenomenon, we write the normalized MSE of uniform quantization as the sum of overload distortion and granular distortion as

$$\mathsf{mse} = \mathsf{mse}_{\mathrm{o}} + \mathsf{mse}_{\mathrm{g}}, \tag{48}$$

where

$$\mathsf{mse}_{\mathrm{o}} := \int_L^\infty (t - L + \ell/2)^2\phi(t)\mathrm{d}t, \tag{49}$$

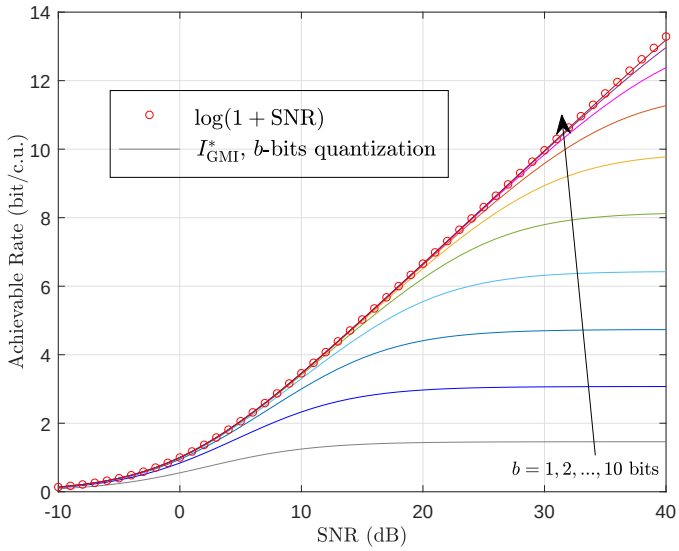


Fig. 2. Achievable rate with uniform output quantization and optimal loading factor.

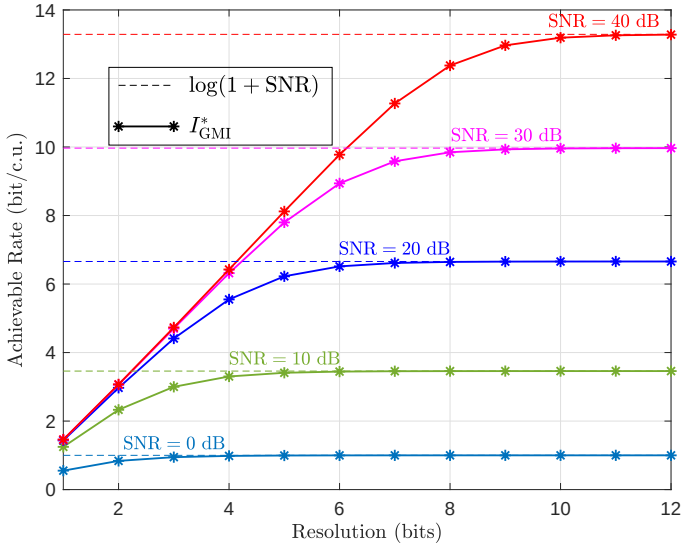


Fig. 3. Convergence of achievable rate to channel capacity.

and

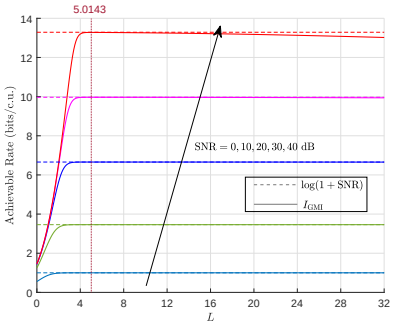
$$\mathsf{mse}_{\mathrm{g}} := \sum_{k=1}^{K} \int_{(k-1)\ell}^{k\ell} \left(t - \left(k - \frac{1}{2} \right) \ell \right)^2 \phi(t) \mathrm{d}t. \tag{50}$$

Clearly, decreasing L reduces the granular distortion but increases the overload distortion. When $L < L^*$ the overload distortion increases quickly, thereby causing the waterfall of achievable rate.

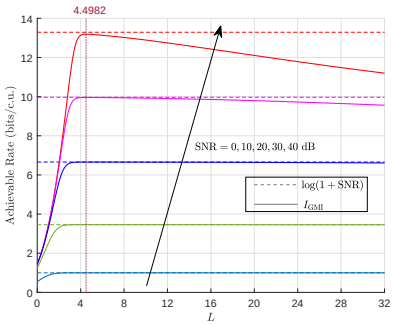
The numerical results in Fig. 4 and Fig. 5 reveal the increasing importance of gain control (realized by an AGC module in practical systems) as the resolution decreases: 1) Under high-resolution output quantization, we only require a rough estimate of the channel gain to guarantee that the loading factor to be *no less than* a predefined threshold, say 4 (from the four-sigma rule of thumb [44], [47]), so that we can stay away from the waterfall; 2) Under low-resolution output quantization, such a simple strategy may increase rate loss considerably, but perfect gain control needs accurate channel estimation which is also challenging in this case.

B. Consistency Between Rate Maximization and MSE Minimization and a GMI-MMSE Formula

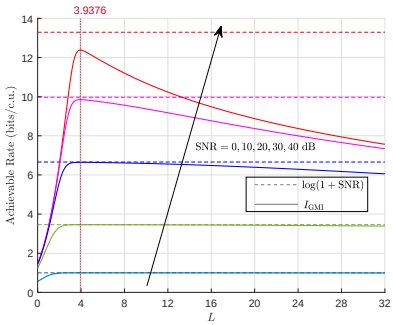
From source quantization theory [46], [50] we know that, for Gaussian input, the MSE of the uniform quantizer is a strictly convex function of the step size (or loading factor) and the minimum is located at the *unique* solution of $\frac{\mathrm{dMSE}}{\mathrm{d}\ell} = 0$, denoted



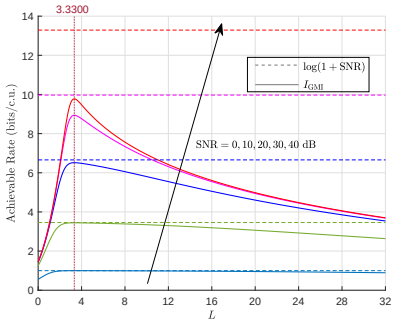
(a) $b = 12$ bits. Simple gain control strategy is always enough.



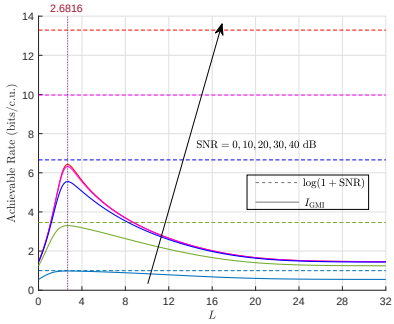
(b) $b = 10$ bits. Simple gain control strategy is usually enough.



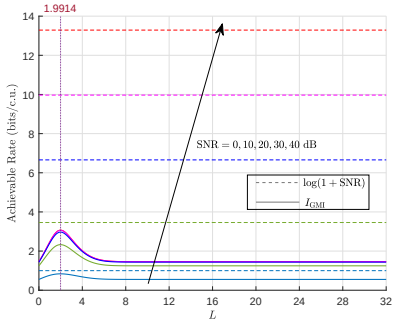
(c) $b = 8$ bits. Four-sigma rule of thumb is convenient.



(d) $b = 6$ bits. Four-sigma rule of thumb causes small loss at high SNR.

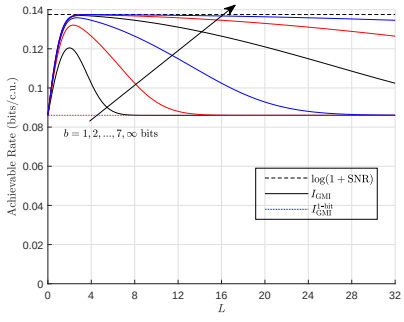


(e) $b = 4$ bits. Four-sigma rule of thumb causes considerable loss.

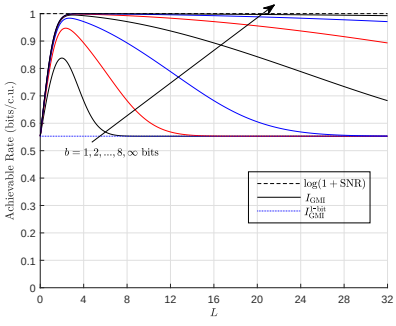


(f) $b = 2$ bits. Four-sigma rule of thumb causes significant loss.

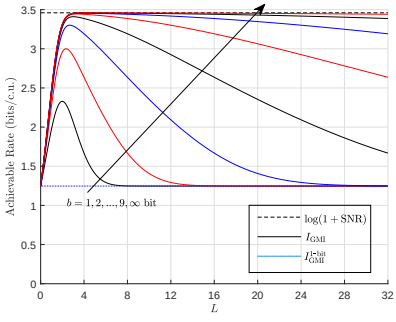
Fig. 4. Impact of loading factor on achievable rate: Fixed resolution, varying SNR. Vertical lines and corresponding values show the optimal loading factors.



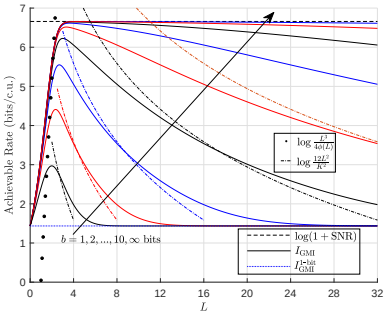
(a) SNR $= -10$dB.



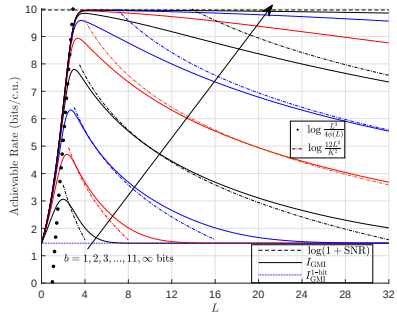
(b) SNR $= 0$dB.



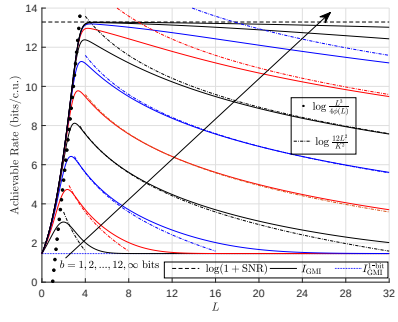
(c) SNR $= 10$dB $\approx$ SNR$_q(2)$.



(d) SNR $= 20$dB $\approx$ SNR$_q(4)$.



(e) SNR $= 30$dB $\approx$ SNR$_q(6)$.



(f) SNR $= 40$dB $\approx$ SNR$_q(8)$.

Fig. 5. Impact of loading factor on achievable rate: Fixed SNR, varying resolution.

by ℓ^*. In fact, that solution is also the unique step size that maximize the GMI; that is, for uniform quantization at the receiver there is a consistency between achievable rate maximization and MSE minimization. This consistency further leads to a GMI-MMSE formula which connects the maximum achievable rate and the MMSE of quantization in a simple and closed form. We first give the following lemma, and then use it to establish the aforementioned findings in Theorem 8.

Lemma 7: *For uniform quantization at the receiver, for $b > 1$ (i.e., except for one-bit quantization), we have*

$$\frac{\mathrm{d}I_{\mathrm{GMI}}}{\mathrm{d}\ell} = 0 \Leftrightarrow \frac{\mathrm{d}\mathsf{mse}}{\mathrm{d}\ell} = 0, \tag{51}$$

and both of them are equivalent to

$$\frac{\mathcal{A}}{\mathcal{B}} = \sqrt{\frac{2}{\pi}}. \tag{52}$$

Proof: See Appendix D. ∎

Theorem 8: *For the channel (4) where $q(\cdot)$ is the uniform quantizer described in Sec. II-A2 with resolution $b > 1$, the input employs i.i.d. complex Gaussian codebook, and the receiver employs nearest neighbor decoding rule (7), we have the following properties.*

- *The loading factor L^* that maximizes the GMI (19) is unique, and it is also the unique loading factor that minimizes the MSE (39), i.e.,*

$$\arg\max_{L} I_{\mathrm{GMI}}(L) = \arg\min_{L} \mathsf{mse}(L). \tag{53}$$

- *The minimum of γ as a function of L, namely $\gamma(L^*)$, is exactly the normalized MMSE (42) and satisfies*

$$0 < \mathsf{mmse} = \gamma(L^*) \le \gamma \le 1 - \frac{2}{\pi}, \tag{54}$$

and the maximum GMI can be written as

$$I^*_{\mathrm{GMI}} = \log(1 + \mathsf{SNR}) - \log\left(1 + \mathsf{mmse} \cdot \mathsf{SNR}\right). \tag{55}$$

Proof: First, in [46, Sec. V-A], it has been proved that the MSE is minimized if and only if the loading factor is set to be the unique solution of $\frac{\mathrm{d}\mathsf{mse}}{\mathrm{d}L} = 0$, denoted by L^*. The uniqueness of L^* is confirmed by showing that the second derivative of the MSE is strictly positive. Lemma 7 implies that the loading factor that satisfies $\frac{\mathrm{d}I_{\mathrm{GMI}}}{\mathrm{d}L} = \frac{\mathrm{d}I_{\mathrm{GMI}}}{K\mathrm{d}\ell} = 0$ is also unique. Noting that the achievable rate I_{GMI} is a continuous and differentiable function of L, from Proposition 4 we can infer that L^* also maximizes $I_{\mathrm{GMI}}(L)$ (it is not difficult to exclude the other possible case that L^* minimizes $I_{\mathrm{GMI}}(L)$). Then the first part of Theorem 8 is proved.

According to Lemma 7, if $L = L^*$ then (52) holds. Combining this fact with (40) and (20), we have

$$\gamma(L^*) = 1 - \frac{\pi}{2}\mathcal{B}(\ell^*) = \mathsf{mmse}, \tag{56}$$

thereby completing the proof of the second part of Theorem 8. ∎

From the proof of Theorem 8, we can infer that $I_{\mathrm{GMI}}(L) > I^{\text{1-bit}}_{\mathrm{GMI}}$ holds for $0 < L < \infty$. Combining this with Proposition 4 yields the following corollary.

Corollary 9: For uniform receiver quantization, the achievable rate given in (19), as a function of the loading factor L, satisfies

$$\inf_{L} I_{\mathrm{GMI}}(L) = \lim_{L\to 0} I_{\mathrm{GMI}}(L) = \lim_{L\to\infty} I_{\mathrm{GMI}}(L) = I^{\text{1-bit}}_{\mathrm{GMI}}, \tag{57}$$

where $I^{\text{1-bit}}_{\mathrm{GMI}}$ is given in (29).

In Theorem 8 we exclude the case $b = 1$. In fact, for symmetric one-bit quantization, we have $\mathcal{A} = \ell/2$, $\mathcal{B} = \pi\ell^2/8$, and $\gamma \equiv 1 - 2/\pi$ for $0 < \ell < \infty$, so that gain control is unnecessary. But when

$$\ell = \frac{4}{\sqrt{2\pi}} = 1.5958, \tag{58}$$

the normalized MSE achieves its minimum as $\mathsf{mmse}^{\text{1-bit}} = 1 - 2/\pi$, corresponding to the upper bound in (54).

Remark: From Theorem 2, the rate loss due to uniform quantization with a loading factor L is given by $\log(1+\gamma(L)\mathsf{SNR})$. For uniform quantization, Theorem 8 characterizes the minimum of rate loss when $L = L^*$. Thus we should distinguish two parts of the total rate loss as follows.

- Irreducible loss:[7] the unavoidable part for given resolution and SNR, given by

$$C - I^*_{\mathrm{GMI}} = \log(1 + \mathsf{mmse} \cdot \mathsf{SNR}). \tag{59}$$

[7]We note that the loss (59) is irreducible in the sense of GMI, and it is not necessarily irreducible in general since the GMI is only a lower bound on the mismatch capacity of the channel (4).

Numerical results in Fig. 2 and Fig. 3 consider only irreducible loss.

- Loading loss: the remaining part, given by

$$I_{\text{GMI}}^* - I_{\text{GMI}}(L) = \log \frac{1+\gamma(L)\mathsf{SNR}}{1+\mathsf{mmse}\cdot\mathsf{SNR}}, \tag{60}$$

which is due to suboptimal loading factor and can be reduced by improving the accuracy of gain control. Numerical results in Fig. 4 and Fig. 5 show the importance of reducing loading loss.

Theorem 8 enables us to utilize known results on uniform quantization for minimizing the MSE. The following result follows immediately from the equivalence (53) and existing results in [46], [67]. In particular, it is shown that the optimal loading factor L^* grows with the resolution b like $2\sqrt{b\ln 2}$.

Corollary 10: *In the channel (4) under i.i.d. complex Gaussian codebook and nearest neighbor decoding rule (7), the optimal step size ℓ^* that maximizes the GMI (19) satisfies*

$$\lim_{K\to\infty} K\ell^* = \infty, \ \lim_{K\to\infty} \ell^* = 0. \tag{61}$$

The optimal loading factor $L^* = K\ell^*$ grows monotonically with K and satisfies

$$\lim_{K\to\infty} \frac{L^*}{2\sqrt{\ln(2K)}} = 1. \tag{62}$$

The corresponding MMSE consisting of the granular distortion mmse_{g} and the overload distortion mmse_{o} satisfies

$$\lim_{K\to\infty} \frac{\mathsf{mmse}}{{\ell^*}^2/12} = \lim_{K\to\infty} \frac{\mathsf{mmse}_{\text{g}}}{{\ell^*}^2/12} = 1; \tag{63}$$

that is, the granular distortion dominates the MMSE as the resolution increases:

$$\lim_{K\to\infty} \frac{\mathsf{mmse}_{\text{o}}}{\mathsf{mmse}_{\text{g}}} = 0. \tag{64}$$

For (61) and (62), we also provide new proofs of them, respectively, by asymptotic results on the achievable rate; see Sec. VI-A.

Remark: In quantization theory, the "SNR" for a quantizer is often defined as the ratio between the variance of the quantizer input and the MSE. For the optimal uniform quantizer we denote

$$\mathsf{SNR}_q = \frac{1}{\mathsf{mmse}}. \tag{65}$$

Theorem 8 implies that the maximum saturation rate can be expressed as

$$\bar{I}_{\text{GMI}}^* = \log \frac{1}{\mathsf{mmse}} = \log \mathsf{SNR}_q, \tag{66}$$

and the corresponding effective SNR is given by

$$\sup_{\mathsf{SNR}>0} \mathsf{SNR}_{\text{e}} = \lim_{\mathsf{SNR}\to\infty} \mathsf{SNR}_{\text{e}} = \mathsf{SNR}_q - 1. \tag{67}$$

For finite SNR we have

$$\mathsf{SNR}_{\text{e}} = \frac{1-\mathsf{mmse}}{\mathsf{mmse}\cdot\mathsf{SNR}+1}\mathsf{SNR} = \frac{(1-\mathsf{mmse})\mathcal{E}_x}{\mathsf{mmse}\cdot\mathcal{E}_x+\sigma^2}. \tag{68}$$

In Table I,[8] we show numerical results of optimal uniform quantization under our transceiver architecture (an approximation of mmse given by $\ln(2K)/3K^2$, and an approximation of $\bar{I}_{\text{GMI}}$, denoted by $\hat{I}_{\text{GMI}}$, will be introduced in Sec. VI-B).

As the SNR decreases, numerical results in Fig. 5 show that the supremum and infimum of I_{GMI} have a decreasing ratio which converges to $\pi/2$, coinciding with the "2 dB loss" result for hard-decision decoding. Thus, a major part of the low-SNR capacity can always be utilized. At moderate-to-high SNR (Figs. (5(c)-5(f))), we may roughly separate different scenarios of receiver quantization as follows, which show that channel estimation and AGC design become more challenging as the resolution decreases.

- SNR-limited (high resolution) scenario: $b > 2\log_{10}\mathsf{SNR} + b_1$ bits and $\mathsf{SNR}_q \gg \mathsf{SNR}$, where b_1 can be 2~3. The irreducible loss (59) is negligible, and significant rate improvement can be obtained by increasing the SNR. In fact, we have

$$\frac{\mathrm{d}(C - I_{\text{GMI}}^*)}{\mathrm{d}\mathsf{SNR}} = \mathsf{mmse} + o(\mathsf{mmse}^2), \tag{69}$$

which suggests that the irreducible loss increases slowly with the SNR. A large overestimate of the optimal gain control factor does not cause significant loading loss, thereby allowing the simple gain control strategy described at the end of

[8]Some numerical results therein have been given in [46, Table II], in which an error occurred in actual SNR computation for Gaussian source, $b = 8$.

TABLE I
NUMERICAL RESULTS OF OPTIMAL PARAMETERS AND PERFORMANCE METRICS FOR UNIFORM QUANTIZATION

b	$2K$	L^*	ℓ^*	$\gamma(L^*)=\mathsf{mmse}$	$\ln(2K)/3K^2$	SNR_q (dB)	$\bar{I}^*_{\mathrm{GMI}}$(bits/c.u.)	$\hat{\bar{I}}_{\mathrm{GMI}}$(bits/c.u.)
1	2	1.5958	1.5958	0.3634	0.2310	4.40	1.4604	2.11
2	4	1.9914	0.9957	0.1188	0.1155	9.25	3.0728	3.11
3	8	2.3441	0.5860	3.7440×10^{-2}	4.3322×10^{-2}	14.27	4.7393	4.53
4	16	2.6816	0.3352	1.1543×10^{-2}	1.4441×10^{-2}	19.38	6.4369	6.11
5	32	3.0102	0.1881	3.4952×10^{-3}	4.5127×10^{-3}	24.57	8.1604	7.79
6	64	3.3300	0.1041	1.0400×10^{-3}	1.3538×10^{-3}	29.83	9.9091	9.53
7	128	3.6395	5.6868×10^{-2}	3.0433×10^{-4}	3.9486×10^{-4}	35.17	11.6821	11.31
8	256	3.9376	3.0762×10^{-2}	8.7686×10^{-5}	1.1282×10^{-4}	40.57	13.4773	13.11
9	512	4.2237	1.6499×10^{-2}	2.4919×10^{-5}	3.1730×10^{-5}	46.03	15.2924	14.94
10	1024	4.4982	8.7855×10^{-3}	6.9970×10^{-6}	8.8138×10^{-6}	51.55	17.1248	16.79
11	2048	4.7614	4.6498×10^{-3}	1.9444×10^{-6}	2.4238×10^{-6}	57.11	18.9722	18.65
12	4096	5.0143	2.4484×10^{-3}	5.3554×10^{-7}	6.6104×10^{-7}	62.71	20.8325	20.53

Sec. IV-A. In this scenario the impact of receiver quantization is not important. There is still room to reduce the resolution if accurate gain control is possible.

- Resolution-limited (low resolution) scenario: $b \le 2\log_{10}\mathsf{SNR}$ bits and $\mathsf{SNR}_q < \mathsf{SNR}$. The achievable rate is seriously limited due to large irreducible loss, while accurate gain control is required to avoid large loading loss. Significant rate improvement can be obtained by increasing the resolution.
- Moderate resolution scenario (the remaining cases): The resolution is enough to maintain a small irreducible loss, but an overestimate of the optimal gain control factor may cause considerable loading loss.

We note that the consistency between rate maximization and MSE minimization does not hold in general; see discussions on nonuniform quantization in Sec. VII.

C. Geometry of Optimal Uniform Quantization at Receiver

We have shown that, in the standard transceiver architecture shown in Fig. 1, if the gain control factor g is set appropriately so that the loading factor of the quantizer $q(\cdot)$ is equal to the optimal value L^*, then the achievable rate attains its maximum given by the GMI-MMSE formula (55). In fact, a geometrical interpretation for such optimal uniform quantization can be established. For simplification we first introduce an equivalent model shown in Fig. 6, where we let $\mathsf{X}=hX$, $\mathsf{Y}=\sigma_v Y = Y\cdot\sqrt{(|h|^2\sigma_x^2+\sigma^2)/2}$, $\mathsf{V}=V^{\mathrm{R}}+\mathrm{j}V^{\mathrm{I}}$, $\mathsf{Z}=Z$, and $\mathsf{W}=\mathsf{Y}-\mathsf{V}$ (we write $\mathsf{W}(\mathsf{V})$ in Fig. 6 to emphasize that W is a function of V), so that

$$\mathsf{Y}=\mathsf{X}+\mathsf{Z}+\mathsf{W}=\mathsf{V}+\mathsf{W}, \tag{70}$$

where the quantization error W satisfies

$$\mathrm{E}\left[|\mathsf{W}|^2\right]=\mathsf{MMSE}=\mathsf{mmse}\cdot\mathrm{E}\left[|\mathsf{V}|^2\right]=2\cdot\mathsf{mmse}\cdot\sigma_v^2. \tag{71}$$

Thus, the high-resolution limit of Y is V, and the high-SNR limit of V is X.

Now we are ready to illustrate the geometry of optimal uniform quantization at the receiver in N-dimensional Euclidean space, as shown in Fig. 7. We use boldface letters to denote codewords or signal vectors in the equivalent model (70), e.g., $\mathbf{X}=[\mathsf{X}_1,...,\mathsf{X}_N]$. Since i.i.d. codebook is considered, the vector $\mathbf{X}$ and other vectors in Fig. 7 are all i.i.d. random vectors. Thus, we have $\mathrm{E}\left[\|\mathbf{X}\|^2\right]=N\mathcal{E}_x$, and the empirical average power of the input codeword converges in probability to $\mathcal{E}_x$, i.e.,

$$\lim_{N\to\infty}\left(\left|\frac{1}{N}\|\mathbf{X}\|^2-\mathcal{E}_x\right|>\epsilon\right)=0,\ \forall\epsilon>0. \tag{72}$$

We thus briefly say that the length of $\mathbf{X}$ (in asymptotic sense) is $\sqrt{\mathrm{E}\left[\|\mathbf{X}\|^2\right]}=\sqrt{N\mathcal{E}_x}$. Similarly, the length of $\mathbf{Z}$ is $\sqrt{N\sigma^2}$. The geometry in Fig. 7 includes two Pythagorean relations as follows.

- For the additive noise channel $\mathsf{V}=\mathsf{X}+\mathsf{Z}$, we have $\mathrm{E}\left[\mathsf{X}\overline{\mathsf{Z}}\right]=0$, implying

$$\mathrm{E}\left[\|\mathbf{V}\|^2\right]=\mathrm{E}\left[\|\mathbf{X}\|^2\right]+\mathrm{E}\left[\|\mathbf{Z}\|^2\right]. \tag{73}$$

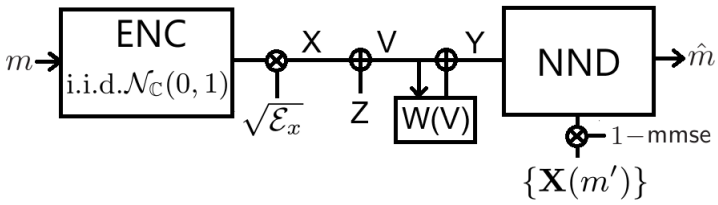


Fig. 6. A simplified equivalent model of the transceiver architecture when $L = L^*$.

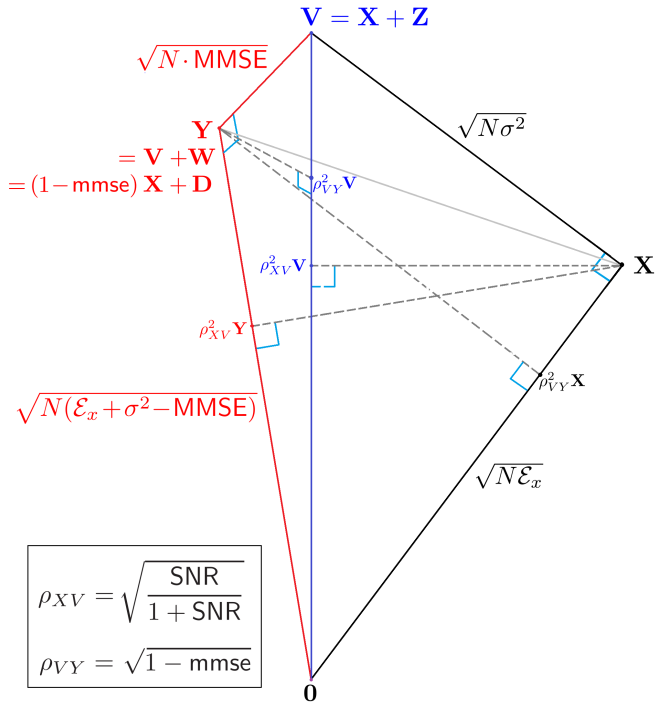


Fig. 7. Geometry of optimal uniform quantization at the receiver.

- For the quantization channel $\mathsf{Y} = \mathsf{V} + \mathsf{W}$, it has been shown in [67] that the MMSE uniform quantization satisfies $\mathrm{E}\left[\mathsf{Y}\overline{\mathsf{W}}\right] = 0$; i.e., the error W is uncorrelated with the *output* of the quantizer, yielding

$$\mathrm{E}\left[\|\mathbf{V}\|^2\right] = \mathrm{E}\left[\|\mathbf{Y}\|^2\right] + \mathrm{E}\left[\|\mathbf{W}\|^2\right]. \tag{74}$$

Thus, the lengths of $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{Y}$ are $\sqrt{N(\mathcal{E}_x+\sigma^2)}$, $\sqrt{N \cdot \mathsf{MMSE}}$, and $\sqrt{N(\mathcal{E}_x+\sigma^2-\mathsf{MMSE})}$, respectively. We then have two right triangles in Fig. 7: the *quantization triangle* $\mathbf{0VY}$, which determines the saturation rate $\bar{I}^*_{\mathrm{GMI}}$ in (66), and the *noise triangle* $\mathbf{0XV}$, which determines the channel capacity C, i.e., the limit of achievable rate under fine quantization. These two triangles jointly determine the triangle $\mathbf{0XY}$, and consequently determine the achievable rate I^*_{GMI}.

We note that $\mathbf{YVX}$ and $\mathbf{0XY}$ are not right triangles. In fact, for finite SNR and finite resolution, $\mathrm{E}\left[|\mathsf{Y}-\mathsf{X}|^2\right]$ is strictly smaller than $\mathrm{E}\left[|\mathsf{Z}|^2\right] + \mathrm{E}\left[|\mathsf{W}|^2\right]$, because the error W is correlated with the noise Z. Treating the error as independent noise always reduces the achievable rate.

In Fig. 7 there are also some perpendicular lines (dashed) which indicate some projections. For example, the projection of $\mathbf{X}$ onto $\mathbf{V}$ is the linear minimum MSE (LMMSE) estimator of $\mathbf{X}$ from $\mathbf{V}$ as

$$\hat{\mathbf{X}}_{\mathrm{LMMSE}}(\mathbf{V}) = \frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{V}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]}\mathbf{V}, \tag{75}$$

i.e., a scaling of $\mathbf{V}$ with a scalar (called Wiener coefficient), which is equal to $\rho^2_{XV} = \mathsf{SNR}/(1+\mathsf{SNR})$ since $\mathrm{E}\left[\mathsf{X}\overline{\mathsf{V}}\right] = \mathrm{E}\left[|\mathsf{X}|^2\right] = \mathsf{SNR} \cdot \sigma^2$. Similarly, the Wiener coefficient of the LMMSE estimator of $\mathbf{Y}$ from $\mathbf{V}$ is $\rho^2_{VY} = 1 - \mathsf{mmse}$. The following result provides some equivalent expressions of the condition (52) for GMI maximization, which also interprets the Wiener coefficients of the other two projections shown in Fig. 7.

Proposition 11: *The following four conditions are equivalent to each other:*

$$\frac{\mathcal{A}}{\mathcal{B}} = \sqrt{\frac{2}{\pi}}, \tag{76}$$

$$\frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{Y}}\right]}{\mathrm{E}\left[|\mathsf{Y}|^2\right]} = \frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{V}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]}, \tag{77}$$

$$\frac{\mathrm{E}\left[\mathsf{Y}\overline{\mathsf{X}}\right]}{\mathrm{E}\left[|\mathsf{X}|^2\right]} = \frac{\mathrm{E}\left[\mathsf{V}\overline{\mathsf{Y}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]}, \tag{78}$$

and

$$\rho_{XY} = \rho_{XV} \cdot \rho_{VY}, \tag{79}$$

where

$$\frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{V}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]} = \frac{\mathsf{SNR}}{1+\mathsf{SNR}} = \rho_{XV}^2, \tag{80}$$

and

$$\frac{\mathrm{E}\left[\mathsf{V}\overline{\mathsf{Y}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]} = 1 - \mathsf{mmse} = \rho_{VY}^2, \tag{81}$$

respectively.

Proof: See Appendix E. ∎

Remark: Proposition 11 shows that the optimal uniform quantization rule should let

$$\hat{\mathbf{X}}_{\mathrm{LMMSE}}(\mathbf{Y}) = \frac{\mathsf{SNR}}{1+\mathsf{SNR}}\mathbf{Y} \tag{82}$$

hold; i.e., it should let the Wiener coefficient be equal to the one in (80). Similarly, the optimal uniform quantization rule should let

$$\hat{\mathbf{Y}}_{\mathrm{LMMSE}}(\mathbf{X}) = (1-\mathsf{mmse})\mathbf{X} \tag{83}$$

hold, where the Wiener coefficient is equal to the one in (81). Moreover, from (83) we obtain an additive uncorrelated noise model as

$$\mathsf{Y} = (1-\mathsf{mmse})\mathsf{X} + \mathsf{D}, \tag{84}$$

where D satisfies $\mathrm{E}\left[\mathsf{X}\overline{\mathsf{D}}\right] = 0$ and

$$\mathrm{E}\left[|\mathsf{D}|^2\right] = (1-\mathsf{mmse})(\mathsf{mmse}\cdot\mathcal{E}_x + \sigma^2). \tag{85}$$

The channel (84) is a scaling of the channel $Y = \alpha X + D$ in the proof of Proposition 1. We can apply Proposition 1 directly to (84) and obtain exactly the same GMI expression in Theorem 8. Correspondingly, the scaling factor for the nearest neighbor decoder in Fig. 7 should be set as $1-\mathsf{mmse}$.

We next discuss how the triangles in Fig. 7 determine the performance via three angles as follows.

1) $\theta_{XV} \in (0, \pi/2)$: it is determined by the SNR and satisfies

$$\mathsf{SNR} = \cot^2\theta_{XV} \tag{86}$$

and

$$\rho_{XV} = \cos\theta_{XV} = \sqrt{\frac{\mathsf{SNR}}{1+\mathsf{SNR}}}. \tag{87}$$

The high-resolution limit of achievable rate (i.e., the capacity) can be expressed as

$$C = \log(1+\mathsf{SNR}) = \log\left(\csc^2\theta_{XV}\right). \tag{88}$$

2) $\theta_{VY} \in \left(0, \arccos\sqrt{\frac{2}{\pi}}\right)$: it is determined by the resolution, or equivalently by mmse, and satisfies

$$\mathsf{mmse} = \sin^2\theta_{VY} \tag{89}$$

and

$$\rho_{VY} = \cos\theta_{VY} = \sqrt{1-\mathsf{mmse}}. \tag{90}$$

The high-SNR limit of achievable rate (i.e., the saturation rate) can be expressed as

$$\bar{I}^*_{\text{GMI}} = \log \frac{1}{\mathsf{mmse}} = \log \left(\csc^2 \theta_{VY}\right). \tag{91}$$

3) $\theta_{XY} \in (\max(\theta_{XV}, \theta_{VY}), \pi/2)$: it is determined by θ_{XV} and θ_{VY} as (see Proposition 11)

$$\cos\theta_{XY} = \cos\theta_{XV} \cos\theta_{VY}, \tag{92}$$

so that

$$\rho_{XY} = \cos\theta_{XY} = \sqrt{\frac{\mathsf{SNR}}{1+\mathsf{SNR}}}\sqrt{1-\mathsf{mmse}}. \tag{93}$$

The maximum achievable rate can be expressed as

$$I^*_{\text{GMI}} = \log \left(\csc^2 \theta_{XY}\right), \tag{94}$$

corresponding to an effective SNR as $\mathsf{SNR}^*_{\text{e}} = \cot^2 \theta_{XY}$.

The preceding relations show that, under varying resolution and fixed SNR (θ_{VY} varies while the triangle $\mathbf{0XV}$ is fixed), according to the relationship (92), the optimal loading factor lets $\cos\theta_{XY}$ be directly proportional to $\cos\theta_{VY}$. For increasingly fine quantization, the limit of the triangle $\mathbf{0XY}$ is the triangle $\mathbf{0XV}$, the limit of $\rho^2_{XY}\mathbf{Y}$ is $\rho^2_{XY}\mathbf{V}$, and the limit of achievable rate is the channel capacity C. Similarly, under varying SNR and fixed resolution (θ_{XV} varies while the triangle $\mathbf{0VY}$ is fixed), the optimal loading factor lets $\cos\theta_{XY}$ be directly proportional to $\cos\theta_{XV}$. For increasing SNR, the limit of the triangle $\mathbf{0XY}$ is the triangle $\mathbf{0VY}$, the limit of $\rho^2_{VY}\mathbf{X}$ is $\rho^2_{VY}\mathbf{V}$, and the achievable rate tends to the saturation rate. The low-SNR slope of the achievable rate is $\cos^2\theta_{VY} = 1 - \mathsf{mmse}$.

Finally we note that a suboptimal loading factor always breaks the geometry in Fig. 7. For example, if we fix the loading factor and let b increase without bound, then the granular distortion vanishes but the overload distortion keeps unchanged, so that $\mathbf{Y}$ cannot converge to $\mathbf{V}$. For a given resolution, as the SNR increases, a suboptimal loading factor breaks the Pythagorean relation (74). In this case the length of $\mathbf{Y}$ can be larger or smaller than $\sqrt{N(\mathcal{E}_x + \sigma^2 - \mathsf{MMSE})}$, but the distance $\mathrm{E}\left[\|\mathbf{Y}-\mathbf{V}\|^2\right]$ and the angle θ_{VY} always become larger, thereby reducing the achievable rate.

V. Asymptotic Analysis: Characterization of $I_{\text{GMI}}(L)$

From previous numerical results, we have noted that adjusting the loading factor by gain control is essential in approaching the maximum achievable rate. To understand the impact of biased gain control, this section investigates the achievable rate $I_{\text{GMI}}(L)$ as a function of the loading factor L when the resolution and the SNR are given. Our basic approach is asymptotic analysis, by which we establish analytical results and interpret some phenomena observed in numerical results. In source quantization, asymptotic analysis yields high-resolution quantization theory, which provides fairly accurate results for resolutions equal to or greater than 3 bits [41]. Our results for receiver quantization exhibit similar accuracy.

In summary, the behavior of $I_{\text{GMI}}(L)$ can be characterized by two regions as follows.

1) Overload Region ($L < L^*$): In this region overload distortion dominates the performance, but the impact of granular distortion is also important when L is close to L^*. The achievable rate behaves like a waterfall except except for the case where L is very close to L^*. For high-resolution quantization at high SNR, we have $I_{\text{GMI}} \approx \log(L^3/4\phi(L))$, which increases approximately linearly with L. On the other hand, as $L \to 0$, the effective resolution of the quantizer reduces to one bit and $I_{\text{GMI}}(L)$ converges to $I^{\text{1-bit}}_{\text{GMI}}$ given in (29).

2) Underload Region ($L > L^*$): In this region granular distortion dominates the performance. As L increases, the rate loss increases like $\log(1 + L^2K^{-2}\mathsf{SNR}/12)$ over a wide range. When $L \gg L^*$, the effective resolution reduces to one bit and $I_{\text{GMI}}(L)$ converges to $I^{\text{1-bit}}_{\text{GMI}}$ again.

Specifically, when the resolution is sufficiently high so that the irreducible loss is negligible, the loading loss (60) can be expressed as $\log(1+\gamma(L)\mathsf{SNR})$, which is characterized as follows.

- When $L \to L^*$ from below (in the overload region), the loading loss decays like $O(4\phi(L)/L^3)\mathsf{SNR}$; i.e., it decays exponentially with L; see Theorem 12.
- When $L \to L^*$ from above (in the underload region), the loading loss decays like $L^2K^{-2}\mathsf{SNR}/12$; i.e., it decays quadratically with L; see Theorem 14.

We first establish the preceding high-resolution asymptotic results for loading loss in Sec. V-A, and then utilize them to characterize the behavior of $I_{\text{GMI}}(L)$ in Sec. V-B.

A. Decay of Rate Loss under High-Resolution Receiver Quantization

We first consider the impact of bias in gain control when the irreducible loss is negligible. In other words, we consider how $I_{\text{GMI}}(L)$ converges to $I_{\text{GMI}}^* \approx C$ as $L \to L^*$. Our method is to characterize asymptotic behaviors of rate loss under finite loading factor and finite step size, respectively, in the high-resolution limit. Then the accuracy of the obtained asymptotic formulas is evaluated numerically for finite-resolution receiver quantization.

The following result shows that, the rate loss due to only overload distortion decays exponentially as the loading factor increases. Asymptotically, the loading loss is directly proportional to the overload distortion, and is also directly proportional to the SNR.

Theorem 12: *In the channel (4) under i.i.d. complex Gaussian codebook and nearest neighbor decoding rule (7), the rate loss due to uniform quantization with a loading factor L satisfies*

$$C - I_{\text{GMI}} = \left(\underline{\mathsf{mse}}_{\text{o}} + (1 + o_L(1)) \frac{4\phi^2(L)}{L^2} \right) \mathsf{SNR} \text{ nats/c.u.} \tag{95a}$$

$$= \left((1 + o_L(1)) \frac{4\phi(L)}{L^3} \right) \mathsf{SNR} \text{ nats/c.u.} \tag{95b}$$

in the high-resolution limit, where

$$\underline{\mathsf{mse}}_{\text{o}} := 2 \int_L^\infty (t - L)^2 \phi(t) \mathrm{d}t \tag{96}$$

is the infimum of the normalized overload distortion mse_{o} *given in (49).*

Proof: Combining (46c) and (47c), we obtain the high-resolution limit of γ as a function of L as

$$\bar{\gamma}(L) := \lim_{K \to \infty} \gamma \tag{97a}$$

$$= \frac{\frac{1}{4} - \int_L^\infty tQ(t)\mathrm{d}t - \left(\frac{1}{2} - Q(L)\right)^2}{\frac{1}{4} - \int_L^\infty tQ(t)\mathrm{d}t} \tag{97b}$$

$$= \frac{4\int_L^\infty \left(\phi(t) - tQ(t)\right) \mathrm{d}t - 4Q^2(L)}{1 - 4\int_L^\infty tQ(t)\mathrm{d}t}. \tag{97c}$$

In the high-resolution limit, granular distortion vanishes and the normalized MSE includes only the overload distortion, namely $\underline{\mathsf{mse}}_{\text{o}}$. From Proposition 5, (46c), and (47c), we obtain another expression of the overload distortion as

$$\underline{\mathsf{mse}}_{\text{o}} = 4 \int_L^\infty \left(\phi(t) - tQ(t)\right) \mathrm{d}t. \tag{98}$$

From [46, Lemma 7 and Eqn. A9], we can infer that[9]

$$\underline{\mathsf{mse}}_{\text{o}} = (1 + o_L(1)) \frac{4\phi(L)}{L^3}. \tag{99}$$

We thus obtain

$$\bar{\gamma}(L) = \frac{\underline{\mathsf{mse}}_{\text{o}} - 4Q^2(L)}{1 - 4\int_L^\infty tQ(t)\mathrm{d}t} \tag{100a}$$

$$= \underline{\mathsf{mse}}_{\text{o}} + (1 + o_L(1)) \frac{4\phi^2(L)}{L^2} \tag{100b}$$

$$= (1 + o_L(1)) \frac{4\phi(L)}{L^3}, \tag{100c}$$

where we utilize the fact

$$Q(t) = (1 + o_t(1)) \frac{\phi(t)}{t}, \tag{101}$$

which follows from bounds for the Q function as [68]

$$\frac{\phi(t)}{t} > Q(t) > \frac{t}{1 + t^2} \phi(t). \tag{102}$$

The proof is completed by combining (100) and (30). ■

[9]The derivation in [46] begins from the original form (96). However, we can also begin from (98) and confirm (99) directly by bounds of Q-function [68], e.g., $\frac{1}{t+1/t}\phi(t) < Q(t) \leq \frac{1}{3t/4+\sqrt{t^2+8}/4}\phi(t)$.

Remark: By the lower bound in (102), one can show that $\phi(t) - tQ(t) < \phi(t)/t^2$. Thus the integral in (98) can be upper bounded by

$$\int_L^\infty \frac{\phi(t)}{t^2}\mathrm{d}t = L^{-1}\phi(L) - Q(L). \tag{103}$$

Using the lower bound in (102) again, we obtain

$$\underline{\mathsf{mse}}_{\mathrm{o}} < \frac{4\phi(L)}{L^3}, \tag{104}$$

which implies that, as L increases, $\underline{\mathsf{mse}}_{\mathrm{o}}$ converges to $4\phi(L)L^{-3}$ from below (cf. (99)).

On the other hand, the rate loss due to only granular distortion decays quadratically as the step size vanishes. To prove this we need the following lemma which is one of the various forms of the Euler-Maclaurin summation formula [60].

Lemma 13: *For a real-valued continuously differentiable function $f(t)$ defined on $[a, b]$, we have*

$$\int_a^b f(t)\mathrm{d}t = \ell\left(\frac{f(a)}{2} + \sum_{k=1}^{K-1} f(a+k\ell) + \frac{f(b)}{2}\right) - \frac{\ell^2}{12}\left(f'(b) - f'(a)\right) + o(\ell^2), \tag{105}$$

where $\ell = \frac{b-a}{K}$.

This lemma characterizes the error of numerical integration using the composite trapezoidal rule with an evenly spaced (uniform) grid. Based on Lemma 13, the following result can be obtained.

Theorem 14: *In the channel (4) under i.i.d. complex Gaussian codebook and nearest neighbor decoding rule (7), the rate loss due to uniform quantization with a step size ℓ satisfies*

$$C - I_{\mathrm{GMI}} = \left(\overline{\mathsf{mse}}_{\mathrm{g}} + o(\ell^2)\right)\mathsf{SNR} \ \text{ nats/c.u.} \tag{106a}$$

$$= \left(\frac{\ell^2}{12} + o(\ell^2)\right)\mathsf{SNR} \ \text{ nats/c.u.} \tag{106b}$$

in the high-resolution limit, where

$$\overline{\mathsf{mse}}_{\mathrm{g}} := \lim_{K\to\infty} \mathsf{mse}_{\mathrm{g}} \tag{107}$$

is the supremum of the normalized granular distortion $\mathsf{mse}_{\mathrm{g}}$ given in (50).

Proof: Applying the Euler-Maclaurin summation formula in Lemma 13, for $\mathcal{A}$ we have

$$\int_0^{K\ell} \exp\frac{-t^2}{2}\mathrm{d}t = \ell\left(\frac{1}{2} + \sum_{k=1}^{K-1}\exp\frac{-k^2\ell^2}{2} + \frac{1}{2}\exp\frac{-K^2\ell^2}{2}\right) + \frac{\ell^2}{12}K\ell\exp\frac{-K^2\ell^2}{2} + o(\ell^2) \tag{108a}$$

$$= \sum_{k=0}^{K-1}\ell\cdot\exp\frac{-k^2\ell^2}{2} - \frac{\ell}{2} + \frac{\ell}{2}\exp\frac{-K^2\ell^2}{2} + \frac{K\ell^3}{12}\exp\frac{-K^2\ell^2}{2} + o(\ell^2) \tag{108b}$$

$$= \mathcal{A} + o(\ell^2), \tag{108c}$$

and for $\mathcal{B}$ we have

$$\int_0^L 2tQ(t)\mathrm{d}t = \ell\left(\sum_{k=1}^{K-1} 2k\ell Q(k\ell) + K\ell Q(K\ell)\right) - \frac{\ell^2}{6}\left(Q(K\ell) - \frac{K\ell}{\sqrt{2\pi}}\exp\frac{-K^2\ell^2}{2} - \frac{1}{2}\right) + o(\ell^2) \tag{109a}$$

$$= \sum_{k=0}^{K-1} 2k\ell^2 Q(k\ell) + \frac{\ell^2}{8} - \frac{\ell^2}{24} + o(\ell^2) \tag{109b}$$

$$= \frac{\mathcal{B}}{\pi} - \frac{\ell^2}{24} + o(\ell^2). \tag{109c}$$

As $K\to\infty$, we have $L = K\ell = \infty$. Then in the high-resolution limit we have $\mathcal{A} = \sqrt{\pi/2} + o(\ell^2)$ and $\mathcal{B}/\pi = 1/2 + \ell^2/24 + o(\ell^2)$, which yield

$$\mathsf{mse} = \overline{\mathsf{mse}}_g = \frac{\ell^2}{12} + o(\ell^2) \tag{110}$$

and

$$\bar{\gamma} = 1 - \frac{1}{\pi}\frac{\frac{\pi}{2} + o(\ell^2)}{\frac{1}{2} + \frac{\ell^2}{24} + o(\ell^2)} = \frac{\ell^2}{12} + o(\ell^2). \tag{111}$$

The proof of (106a) is completed by combining (110) and (30), while the proof of (106b) is completed by combining (111) and (30). ∎

In Theorem 12 and Theorem 14 we assume output quantization with *unlimited* resolution. We now check whether they provide useful approximations of rate loss under finite resolutions. In Fig. 8, we show the rate loss $C - I_{\text{GMI}}$ due to 12 bits output quantization. According to Theorem 12, a small irreducible loss in overload region can be approximated by

$$C - I_{\text{GMI}}(L) \approx 4\phi(L)L^{-3}\mathsf{SNR} \text{ nats/c.u.}, \tag{112}$$

and according to Theorem 14, a small irreducible loss in underload region can be approximated by

$$C - I_{\text{GMI}}(L) \approx L^2\mathsf{SNR}/12K^2 \text{ nats/c.u.} \tag{113}$$

Clearly, these approximations successfully capture the decay of rate loss when it is dominated by the overload distortion. When the loading factor L exceeds 4, the increasing granular distortion kicks in and dominates the performance quickly. In this case the achievable rate is well approximated by $L^2\mathsf{SNR}/12K^2$. When the resolution decreases, e.g., in Fig. 9 where the resolution is 6 bits, the approximation (112) becomes less useful, especially at high SNR. But the approximation (113) is still satisfactory. For higher accuracy, by considering both the overload distortion and the granular distortion, we propose an approximation given by

$$C - I_{\text{GMI}}(L) \approx \log\left(1 + \widehat{\mathsf{mse}} \cdot \mathsf{SNR}\right), \tag{114}$$

where

$$\widehat{\mathsf{mse}} = \frac{4\phi(L)}{L^3} + \frac{4\phi^2(L)}{L^2} + \frac{L^2}{12K^2}, \tag{115}$$

in which the first two terms of the RHS come from (95a), and the last term comes from (106). As shown in Fig. 9, the proposed formula well approximates the transition from overload region to underload region.

B. Properties and Approximations of $I_{GMI}(L)$

We first discuss some general properties of $I_{\text{GMI}}(L)$. The derivative of the achievable rate

$$\frac{\mathrm{d}I_{\text{GMI}}}{\mathrm{d}L} = \frac{-\mathsf{SNR}}{\gamma\mathsf{SNR}+1}\frac{\mathrm{d}\gamma}{\mathrm{d}L} \tag{116}$$

shows that, for the SNR-limited scenario (defined in Sec. IV-B), where $\gamma\mathsf{SNR} \ll 1$ over a wide range of L, we have

$$\frac{\mathrm{d}I_{\text{GMI}}}{\mathrm{d}L} \approx -\mathsf{SNR}\frac{\mathrm{d}\gamma}{\mathrm{d}L}, \tag{117}$$

which implies that the penalty of a small bias of gain control increases approximately linearly with the SNR; see Fig. 8 and Fig. 9. This is consistent with Theorem 12 and Theorem 14. For the resolution-limited scenario (also defined in Sec. IV-B), where $\gamma\mathsf{SNR} \gg 1$, we have

$$\frac{\mathrm{d}I_{\text{GMI}}}{\mathrm{d}L} \approx -\frac{1}{\gamma}\frac{\mathrm{d}\gamma}{\mathrm{d}L}, \tag{118}$$

which does not varies with the SNR, implying that the achievable rate saturates; see high-SNR curves in Fig. 4(e) and Fig. 4(f).

Apart from the aforementioned cases, in general it is not easy to establish a global property for $I_{\text{GMI}}(L)$ from the derivative. We next propose approximations for $I_{\text{GMI}}(L)$ by asymptotic analysis of $\gamma(L)$.

1) Overload Region: In this region it is not easy to find a simple and accurate approximation for $I_{\text{GMI}}(L)$. In the range of resolution of practical interest (e.g., $b \le 12$ bits), the loading factor in the overload region is small, so that the asymptotics (101) can be inaccurate and the asymptotic expression of $\bar{\gamma}(L)$ in (100) is less useful. For example, Fig. 10 and Fig. 11 show that $C - \log\left(1 + 4\phi(L)L^{-3}\mathsf{SNR}\right)$ is not a satisfactory approximation. Instead, using

$$C - \log\left(1 + \underline{\mathsf{mse}}_{\mathsf{o}}\mathsf{SNR}\right). \tag{119}$$

may approximate the waterfall better, but it cannot be expressed in a closed form. A special case is high-resolution quantization at high SNR, where the waterfall in the overload region can be approximated by the high-resolution limit of the saturation rate $\log(1/\gamma)$ as

$$\log\frac{1}{\bar{\gamma}(L)} = \log\left((1+o_L(1))\frac{L^3}{4\phi(L)}\right) \tag{120a}$$

$$\approx \log\left(\frac{L^3}{4\phi(L)}\right) \tag{120b}$$

$$= L^2/2 + 3\ln L + \ln\frac{\sqrt{2\pi}}{4} \text{ nats/c.u.}, \tag{120c}$$

which does not depend on resolution or SNR; see Figs. (5(d)-5(f)).

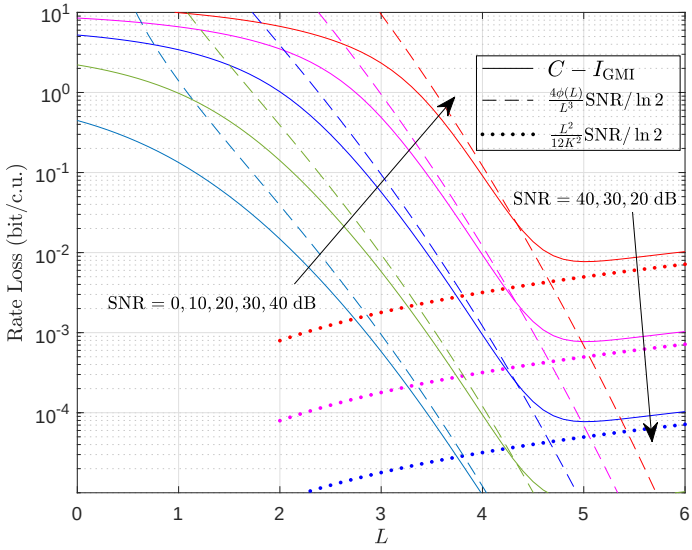


Fig. 8. Approximations of small irreducible rate loss: $b = 12$ bits.

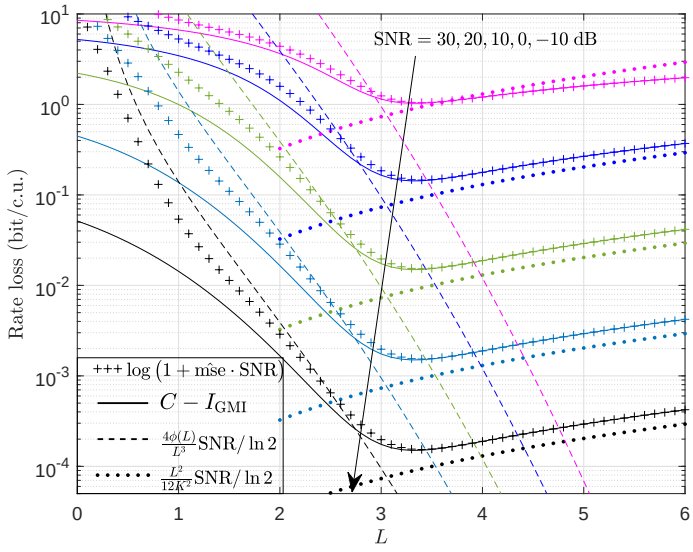


Fig. 9. Approximations of small irreducible rate loss: $b = 6$ bits.

2) Underload Region: It is easier to approximate $I_{\text{GMI}}(L)$ in this region. We utilize the asymptotic expression (111) of $\bar{\gamma}$ and let the resolution be finite with $2K$ levels, yielding

$$I_{\text{GMI}}(L) \approx C - \log\left(1 + \frac{12L^2}{K^2}\text{SNR}\right), \tag{121}$$

which is highly accurate even when the resolution is low; see Fig. 10 and Fig. 11. We have also considered an even simpler approximation $C - 12L^2K^{-2}\text{SNR}$ nats/c.u. in the high-resolution scenario; see Fig. 8 and Fig. 9. However, it becomes less useful as the loading loss increases, as shown in Fig. 10 and Fig. 11. The approximation (121) also implies that the saturation rate can be approximated by

$$\bar{I}_{\text{GMI}}(L) = \log(1/\gamma) \approx \log\frac{K^2}{12L^2}. \tag{122}$$

In Figs. 5(d)-5(f), we can observe that (122) well approximates I_{GMI} in the resolution-limited scenario.

VI. Asymptotic Analysis: Characterization of L^* and I^*_{GMI}

This section focuses on I^*_{GMI}, the maximum achievable rate of the transceiver architecture described in Sec. II-A when the receiver quantization is uniform with the optimal loading factor L^*. In Sec. VI-A, we consider approximations of the optimal

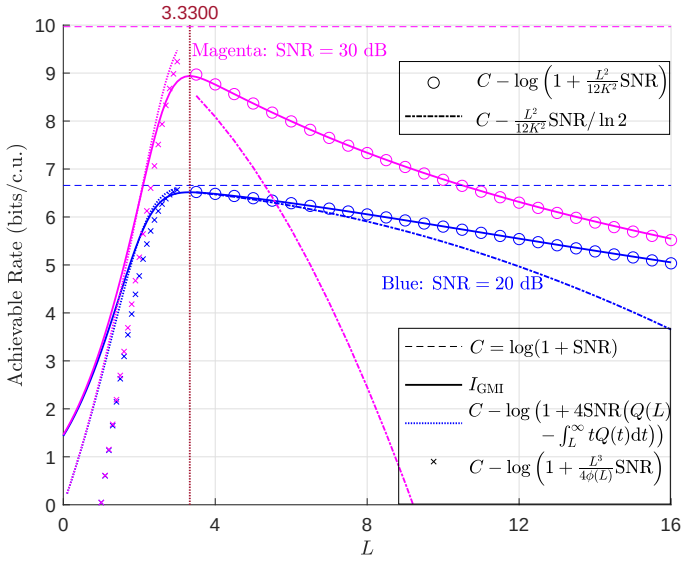


Fig. 10. Approximations of $I_{\mathrm{GMI}}(L)$: $b = 6$ bits.

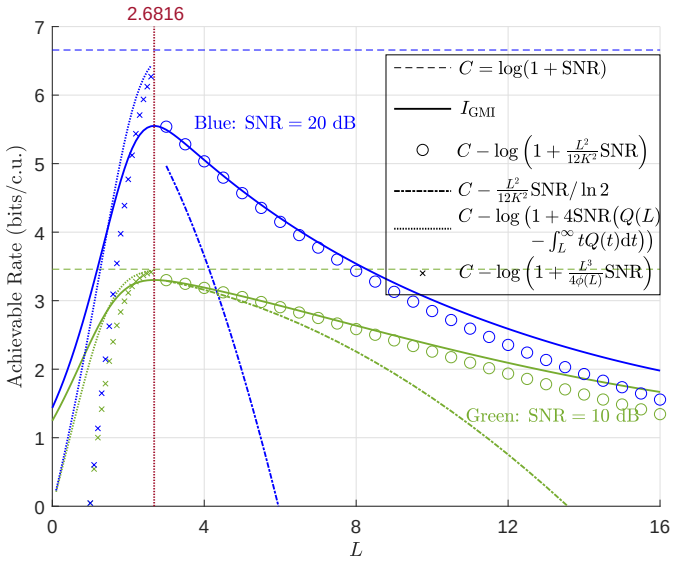


Fig. 11. Approximations of $I_{\mathrm{GMI}}(L)$: $b = 4$ bits.

loading factor L^*, which is essential for achieving I^*_{GMI} by accurate gain control. In Sec. VI-B, we characterize I^*_{GMI} as a function of the resolution b, and further provide some per-bit rules for different performance metrics such as saturation rates and irreducible rate loss.

A. Approximations of Optimal Loading Factor

We first provide a new proof of the property (61) for the step size and loading factor of the optimal uniform quantizer. In [67], it was shown that the uniform quantizer that minimizes the MSE must satisfy (61) if the input density has infinite support. Here we show that (61) is a natural corollary of the achievable rate results in Theorem 12 and Theorem 14.

New Proof of (61): We consider two possible cases other than (61) and exclude them respectively. First, if the high-resolution limit of the optimal step size, namely $\lim_{K\to\infty} \ell^*$, is strictly larger than zero, then $\lim_{K\to\infty} L^* = \infty$. However, we can infer from Theorem 14 that, for a step size bounded away from zero, if the resolution is high enough, then there must exist a smaller step size that reduces the loss in achievable rate. Thus $\lim_{K\to\infty} \ell^*$ must be equal to zero. Second, on the other side, if $\lim_{K\to\infty} L^* < \infty$, then $\lim_{K\to\infty} \ell^* = 0$. However, we can infer from Theorem 12 that, given a finite L, if the resolution is high enough, then there must exist a larger L that reduces the loss in achievable rate. Thus $\lim_{K\to\infty} L^*$ cannot be finite. Therefore, we conclude that $\lim_{K\to\infty} \ell^* = 0$ and $\lim_{K\to\infty} L^* = \infty$ hold simultaneously. ■

The scaling law (62) was obtained in [46] (as a special case of a more general result) by analyzing the derivative of the MSE. Here we give a simpler proof of (62) from the condition (52) for achievable rate maximization. The proof utilizes the Euler-Maclaurin formula.

New Proof of (62): In the high-resolution regime, by combining (108), (109), and (52) we obtain

$$\frac{\int_{K\ell^*}^{\infty} \exp\frac{-t^2}{2}\mathrm{d}t + o(\ell^{*2})}{\int_{K\ell^*}^{\infty} 2tQ(t)\mathrm{d}t - \frac{1}{24}\ell^{*2} + o(\ell^{*2})} = \sqrt{2\pi}, \tag{123}$$

yielding

$$\lim_{K\to\infty} \frac{Q(L^*)}{\int_{L^*}^{\infty} 2tQ(t)\mathrm{d}t - \frac{1}{24K^2}L^{*2}} = 1. \tag{124}$$

Noting that

$$\lim_{K\to\infty} \frac{Q(L^*)}{\int_{L^*}^{\infty} tQ(t)\mathrm{d}t} = \lim_{K\to\infty} \frac{-\phi(L^*)}{-L^*Q(L^*)} = 1, \tag{125}$$

where the second equality follows from (102), we obtain

$$\lim_{K\to\infty} \frac{24K^2 \exp\frac{-L^{*2}}{2}}{\sqrt{2\pi}L^{*3}} = 1. \tag{126}$$

The proof is completed by noting that (126) implies (62). ■

Although (62) provides a simple approximation of L^* as $\hat{L}_1 = 2\sqrt{\ln(2K)}$, it can be refined by more elaborate techniques. In [46, Sec. V] three approximations have been proposed. For Gaussian input, the first is just $\hat{L}_1$, and the other two satisfy a stronger condition

$$\lim_{K\to\infty} \left\{L^* - \hat{L}_i\right\} = 0,\ i = 2, 3, \tag{127}$$

where

$$\hat{L}_2 = \sqrt{4\ln(2K) - 3\ln\ln(2K) - \ln\frac{32\pi}{9}} \tag{128}$$

and

$$\hat{L}_3 = \sqrt{\hat{L}_2^2 + \epsilon},\ K > 1, \tag{129}$$

with an error term

$$\epsilon = 2\ln\left[\left(1 + \frac{4\ln(2K)}{2K}\right)\left(1 - \frac{3}{4\ln(2K)}\right)\cdot\left(1 + \frac{1}{2\ln(2K)}\left(\frac{3}{2}\ln\ln(2K) + \ln\frac{4\sqrt{2\pi}}{3}\right)\right)^{3/2}\right] \tag{130}$$

included, which satisfies $\lim_{K\to\infty}\epsilon = 0$.

In the following result, we adopt a new way to approximate L^*, which begins from (124) in the preceding new proof of (62).

Proposition 15: *In the channel (4) under i.i.d. complex Gaussian codebook and nearest neighbor decoding rule (7), the optimal loading factor L^* that maximizes the GMI (19) satisfies*

$$\lim_{K\to\infty} \left\{L^* - \hat{L}_0\right\} = 0, \tag{131}$$

where $\hat{L}_0$ is the unique real-valued solution of the transcendental equation[10]

$$L^2 + 6\ln L - \ln\frac{18}{\pi} = 4\ln(2K). \tag{132}$$

Proof: According to (61) and (123), the difference between the numerator and the denominator of (124) vanishes as K tends to infinity. Therefore, L^* can be approximated with vanishing error by the solution of

$$\int_{L}^{\infty} 2tQ(t)\mathrm{d}t - Q(L) = \frac{L^2}{24K^2}, \tag{133}$$

[10]A transcendental equation $x + \ln x = \ln a$ can be solved by the Lambert W function as $x = W(a)$. Similarly, the solution $\hat{L}$ in Proposition 15 can be given by a special case of the generalized Lambert W function, although less is known about this function [69].

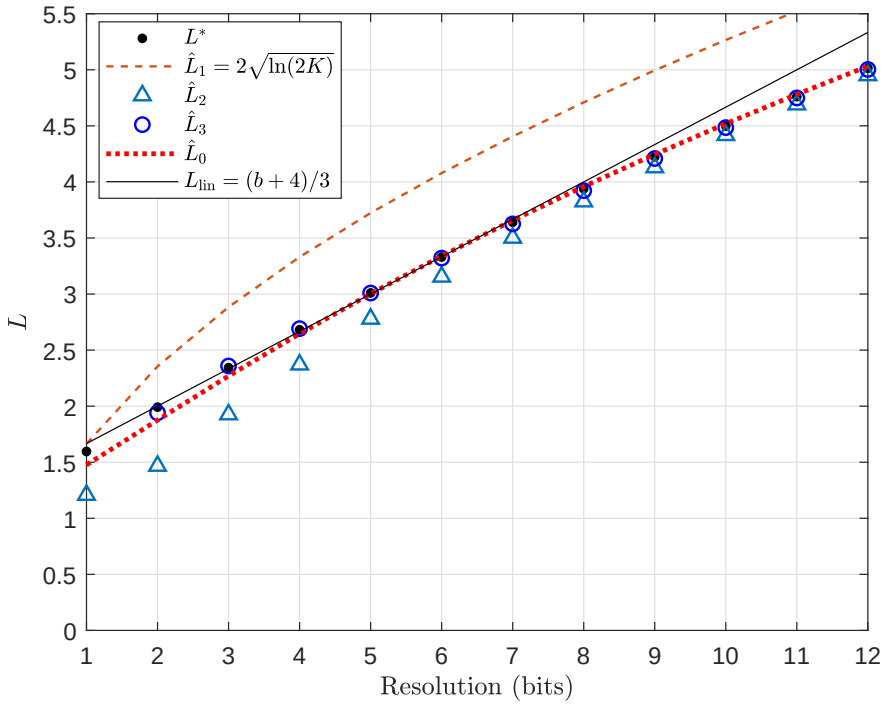


Fig. 12. Numerical results, asymptotics, and approximations of L^*.

since its LHS and RHS are continuous functions of L. According to (102) and its variation

$$\left(1 - \frac{1}{1+t^2}\right)\phi(t) < 2tQ(t) - \phi(t) < \phi(t), \tag{134}$$

we note that, instead of (133), we can turn to the solution of equation

$$24K^2\phi(L) = L^3, \tag{135}$$

which is equivalent to (132). The uniqueness of $\hat{L}$ is clear since the LHS of (132) is a smooth and monotonically increasing function of L. ■

Remark: In fact, the equation (135) can also be derived from analysis of the MSE: Just let the derivative of the MSE approximation (115) be zero, and omit the exponential terms therein except for the dominated one.

In Fig. 12, we show L^*, its approximation $\hat{L}_0$ from (131) in Proposition 15, the simple approximation $2\sqrt{\ln(2K)}$, and the approximations $\hat{L}_2$ and $\hat{L}_3$ from [46]. Note that the case $b = 1$ does not require gain control. We see that the new approximation $\hat{L}_0$ is accurate for all resolutions $b \geq 2$. It is much better than $\hat{L}_2$ and close to the more complicated approximation $\hat{L}_3$.

Remark: To approach I^*_{GMI}, it is sufficient to use a simple linear approximation of L^* as

$$\hat{L}_{\text{lin}} = \frac{b+4}{3}. \tag{136}$$

Fig. 12 shows that it is accurate for $2 \leq b \leq 7$. When $b \geq 8$, its accuracy degrades, but the loading loss caused is negligible since it grows quadratically with L; see numerical results in Fig. 4.

B. Maximum Achievable Rate Approximations and Per-Bit Rules

We have found a simple relationship between I^*_{GMI} and mmse in Theorem 8, and we have also given numerical results of mmse for $b = 1, 2, ..., 12$ in Table I. However, a direct connection between I^*_{GMI} and the resolution is still very useful. The following result provides such a connection via approximating mmse by b.

Proposition 16: *The achievable rate given in (55) satisfies*

$$I^*_{\text{GMI}} = \hat{I}_{\text{GMI}}(b) + o_b(1), \tag{137}$$

where

$$\hat{I}_{\text{GMI}}(b) = C - \log\left(1 + \frac{4b\ln 2}{3 \cdot 4^b}\mathsf{SNR}\right). \tag{138}$$

Proof: From (62) and (63) we obtain

$$\lim_{K\to\infty} \frac{3K^2}{\ln(2K)}\mathsf{mmse} = 1, \tag{139}$$

which yields a high-resolution approximation of MMSE given by

$$\mathsf{mmse} = (1+o_b(1))\frac{4b\ln 2}{3\cdot 4^b} \approx \frac{4b\ln 2}{3\cdot 4^b}. \tag{140}$$

If we replace mmse in (55) by $\frac{4b\ln 2}{3\cdot 4^b}$, then we obtain $\hat{I}_{\text{GMI}}(b)$. According to (140), it is straightforward to show that the gap between I^*_{GMI} and $\hat{I}_{\text{GMI}}(b)$ is $o_b(1)$ for an arbitrary finite SNR, thereby completing the proof. ■

The approximation (140) implies a 6-dB-per-bit-rule as

$$10\log_{10}\mathsf{SNR}_q \approx 6.02b - 10\log_{10} b + 0.34 \text{ (dB)}, \tag{141}$$

which has been known since [46]. Therefore, in the high-resolution regime, each additional bit in resolution reduces the MSE by four times:

$$\lim_{b\to\infty} \frac{\mathsf{mmse}(b)}{\mathsf{mmse}(b-1)} = 4. \tag{142}$$

The approximation (140) is still not accurate enough for moderate to low resolutions; see Table I. More approximations for MMSE or SNR_q can be found in [46, Sec. IV] based on refined asymptotic formulas of L^* and its asymptotic relationship with MMSE. However, using (140) is enough to get the simple and useful approximation of I^*_{GMI} in Proposition 16. In Fig. 13 we compare I^*_{GMI} with its approximation $\hat{I}_{\text{GMI}}$. It is shown that the accuracy is acceptable when $b \geq 2$.

We next introduce some per-bit rules for other performance metrics. The following one is obtained from Proposition 16 immediately.

- *A 2-bpcu-per-bit rule for saturation rate*: The saturation rate $\bar{I}^*_{\text{GMI}}(b)$ given in (66) satisfies

$$\bar{I}^*_{\text{GMI}}(b) = \hat{\bar{I}}_{\text{GMI}}(b) + o(1), \tag{143}$$

where

$$\hat{\bar{I}}_{\text{GMI}}(b) = 2b - \log_2 b + 0.11 \text{ bits/c.u.}, \tag{144}$$

which implies that

$$\lim_{b\to\infty} \left(\bar{I}^*_{\text{GMI}}(b) - \bar{I}^*_{\text{GMI}}(b-1)\right) = 2 \text{ bits/c.u.} \tag{145}$$

Numerical evaluations of $\hat{\bar{I}}_{\text{GMI}}(b)$ for $1 \leq b \leq 12$ are shown in Table I. In Fig. 14, it is shown that mmse decreases exponentially as the resolution increases, and correspondingly, the saturation rate $\bar{I}^*_{\text{GMI}}$ grows approximately linearly with the resolution. The improvement per bit is 1.6-1.9 bits/c.u. in our range of interest, although the high-resolution limit is 2 bits/c.u.

The second per-bit rule can be observed from results in Table I, suggesting a roughly 5-dB-per-bit increase of SNR_q:

$$10\log_{10}\mathsf{SNR}_q \approx 5b. \tag{146}$$

Since the rate loss in (55) is determined by $\mathsf{SNR}/\mathsf{SNR}_q$, (146) implies that, in our range of interest, a 5 dB increase of SNR requires an extra bit of resolution to maintain the same rate loss. More specifically, we have the following rule.

- *A 5-dB-per-bit rule for irreducible rate loss*: We require a resolution of at least

$$2\log_{10}\mathsf{SNR} + b_0 \tag{147}$$

bits so that the irreducible loss $C - I^*_{\text{GMI}}$ can be as small as 10^{-b_0} bits/c.u., where $b_0 \geq 0$.

This rule can be confirmed by numerical results in Fig. 15. Although the 6-dB-per-bit rule is well-known in quantization theory and gives the correct asymptotics (as shown in (141)), it is less accurate unless the rate loss is extremely small, see Fig. 16.

VII. Discussions on Further Quantization Rules at the Receiver

Based on the analytical framework given in Theorem 2, a major part of this paper has focused on the simplest receiver quantization scheme, namely scalar uniform quantization with mid-rise levels. This section briefly discusses nonuniform and other types of quantization rules in several aspects.

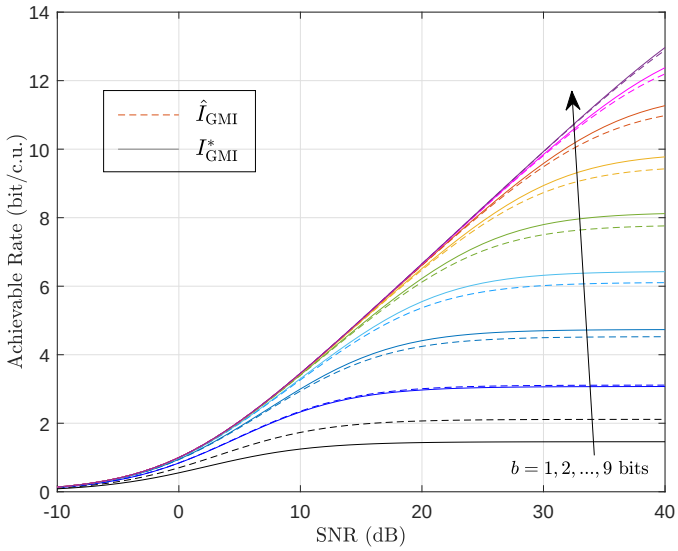


Fig. 13. Approximation of I^*_{GMI}.

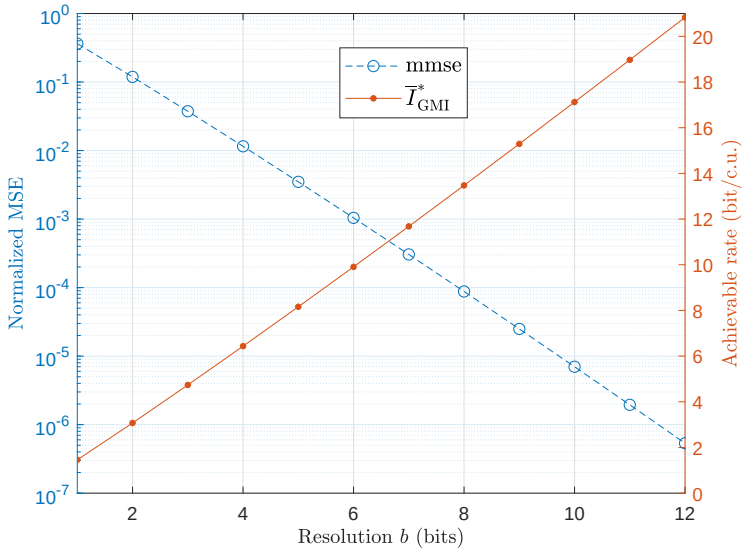


Fig. 14. Normalized MMSE and saturation rate under different resolutions.

A. Relationship Between MSE and Achievable Rate: Numerical Examples

For the channel (4) with a uniform quantizer $q(\cdot)$ whose resolution is at least two bits, we have shown that the unique gain control factor that minimizes the MSE also maximizes the GMI. When non-uniform quantization rules are allowed, one may vary the thresholds $\{\ell_k\}$ and levels $\{y_k\}$ of $q(\cdot)$ (possibly under some constraints) to reduce γ so that the GMI (19) can be optimized, yielding thresholds $\{\ell_k^*\}$ and levels $\{y_k^*\}$. However, when this optimized quantization rule is applied in a given channel, to achieve that optimized GMI, we need to set the gain control factor g appropriately according to σ_v (the standard deviation of the quantizer input) to guarantee that the thresholds satisfy $\ell_k = l_k/g = \ell_k^*$ for $1 \le k < K$. Otherwise, the performance will be degraded. We next explore the impact of gain control on the MSE and the GMI of the channel (4) under nonuniform quantization rules by examples; see numerical results given in Fig. 17, where we set $\mathsf{SNR} = 10$ dB. The examples are described as follows, each of which satisfies $2K = 4$, i.e., $b = 2$.

- The uniform quantizer (with equispaced thresholds and mid-rise levels); see Fig. 17(a) and Fig. 17(d).
- The optimal nonuniform quantizer: the thresholds and levels are optimized to maximize the GMI; see Fig. 17(b).
- Two nonuniform quantizers without optimization: the first one has monotonically increasing levels, and the second one is highly nonlinear since its levels are no longer monotonically increasing; see Fig. 17(c) and Fig. 17(f), respectively.
- An optimized quantizer with equispaced thresholds: its thresholds and levels are optimized to maximize the GMI under

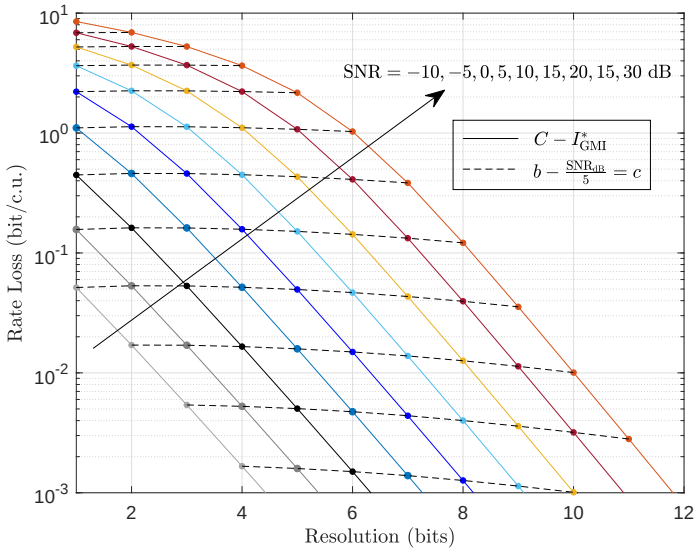


Fig. 15. A 5-dB-per-bit rule for irreducible rate loss.

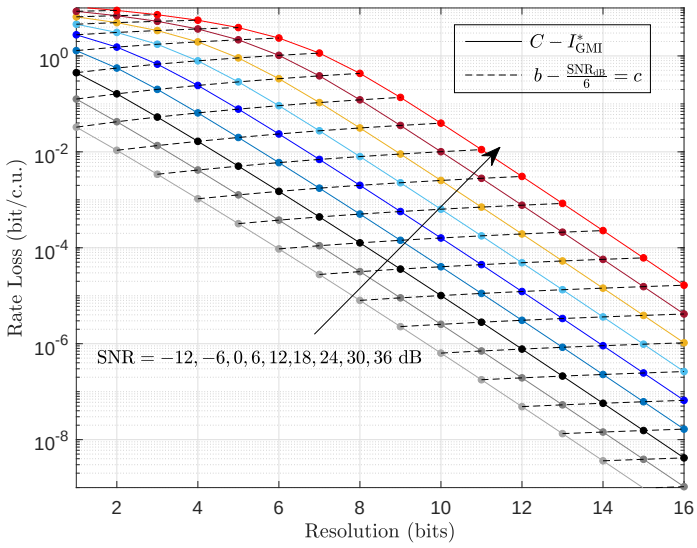


Fig. 16. For irreducible rate loss, a 6-dB-per-bit rule is inaccurate.

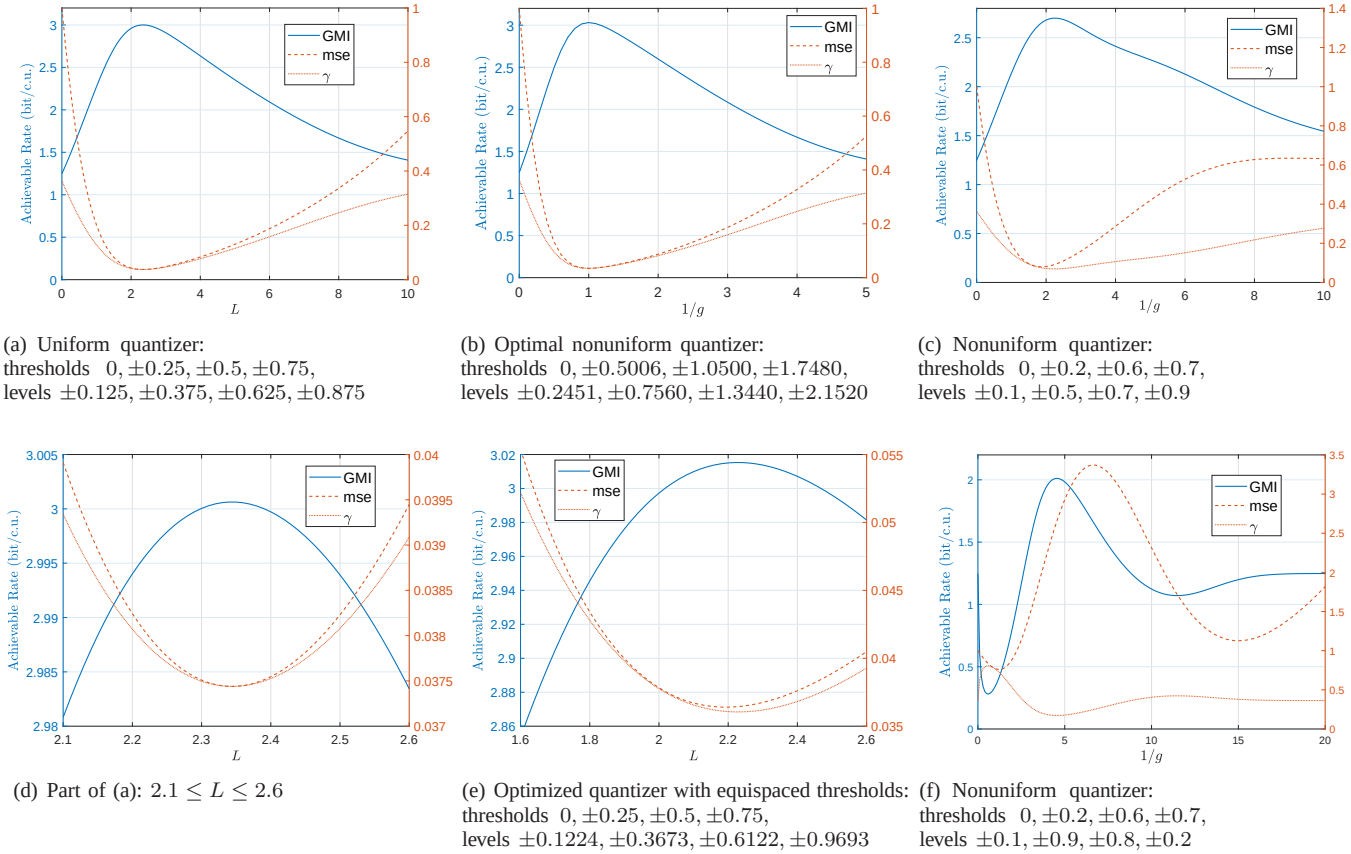
the constraint that the thresholds must be equispaced; see Fig. 17(e).

From Fig. 17(a), we confirm that, under uniform quantization, the parameter γ converges to mmse when the loading factor approaches its optimal value L^* (in this example it is 2.3441); see also Fig. 17(d). Interestingly, this consistency also occurs in Fig. 17(b), where we use the optimal 8-level quantizer obtained numerically in [30] by Lloyd's algorithm [70]. But for the nonuniform quantizer given in Fig. 17(c) and the optimized quantizer with equispaced thresholds (also obtained in [30]) given in Fig. 17(e), the gain control factor that maximizes the GMI and the one that minimizes the MSE are different. The highly nonlinear quantizer given in Fig. 17(f) provides an example of unusual behavior of $I_{\mathrm{GMI}}(g)$, which achieves its maximum when the MSE is relatively large.

Remark: In achievable rate evaluation, if we use the additive noise model of the quantizer, i.e., treating the MSE as additive Gaussian noise, then we obtain an estimate as

$$\hat{R} = \log\left(1 + \frac{\mathcal{E}_x}{\sigma^2 + 2 \cdot \mathsf{mse} \cdot \sigma_v^2}\right) \tag{148}$$

$$= \log(1 + \mathsf{SNR}) - \log\left(1 + \frac{\mathsf{mse}}{1 + \mathsf{mse}}\mathsf{SNR}\right), \tag{149}$$



(a) Uniform quantizer: thresholds $0, \pm 0.25, \pm 0.5, \pm 0.75$, levels $\pm 0.125, \pm 0.375, \pm 0.625, \pm 0.875$

(b) Optimal nonuniform quantizer: thresholds $0, \pm 0.5006, \pm 1.0500, \pm 1.7480$, levels $\pm 0.2451, \pm 0.7560, \pm 1.3440, \pm 2.1520$

(c) Nonuniform quantizer: thresholds $0, \pm 0.2, \pm 0.6, \pm 0.7$, levels $\pm 0.1, \pm 0.5, \pm 0.7, \pm 0.9$

(d) Part of (a): $2.1 \leq L \leq 2.6$

(e) Optimized quantizer with equispaced thresholds: thresholds $0, \pm 0.25, \pm 0.5, \pm 0.75$, levels $\pm 0.1224, \pm 0.3673, \pm 0.6122, \pm 0.9693$

(f) Nonuniform quantizer: thresholds $0, \pm 0.2, \pm 0.6, \pm 0.7$, levels $\pm 0.1, \pm 0.9, \pm 0.8, \pm 0.2$

Fig. 17. GMI, MSE and γ under gain control: $\mathsf{SNR} = 10$ dB.

which may cause underestimate or overestimate. In particular, it overestimates the achievable rate when the optimal uniform quantization is used, while it may significantly underestimate the achievable rate when mse is much larger than γ, which is possible; see the numerical results in Fig. 17 when the gain control factor (or loading factor) is far away from its optimal value.

B. On Improving Uniform Quantization by Post-Processing

In [39], [57], it has been shown that, the GMI (14) given in Proposition 1 can be improved by introducing a post-processing of the channel output as $Y \to \tilde{Y}$, thereby extending the channel $X \to Y$ to $X \to \tilde{Y}$; The GMI is maximized by letting

$$\tilde{Y} = \mathrm{E}\left[X|Y\right], \tag{150}$$

i.e., using a conditional expectation operator as the post-processor, which gives an MMSE estimation of X under the observation Y. Thus we can infer that the optimized quantizer with equispaced thresholds considered in Sec. VII-A is equivalent to a uniform quantizer combined with an MMSE post-processing of its output. In fact, for a quantizer with ℓ-spaced thresholds, the optimal levels have been shown to be [30]

$$y_k^* = \frac{\phi((k-1)\ell) - \phi(k\ell)}{Q((k-1)\ell) - Q(k\ell)},\ k = 1, ..., K-1 \tag{151}$$

and

$$y_K^* = \frac{\phi((K-1)\ell)}{Q((K-1)\ell)}. \tag{152}$$

Such levels satisfy $y_k^* = \mathrm{E}\left[V|V \in [(k-1)\ell, k\ell)\right],\ k = 1, ..., K-1$ and $y_K^* = \mathrm{E}\left[V|V \geq (K-1)\ell\right]$, where $V \sim \mathcal{N}(0, \sigma_v^2)$ is the input of the quantizer; i.e., the optimal level y_k^* is the *centroid* of the corresponding interval. This is consistent with the aforementioned result that a conditional expectation operator maximizes the GMI. Comparing Fig. 17(e) and Fig. 17(d), we see that, under the same loading factor $L = k\ell$, such a post-processing indeed improves the GMI slightly. Interestingly, the optimal loading factor that maximizes the GMI in Fig. 17(e) is smaller than the optimal loading factor in Fig. 17(d). In other words, when post-processing is available, the optimal step size of the uniform quantizer becomes smaller, and a concatenation

of an MMSE uniform quantizer and an MMSE post-processor is strictly worse than the optimized quantizer with equispaced thresholds.

Since the example Fig. 17(e) shows that the gain of the MMSE post-processing is limited for low-resolution uniform quantization, we may infer that the gain is also limited under high resolution. In fact, as the step size decreases, the centroid of the interval $[(k-1)\ell, k\ell)$ given in (151) tends to be its midpoint.[11] To see this, note that for an arbitrary $c > \ell/2$,

$$\lim_{\ell\to 0} \frac{\phi\left(c-\frac{\ell}{2}\right)-\phi\left(c+\frac{\ell}{2}\right)}{Q\left(c-\frac{\ell}{2}\right)-Q\left(c+\frac{\ell}{2}\right)} = \lim_{\ell\to 0} \frac{\phi\left(c-\frac{\ell}{2}\right)-\phi\left(c+\frac{\ell}{2}\right)}{\int_{c-\frac{\ell}{2}}^{c+\frac{\ell}{2}} \phi(t)\mathrm{d}t} \tag{153a}$$

$$= \lim_{\ell\to 0} \frac{\frac{c-\frac{\ell}{2}}{2}\phi\left(c-\frac{\ell}{2}\right)+\frac{c+\frac{\ell}{2}}{2}\phi\left(c+\frac{\ell}{2}\right)}{\frac{1}{2}\left(\phi\left(c+\frac{\ell}{2}\right)+\phi\left(c-\frac{\ell}{2}\right)\right)} \tag{153b}$$

$$= c, \tag{153c}$$

which follows from L'Hôpital's rule. Moreover, according to (102), the largest level y_K^* given in (152) satisfies

$$(K-1)\ell > y_K^* > (K-1)\ell + \frac{1}{(K-1)\ell}, \tag{154}$$

implying that y_K^*/ℓ converges to $K-1$ as $\ell \to 0$. From these facts we can conclude that the gain of MMSE post-processing asymptotically vanishes as the resolution increases.

C. On Possible Gain of Futher Quantization Rules

For communication receivers, nonuniform quantization rules can be realized directly, or indirectly by combining a uniform quantizer with a pre-processing of its input (e.g., companding in pulse-code modulation (PCM)), depending on the cost of implementation. There have been some works on numerical optimization of nonuniform quantization rules with respect to different performance measures, e.g., cutoff rate [71], mutual information [8], [72], GMI [30], and MSE [21], [23]. But no general characterization of the performance gain has been given.

As shown in Fig. 17(b), for nonuniform quantizer, there is also a consistency between GMI maximization and MSE minimization, but we must optimize all the thresholds and levels, rather than only a single factor (g or L). In [30, Appendix E] a proof of this consistency has been given. Thus, the maximum achievable rate can be expressed as

$$I_{\mathrm{GMI}}^* = \log(1+\mathsf{SNR}) - \log(1+\mathsf{mmse}^{\mathrm{nu}}\cdot\mathsf{SNR}), \tag{155}$$

where $\mathsf{mmse}^{\mathrm{nu}}$ is the normalized minimum MSE of all possible nonuniform quantizers. This result can also be understood via a similar geometrical interpretation as that given in Sec. IV-C. From (155) we can infer that, under our setting of transceiver architecture, the maximum gain in achievable rate from replacing the optimal uniform quantizer by the optimal nonuniform quantizer is determined by the reduction of the normalized MMSE from mmse to $\mathsf{mmse}^{\mathrm{nu}}$. The numerical results in Fig. 17(a) and Fig. 17(b) show that the rate gain is marginal for $b=2$. For higher resolutions the rate gain grows slowly. When $b=4$, according to the numerical results in [30], the optimal non-uniform quantization improves the saturation rate $\log(1/\gamma)$ by no more than 4.4%, and optimized quantization with equispaced thresholds improves the saturation rate by 1.5%. In fact, for nonuniform quantization with Gaussian input, the asymptotic quantization theory, in particular the Panter-Dite formula [73], yields a "6-dB-per-bit" rule[12] for the maximum SNR as [41]

$$10\log_{10}\mathsf{SNR}_q^{\mathrm{nu}} = 6.02b - 4.35 + o(1), \tag{156}$$

where $\mathsf{SNR}_q^{\mathrm{nu}} = 1/\mathsf{mmse}^{\mathrm{nu}}$. Therefore, under our setting, the rate gain increases logarithmically with the resolution b, and for $b \le 10$ the gain is limited (cf. (141)).

According to (156), even if the best nonuniform scalar quantization is used, the MSE achieved in the high-resolution regime is still 2.72 times larger than the theoretical limit implied by the rate-distortion function of a Gaussian source [42].[13] If we allow *coded* uniform scalar quantization (i.e., representing the quantizer output by a variable-rate lossless code), the gap can be reduced to only 1.53 dB ($\pi e/6$) worse than the theoretical limit (see [41] and references therein). Unfortunately, coded quantization does not help in communication receivers since the bottleneck therein is the limited resolution of the quantizer rather than the cost of representing its output. Applying vector quantization to communication receivers is a more challenging topic. Nevertheless, the gap to the fundamental limit reminds us that it is possible to alleviate the ADC resolution bottleneck in communication receivers by exploring new quantization mechanisms.

[11] In [54] the asymptotic convergence of centroids to midpoints was proved under uniform quantization with infinitely many levels.

[12] There is another version of the 6-dB-per-bit rule as $B_{\mathrm{eff}} = (\mathsf{SNR}_q - 1.76)/6.02$, which has been widely used in practice to calculate the effective number of bits (ENoB) of a quantizer [2], [3]. It is derived under assumptions that the granular distortion is uniformly distributed over $[-\ell/2, \ell/2]$ and there is no overload distortion; see [2] for more details.

[13] The comparison is made under the assumption that the signal at the receiver front-end is represented by b bits per channel use before feeding into the decoder. The MSE of the representation can be reduced by lossy source coding (treating the received signal as a source). A scalar quantizer is a very simple lossy source coding scheme.

VIII. Concluding Remarks

The goal of this study is to evaluate the impact of resolution reduction on information-theoretic limits of communications with receiver quantization. Leveraging the GMI as a basic tool, which enables us to take the decoding rule into consideration, we establish an array of exact and asymptotic results under a standard transceiver architecture. Our results indicate a critical issue in system design that arises as the resolution decreases, namely, optimizing the loading factor by gain control, which minimizes the loss in achievable rate by eliminating the loading loss. The remaining irreducible loss is an appropriate evaluation of the inherent robustness of the considered transceiver architecture. Our results also establish explicit connections between the MSE (affected by the gain control) and the rate loss. Although for general receiver quantization rules, smaller MSE does not necessarily imply smaller rate loss, for the commonly used uniform quantizer we prove that the unique loading factor that minimizes the MSE also maximizes the GMI (i.e., achieves zero loading loss). For perfect gain control, we show that the irreducible loss is determined by the product of the normalized MMSE and the SNR, and provide a geometrical interpretation for this result. Performance approximations and per-bit rules in this case are also given. For imperfect gain control, to understand its impact, we characterize the decay of small rate loss in the high-resolution regime, and propose approximations of the achievable rate as a function of the loading factor which is fairly accurate for moderate resolutions. These results provide insight into transceiver design with nonnegligible quantization effect, especially choice of quantizer resolution and design of AGC.

A limitation of this work is that the obtained analytical results are derived under the Gaussian codebook. Like many classical information-theoretic results given by mutual information, for the GMI, it is also not easy to get closed-form or analytical results without assuming Gaussian input. For practical systems with finite alphabet, we can expect that our result applies well for Gaussian-like constellations that comes from constellation shaping. But it is natural to ask whether our results are still useful in the cases of QAM, phase-shift keying (PSK), and other commonly used inputs, which are all bounded (unlike the unbounded Gaussian input). Since in these cases Proposition 1 does not apply, the GMI evaluation may rely heavily on numerical computation. We left this problem to future study. Nevertheless, the insight gained in this work will be very helpful for the finite-alphabet case.

Finally, we list some topics that can be addressed following our information-theoretic framework of receiver quantization.

- The orthogonal frequency division multiplexing (OFDM), which typically generate Gaussian-like signal. Therein, a new effect is that the nonlinear distortion due to quantization leads to intercarrier interference. The problem becomes more complicated when the channel introduces time-dispersion.
- Multiuser channels, especially the Gaussian multiple-access channel. In this case a new phenomenon due to receiver quantization is that, when successive interference cancellation (SIC) is used to decode a user, there exists residual interference from other users which may significantly reduces the achievable rate of that user.
- Multiantenna channels. We need an extended version of the proposed analytical framework which applies for different receiver architectures (e.g., MMSE receiver, maximal ratio combining receiver).

Appendix A
Proof of Proposition 4

We first note that all the adjustable levels in $\{\ell_k = l_k/g,\ k = 1, ..., K-1\}$ tend to zero as $g \to \infty$, and they tend to infinity as $g \to 0$. For the case of $g \to \infty$, from (21) and (22) we have $\lim_{g\to\infty} \mathcal{A}/y_K = \sqrt{2\pi}\phi(0)$ and $\lim_{g\to\infty} \mathcal{B}/y_K^2 = \pi Q(0)$. For the case of $g \to 0$, noting that $\mathcal{A}$ and $\mathcal{B}$ can be expressed by

$$\mathcal{A} = y_1 + \sqrt{2\pi} \sum_{k=1}^{K-1} \phi(\ell_k)(y_{k+1} - y_k) \tag{157}$$

and

$$\mathcal{B} = \frac{\pi}{2} y_1^2 + \pi \sum_{k=1}^{K-1} Q(\ell_k)(y_{k+1}^2 - y_k^2), \tag{158}$$

respectively, it is direct to check that $\lim_{g\to 0} \mathcal{A}/y_1 = 1$ and $\lim_{g\to 0} \mathcal{B}/y_1^2 = \pi/2$. Therefore, in both cases γ tends to $1 - 2/\pi$ and the limits of the GMI are the same, $I_{\text{GMI}}^{\text{1-bit}}$.

Appendix B
Proof of Proposition 5

We begin from the definition (39) in which the quantization rule is given by (8) and is equivalent to

$$q(gV) = y_k \cdot \text{sgn}(V),\ \text{if}\ \ell_{k-1} \le v < \ell_k, \tag{159}$$

where $v = |V|/\sigma_v \sim \mathcal{N}(0,1)$. Utilizing the symmetries of the input distribution and the quantization rule with respect to the origin, we have

$$\mathsf{mse} = 2\sum_{k=1}^{K}\int_{\ell_{k-1}}^{\ell_k} y_k^2\phi(v)\mathrm{d}v - 2\sum_{k=1}^{K}\int_{\ell_{k-1}}^{\ell_k} 2y_k v\phi(v)\mathrm{d}v + 1 \tag{160a}$$

$$= 1 - 4\sum_{k=1}^{K} y_k\left(\phi(\ell_{k-1}) - \phi(\ell_k)\right) + 2\sum_{k=1}^{K} y_k^2\left(Q(\ell_{k-1}) - Q(\ell_k)\right) \tag{160b}$$

$$= 1 - \frac{2}{\pi}\left(\sqrt{2\pi}\mathcal{A} - \mathcal{B}\right). \tag{160c}$$

The lower bound and condition of equality follow from

$$\mathsf{mse} - \gamma = \frac{1}{\mathcal{B}}\left(\mathcal{A} - \sqrt{\frac{2}{\pi}}\mathcal{B}\right)^2 \geq 0. \tag{161}$$

Appendix C
Proof of Proposition 6

According to (21), we can derive (44) by straightforward derivation as

$$\mathcal{A} = \sum_{k=1}^{K-1}\left(k\ell - \frac{\ell}{2}\right)\left(\exp\frac{-(k-1)^2\ell^2}{2} - \exp\frac{-k^2\ell^2}{2}\right) + \left(K\ell - \frac{\ell}{2}\right)\exp\frac{-(K-1)^2\ell^2}{2} \tag{162a}$$

$$= \sum_{k=1}^{K-1} k\ell\left(\exp\frac{-(k-1)^2\ell^2}{2} - \exp\frac{-k^2\ell^2}{2}\right) - \frac{1}{2}\sum_{k=1}^{K-1}\ell\left(\exp\frac{-(k-1)^2\ell^2}{2} - \exp\frac{-k^2\ell^2}{2}\right)$$
$$+ K\ell\exp\frac{-(K-1)^2\ell^2}{2} - \frac{\ell}{2}\exp\frac{-(K-1)^2\ell^2}{2} \tag{162b}$$

$$= \sum_{k=0}^{K-1}\ell\cdot\exp\frac{-k^2\ell^2}{2} - \frac{\ell}{2}, \tag{162c}$$

and according to (22), we obtain (45) as

$$\mathcal{B} = \pi\sum_{k=1}^{K-1}\left(k - \frac{1}{2}\right)^2\ell^2\left(Q((k-1)\ell) - Q(k\ell)\right) + \pi\left(K - \frac{1}{2}\right)^2\ell^2 Q((K-1)\ell) \tag{163a}$$

$$= \pi\ell^2\Bigg(\sum_{k=1}^{K-1}k^2\left(Q((k-1)\ell) - Q(k\ell)\right) - \sum_{k=1}^{K-1}k\left(Q((k-1)\ell) - Q(k\ell)\right) + \sum_{k=1}^{K-1}\frac{1}{4}\left(Q((k-1)\ell) - Q(k\ell)\right)\Bigg)$$
$$+ \pi\left(K - \frac{1}{2}\right)^2\ell^2 Q((K-1)\ell) \tag{163b}$$

$$= \pi\ell^2\left(\sum_{k=0}^{K-1}(2k+1)Q(k\ell) - \sum_{k=0}^{K-1}Q(k\ell) + \frac{1}{8}\right) \tag{163c}$$

$$= \pi\sum_{k=0}^{K-1}2k\ell^2 Q(k\ell) + \frac{1}{8}\pi\ell^2. \tag{163d}$$

Appendix D
Proof of Lemma 7

Let $K > 1$. According to Theorem 2, the LHS of (51) is equivalent to $\frac{\mathrm{d}\gamma}{\mathrm{d}\ell} = 0$, or

$$\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}\ell} = \frac{\mathcal{A}}{2\mathcal{B}}\frac{\mathrm{d}\mathcal{B}}{\mathrm{d}\ell}. \tag{164}$$

According to Proposition 5, the RHS of (51) is equivalent to

$$\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}\ell} = \frac{1}{\sqrt{2\pi}}\frac{\mathrm{d}\mathcal{B}}{\mathrm{d}\ell}. \tag{165}$$

Let

$$\mathcal{C} = \sum_{k=0}^{K-1} \ell \cdot \left(k^2\ell^2 \cdot \exp \frac{-k^2\ell^2}{2} \right), \tag{166}$$

which is strictly positive when $K > 1$. Noting that

$$\ell \cdot \frac{\mathrm{d}\mathcal{A}}{\mathrm{d}\ell} = \sum_{k=0}^{K-1} \ell \cdot \exp \frac{-k^2\ell^2}{2} - \frac{\ell}{2} - \sum_{k=0}^{K-1} \ell \cdot \left(k^2\ell^2 \cdot \exp \frac{-k^2\ell^2}{2} \right) \tag{167a}$$

$$= \mathcal{A} - \sum_{k=0}^{K-1} \ell \cdot \left(k^2\ell^2 \cdot \exp \frac{-k^2\ell^2}{2} \right) \tag{167b}$$

$$= \mathcal{A} - \mathcal{C} \tag{167c}$$

and

$$\frac{1}{\sqrt{2\pi}} \ell \cdot \frac{\mathrm{d}\mathcal{B}}{\mathrm{d}\ell} = \sqrt{2\pi} \sum_{k=0}^{K-1} 2k\ell^2 Q(k\ell) + \frac{\sqrt{2\pi}}{8}\ell^2 - \sum_{k=0}^{K-1} \ell \cdot \left(k^2\ell^2 \exp \frac{-k^2\ell^2}{2} \right) \tag{168a}$$

$$= \sqrt{\frac{2}{\pi}}\mathcal{B} - \sum_{k=0}^{K-1} \ell \cdot \left(k^2\ell^2 \exp \frac{-k^2\ell^2}{2} \right) \tag{168b}$$

$$= \sqrt{\frac{2}{\pi}}\mathcal{B} - \mathcal{C}, \tag{168c}$$

it is direct to check that both (164) and (165) are equivalent to (52), thereby completing the proof.

Appendix E
Proof of Proposition 11

First, (80) and (81) can be obtained by the Pythagorean relations (73) and (74), respectively. We begin from (24c) and (25) and rewrite them as

$$\mathrm{E}\left[\mathsf{X}\overline{\mathsf{Y}}\right] = \mathrm{E}\left[|\mathsf{X}|^2\right] \sqrt{\frac{2}{\pi}}\mathcal{A} = \mathrm{E}\left[\mathsf{X}\overline{\mathsf{V}}\right] \sqrt{\frac{2}{\pi}}\mathcal{A} \tag{169}$$

and

$$\mathrm{E}\left[|\mathsf{Y}|^2\right] = \mathsf{E}\left[|\mathsf{V}|^2\right] \frac{2}{\pi}\mathcal{B}, \tag{170}$$

respectively. Combining them with (80) we obtain

$$\frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{Y}}\right]}{\mathrm{E}\left[|\mathsf{Y}|^2\right]} = \frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{V}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]} \sqrt{\frac{\pi}{2}} \frac{\mathcal{A}}{\mathcal{B}}, \tag{171}$$

thereby implying that (52) is equivalent to (78). According to (81), we have

$$\mathrm{E}\left[\mathsf{V}\overline{\mathsf{Y}}\right] = (\mathcal{E}_x + \sigma^2)(1 - \mathsf{mmse}) = \mathrm{E}\left[|\mathsf{Y}|^2\right]. \tag{172}$$

Substituting (172) into (171) yields

$$\frac{\mathrm{E}\left[\mathsf{X}\overline{\mathsf{Y}}\right]}{\mathrm{E}\left[|\mathsf{X}|^2\right]} = \frac{\mathrm{E}\left[\mathsf{V}\overline{\mathsf{Y}}\right]}{\mathrm{E}\left[|\mathsf{V}|^2\right]} \sqrt{\frac{\pi}{2}} \frac{\mathcal{A}}{\mathcal{B}}, \tag{173}$$

thereby implying that (52) is equivalent to (77). Combine (171) and its equivalent form (173), we obtain

$$\rho_{XY}^2 = \rho_{XV}^2 \rho_{VY}^2 \frac{\pi\mathcal{A}^2}{2\mathcal{B}^2}, \tag{174}$$

thereby implying that (52) is equivalent to (79).

REFERENCES

[1] J. Zhou, S. Pang, and W. Zhang, "A high-resolution analysis of receiver quantization in communication," *IEEE Int. Symp. Inf. Theory (ISIT)*, Ann Arbor, MI, USA, June 2025.

[2] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, Apr. 1999.

[3] B. Murmann, "ADC Performance Survey 1997-2023," [Online]. Available: https://github.com/bmurmann/ADC-survey

[4] F. J. Bloom, S. S. L. Chang, B. Harris, A. Hauptschein, and K. C. Morgan, "Improvement of binary transmission by null-zone reception," *Proc. IRE*, vol. 45, no. 7, pp. 963–975, July 1957.

[5] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York, NY, USA: Wiley, 1965.

[6] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*. New York, NY, USA: McGraw-Hill, 1979.

[7] T. Koch and A. Lapidoth, "At low SNR, asymmetric quantizers are better," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5421–5445, Sep. 2013.

[8] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Trans. Commun.*, vol. 57, no. 12, pp. 3629–3639, Dec. 2009.

[9] B. M. Murray and S. Reisenfeld, "Maximizing the cutoff rate in a quantized MIMO wireless systems with AGC," in *Proc. 1st Int. Conf. Wireless Broadband Ultra Wideband Commun. (AusWireless'06)*, Sydney, 2006.

[10] B. M. Murray and I. B. Collings, "AGC and quantization effects in a zero-forcing MIMO wireless system," in *Proc. IEEE 63rd Veh. Technol. Conf.*, 1802–1806, May 2006.

[11] J. A. Nossek and M. T. Ivrlač, "Capacity and coding for quantized MIMO systems," in *Proc. Intl. Conf. Wireless Commun. Mobile Computing (IWCMC)*, 2006, Vancouver, Canada, pp. 1387–1391.

[12] A. Mezghani, J. A. Nossek, and A. L. Swindlehurst, "Low SNR asymptotic rates of vector channels with one-bit outputs," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7615–7634, Dec. 2020.

[13] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.

[14] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, "Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate tradeoffs," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 4, pp. 2274–2287, April 2017.

[15] S. Rini, L. Barlett, E. Erkip, and Y. C. Eldar, "A general framework for MIMO receivers with low-resolution quantization," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Kaohsiung, Taiwan, Nov. 2017, pp. 599–603.

[16] A. Khalili, S. Rini, L. Barletta, E. Erkip, and Y. C. Eldar, "On MIMO channel capacity with output quantization constraints," In *Proc. 2018 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1355–1359.

[17] K. Gao, J. N. Laneman, and B. Hochwald, "Capacity of multiple one-bit transceivers in a Rayleigh environment," in *Proc. IEEE Wireless Commun. Net. Conf. (WCNC)*, May 2018.

[18] Y. Nam, H. Do, Y. Jeon, and N. Lee, "On the capacity of MISO channels with one-bit ADCs and DACs," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 9, pp.2132–2145, Sept. 2019.

[19] N. I. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Capacity bounds for one-bit MIMO Gaussian channels with analog combining," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7224–7239, Nov. 2022.

[20] A. Mezghani and J. A. Nossek, "Capacity lower bound of MIMO channels with output quantization and correlated noise," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*, July 2012, Cambridge, MA, USA, July 2012, pp. 1732-1736. Available: https://mediatum.ub.tum.de/doc/1171263/1171263.pdf (not available on IEEE Xplore.).

[21] L. Fan, S. Jin, C. K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Dec. 2015.

[22] K. Roth and J. A. Nossek, "Achievable rate and energy efficiency of hybrid and digital beamforming receivers with low resolution ADC," *IEEE J. Select. Areas Commun.*, vol. 35, no. 9, pp. 2056-2068, Sept. 2017.

[23] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. on Wirel. Commun.*, vol. 16, no. 6, pp. 4038–4051, June 2017.

[24] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal. Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[25] S. Dutta, C. N. Barati, D. Ramirez, A. Dhananjay, J. F. Buckwalter, and S. Rangan, "A case for digital beamforming at mmWave," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 2, pp. 756–769, Feb. 2020.

[26] Q.-U.-A. Nadeem and A. Chaaban, "Analysis of one-bit quantized linear precoding schemes in multi-cell massive MIMO downlink," *IEEE Trans. Commun.*, vol. 72, no. 5, pp. 2577–2594, May 2024.

[27] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Technical Report No. 216, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, Mar. 1952.

[28] Ö. T. Demir and E. Björnson, "The Bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions [lecture notes]," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 131–136, Jan. 2021.

[29] A. Lozano and S. Rangan, "Spectral vs energy efficiency in 6G: Impact of the receiver front-end," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 1, pp. 41–53, March 2023.

[30] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, Feb. 2012.

[31] N. Liang and W. Zhang, "Mixed-ADC massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 983–997, Apr. 2016.

[32] J. Scarlett, A. G. i Fabregas, A. Somekh-Baruch, and A. Martinez, "Information-theoretic foundations of mismatched decoding," *Found. Trends Commun. Inf. Theory*, vol. 17, no. 2, Mar. 2020.

[33] A. Ganti, A. Lapidoth, and İ. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[34] A. Martinez, A. G. i Fábregas, G. Caire, and F. M. J. Willems, "Bit-interleaved coded modulation revisited: A mismatched decoding perspective," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2756–2765, June 2009.

[35] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?" *IEEE Trans. Inf. Theory*, vol 48, no. 5, pp. 1118–1134, May 2002.

[36] H. Weingarten, Y. Steinberg, and S. Shamai(Shitz), "Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1665–1686, Aug. 2004.

[37] G. Kramer, "Information rates for channels with fading, side information and adaptive codewords," *Entropy*, vol. 25, no. 5, May 2023.

[38] M. Secondini and E. Forestieri, "Scope and limitations of the nonlinear Shannon limit (invited paper)," *J. Lightw. Technol.*, vol. 35, no. 4, pp. 893-902, Feb. 2017.

[39] Y. Wang and W. Zhang, "Generalized nearest neighbor decoding," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 5852–5865, Sept. 2022.

[40] B. Li, N. Liang, and W. Zhang, "On transmission model for massive MIMO under low-resolution output quantization," in *Proc. 2017 IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Sydney, NSW, Australia, Jun. 2017.

[41] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, Oct. 1998.

[42] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[43] B. M.Oliver, J. R. Pierce, and C. E. Shannon, "The philosophy of PCM," *Proc. IRE*, vol. 36, no. 11, pp. 1324–1331, Nov. 1948.
[44] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, 1948, vol. 27, no. 3, pp. 446–472.
[45] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inf. Theory*, vol. IT-14, pp. 676–683, Sept. 1968.
[46] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, March 2001.
[47] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.
[48] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[49] S. Na and D. L. Neuhoff, "Asymptotic MSE distortion of mismatched uniform scalar quantization," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3169–3171, May 2012.
[50] S. Na and D. L. Neuhoff, "On the convexity of the MSE distortion of symmetric uniform scalar quantization," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2626–2638, Apr. 2018.
[51] S. Na and D. L. Neuhoff, "Monotonicity of step sizes of MSE-optimal symmetric uniform scalar quantizers," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1782–1792, Mar. 2019.
[52] F. Sun, J. Singh, and U. Madhow, "Automatic gain control for ADC-limited communication," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Miami, FL, USA, Dec. 2010.
[53] S. Krone and G. Fettweis, "Optimal gain control for single-carrier communications with uniform quantization at the receiver," in *Proc. 2010 Int. Conf. Acoust. Speech Sig. Process. (ICASSP'10)*, Dallas, USA, Mar. 2010.
[54] D. Marco and D. L. Neuhoff, "The validity of the additive noise model for uniform scalar quantizers," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1739–1755, May 2005.
[55] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, New York: Cambridege Unviersity Press, 2005.
[56] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1520–1529, May 1996.
[57] W. Zhang, Y. Wang, C. Shen, and N. Liang, "A regression approach to certain information transmission problems," *IEEE J. Select. Areas Commun.*, vol. 37, no. 11, pp. 2517–2531, Nov. 2019.
[58] Y. Wu, L. M. Davis, and R. Calderbank, "On the capacity of the discrete-time channel with uniform output quantization," in *Proc. 2009 IEEE Int. Symp. Inf. Theory (ISIT'09)*, Seoul, Korea, 2009, pp. 2194–2198.
[59] G. D. Forney and G. Ungerboeck, "Modulation and coding for linear Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 44, no.6, pp. 2384–2415, Oct. 1998.
[60] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*. 3rd edition. New York, NY: Springer.
[61] E. N. Gilbert, "Increased information rate by oversampling," *IEEE Trans. Inf. Theory*, vol. 39, no. 6, pp. 1973–1976, Nov. 1993.
[62] S. Shamai (Shitz), "Information rates by oversampling the sign of a bandlimited process," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1230–1236, July 1994.
[63] T. Koch and A. Lapidoth, "Increased capacity per unit-cost by oversampling," in *Proc. IEEE 26th Conv. Elect. Electron. Eng. Israel*, Eilat, Israel, Nov. 2010, pp. 684–688.
[64] L. T. N. Landau, M. Dörpinghaus, and G. P. Fettweis, "1-bit quantization and oversampling at the receiver: Sequence-based communication," *EURASIP J. Wirel. Commun. Netw.*, vol. 2018, no. 1, pp. 83, Apr. 2018.
[65] R. Deng, J. Zhou, and W. Zhang, "Bandlimited communication with one-bit quantization and oversampling: Transceiver design and performance evaluation," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 845-862, Feb. 2021.
[66] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no.4, pp. 951–963, Apr. 2003.
[67] J. A. Bucklew and N. C. Gallagher, "Some properties of uniform step size quantizers," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 5, pp. 610–613, Sept. 1980.
[68] P. Borjesson and C.-E. Sundberg, "Simple approximations of the error function Q (x) for communications applications," *IEEE Trans. Commun.*, vol. 27, no. 3, pp. 639–643, Mar. 1979.
[69] I. Mezö and Á. Baricz, "On the generalization of the Lambert W function," *Trans. Amer. Math. Soc.*, vol. 369, no. 11, pp. 7917–7934, Nov. 2017.
[70] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[71] J. L. Massey, "Coding and modulation in digital communication," in *Proc. Zurich Sem. Digital Commun.*, Vol. 2, No. 1, 1974.
[72] Q. Yu and M. Médard, "The asymptotic solutions of the capacity maximal quantization problem," in *Proc. 2015 IEEE 82nd Veh. Technol. Conf. (VTC2015-Fall)*, 2015.
[73] P. F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, pp. 44–48, Jan. 1951.