RZ 3385 (# 93431) 12/10/01
Electrical Engineering 17 pages

Research Report

A Duality of Shaping over Discrete Memoryless Channels

Xiao–Yu Hu

IBM Research
Zurich Research Laboratory
8803 Rüschlikon
Switzerland
xhu@zurich.ibm.com

A Duality of Shaping over Discrete Memoryless Channels

Xiao–Yu Hu

IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland
xhu@zurich.ibm.com

Abstract

We formulate an optimization problem for designing nonuniformly spaced constellations for equiprobable M-ary coded modulation schemes. Given the average channel SNR constraint, optimum (equiprobable) signal sets are obtained in an attempt to maximize the mutual information between channel input and output. It is proved that the optimum signal set can asymptotically achieve the ultimate Shannon capacity over an AWGN channel without requiring any shaping technique. Extensive comparisons between the optimum signal set, a geometrically Gaussian-like signal set, and a uniformly spaced signal set are provided. Extensions to Rayleigh fading channels are also investigated. Rather than using a trellis code to exploit the performance gain of the optimum signal constellation, we investigate powerful low-density parity-check codes over GF(2^b) in conjunction with 2^b-ary modulation schemes. Simulation results show that such a coding scheme can exploit almost the entire performance gain promised by the information-theoretic argument.

I. INTRODUCTION

For uncoded modulation schemes, the principle of designing signal constellations (sets) for an additive white Gaussian noise (AWGN) channel is to maximize the squared Euclidean distance between signal points under average power constraint, leading to uniformly spaced signal sets. Classical coded modulations such as the set-partitioning technique of Ungerboeck [1] and the multilevel coding approach of Imai and Hirakawa [2] use a trellis code to maximize the minimum Euclidean distance between code sequences, in which the signal sets still remain uniformly spaced and equiprobable. Forney *et al.* [3] pointed out that the channel capacity predicted by Shannon for the AWGN channel cannot be achieved by equiprobable, uniformly spaced signal sets and that there exists a gap of $\pi e/6$ (≈1.53 dB) asymptotically.

One way to reduce this gap is *shaping*. In a conventional, uniformly spaced constellation, every point is used with equal likelihood. The objective of shaping is to achieve a Gaussian-like distribution over a somewhat expanded constellation, so as to reduce the average signal power while retaining the data rate. The power reduction is called shaping gain. By going to higher dimensions and choosing signal points from an N-dimensional sphere rather than an N-cube, the constellation implies a nonequiprobable distribution on the projective one-dimensional constellation. In the limit as $N \to \infty$, an N-sphere can ultimately achieve a shaping gain of 1.53 dB, and enforces a truncated Gaussian distribution on the projective one-dimensional constellation. For further details on shaping techniques and nonequiprobable signaling, we refer the reader to [4]-[8] and the references therein.

A duality of the shaping, i.e., the use of *nonequiprobable*, uniformly spaced signal sets, is to use equiprobable, *nonuniformly* spaced signal sets. It has been shown theoretically by Sun and van Tilborg [9] that at any given channel signal-to-noise ratio (SNR), an equiprobable, geometrically Gaussian-like signal set can asymptotically achieve the channel capacity of an AWGN channel. However, this optimality property holds true only for the limiting case in which the code rate tends to zero (the size of the signal set growing to infinity), and the associated problem of finding a suitable coding scheme remains intact. To our knowledge, the first work to demonstrate the advantage of a nonuniformly spaced signal set was done by Divsalar *et al.* [10], who showed that by designing an asymmetric (nonuniformly spaced) M-PSK signal constellation, a higher coding gain can be achieved in many cases than with the conventional M-PSK combined with trellis coding. Several approaches followed this line of thought, as documented by [11]-[14]. Nonuniformly spaced signal sets have also been used against signal-dependent channel impairments affecting the outer points of a signal constellation [15]. By transforming (sometimes called warping) the uniformly spaced signal constellation in such a way that points near the perimeter are spaced further apart than points near the center, a performance gain of about 0.25 dB has been achieved by PCM modems. However, these results all rely heavily on the use of a trellis code with the design criterion to maximize the minimum Euclidean distance between code sequences. Hence the fact that a nonuniformly spaced signal set, which actually decreases the minimum Euclidean distance between signal points under average power constraint, provides an advantage is somewhat perplexing. A heuristic argument [15] is that the performance is governed not only by the minimum Euclidean distance between signal sequences but also by the average number of nearest neighbors to the transmitted signal sequence. However, this explanation is neither fundamental nor convincing. Moreover, a systematic approach to design a good, nonuniformly spaced signal constellation and an appropriate coding scheme capable of exploiting its advantage has not yet been found.

This paper focuses on the optimum design of nonuniformly spaced signal constellations as well as on coding techniques to exploit potential performance improvements. We propose that the optimum equiprobable, nonuniformly spaced signal constellation should be chosen such that, under the equiprobable-input constraint and given the channel SNR, the mutual information between the channel input and output is maximized. This gives rise to a constrained, nonlinear cost-function optimization problem that can be solved numerically. It is shown numerically that the resulting optimum, nonuniformly spaced signal constellation leads to an increased information rate at a specific channel SNR compared with the conventional, uniformly spaced signal constellation. In other words, a nonnegligible performance gain in terms of SNR is expected, which we believe to be the fundamental reason for the preference of nonuniformly spaced signal constellations in many scenarios.

Instead of using a trellis code to exploit the performance gain of a nonuniformly spaced signal constellation, we consider a class of more powerful, low-density parity-check (LDPC) codes [16]-[18] together with the progressive-edge growth (PEG) construction [19]. Specifically we investigate nonbinary LDPC codes defined over $GF(2^b)$ [20] in conjunction with 2^b-ary modulation schemes, in which each transmitted sample or symbol carrying b bits corresponds to exactly one element of a codeword in the nonbinary LDPC code. Simulation results show that such a coding scheme is capable of achieving almost the entire performance gain of an equiprobable, nonuniformly spaced signal constellation promised by the information-theoretic argument.

Closely related to this work are the empirical results reported in [21],[22], in which equally likely signals with nonuniformly spaced constellations designed empirically such that the output signal approximates the Gaussian distribution, have been investigated. There, Turbo codes combined with BICM [23] were utilized to fulfill the so-called shaping gain due to nonuniform constellations.

II. An Information-Theoretical View

Consider a power- and bandwidth-efficient digital communication system based on M-ary modulation schemes such as M-PAM, M-QAM, and M-PSK, where $M = 2^b$, and b is a positive integer larger than 1. For simplicity, we shall focus on M-PAM so that only real-valued input and output are involved. Extensions to other modulation schemes clearly are straightforward. Assume that the signal constellation is chosen as the finite signal set $A = \{a_0, a_1, \ldots, a_{M-1}\}$, a_i being real-valued, over an intersymbol interference-free band-limited channel with AWGN. Note that A may not necessarily be uniformly spaced. With perfect timing and carrier-phase synchronization, we sample at time $jT + \tau_s$, where T is the modulation interval and τ_s the appropriate sampling phase. The output of the modulation channel becomes

$$y_j = x_j + w_j, \tag{1}$$

where x_j denotes a real-valued discrete channel input selected from signal constellation A that is transmitted at modulation time jT, and w_j is an independent Gaussian-distributed noise sample with zero mean and variance σ^2. The average SNR of the channel is defined as

$$\mathrm{SNR} = \frac{1}{M\sigma^2} \sum_{i=0}^{M-1} a_i^2 \ . \tag{2}$$

The mutual information between the channel input X (discrete) and output Y (continuous) can readily be written as [1]

$$I_A(X;Y) = \sum_{k=0}^{M-1} P(a_k) \int_{-\infty}^{+\infty} p(y|a_k) \log_2 \left\{ \frac{p(y|a_k)}{\sum\limits_{i=0}^{M-1} P(a_i) p(y|a_i)} \right\} dy \tag{3}$$

in bit/T, where $P(a_k)$ denotes the *a priori* probability associated with a_k, and because of the assumption of AWGN, we know that

$$p(y|a_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-a_k)^2/2\sigma^2} \ . \tag{4}$$

With the further assumption that only codes with equiprobable occurrence of channel input signals are of interest, we can substitute in (3) with $P(a_k) = 1/M$, which yields

$$\begin{aligned} I_A(X;Y) &= \sum_{k=0}^{M-1} \frac{1}{M} \int_{-\infty}^{+\infty} p(y|a_k) \log_2 \left\{ \frac{M p(y|a_k)}{\sum_{i=0}^{M-1} p(y|a_i)} \right\} dy \\ &= \log_2 M - \sum_{k=0}^{M-1} \frac{1}{M} \int_{-\infty}^{+\infty} p(y|a_k) \log_2 \left\{ \frac{\sum_{i=0}^{M-1} p(y|a_i)}{p(y|a_k)} \right\} dy \\ &= \log_2 M - \frac{1}{M} \sum_{k=0}^{M-1} E_w \left\{ \log_2 \sum_{i=0}^{M-1} \exp\left[-\frac{|a^k + w - a^i|^2 - |w|^2}{2\sigma^2} \right] \right\} \ , \end{aligned} \tag{5}$$

where E_w denotes expectation over the Gaussian-distributed noise variable w with zero mean and variance σ^2. For any given signal set A, the associated information rate, $I_A(X;Y)$, can be numerically evaluated according to (5) via the Monte Carlo method.

One of the main contributions of this paper is the design criterion for an optimum equally-like signal set over a finite discrete-input memoryless channel based on the following proposition.

Proposition 1: The channel capacity of an equiprobable M-ary discrete input memoryless channel C^*_{equ} is equivalent to the supremum channel-input and -output mutual information over all equally-like signal sets, namely,

$$\begin{aligned} C^*_{\text{equ}} &= \max_A I_A(X;Y) \\ &= I_{A^*}(X;Y) \ , \end{aligned} \tag{6}$$

where A^* denotes the optimum equally-like signal set under a specific channel SNR constraint.

Proof: A general formula for the capacity of arbitrary single-user channels without feedback has been proved in [24], asserting that the capacity is equal to the supremum over all input processes of the input-output information rate. Therefore, the channel capacity C^* for a memoryless channel with M-level discrete input constraint can be formulated as

$$\begin{aligned} C^* &= \max_X I_A(X;Y) \\ &= I_{X^*}(X;Y) \ , \end{aligned} \tag{7}$$

with X^* denoting the optimum input process that maximizes the M-level discrete input-constrained information rate $I_A(X;Y)$. The determination of the optimum input process X^* involves investigating which signal constellation is chosen and how frequently each individual signal point in the constellation is used. Given the further constraint that the channel input is equally likely, the problem of choosing the optimum input process X^* reduces to designing an optimum signal constellation A^*, thereby giving rise to (6). ■

Finding the optimum equally-like signal set over a finite discrete-input memoryless channel is in general not a trivial task. The underlying problem is a constrained nonlinear-cost-function M-dimensional optimization problem in which the M signal points in the constellation serve as optimization parameters. Formally this is a standard optimization problem: Maximize

$$I_A(X;Y)$$

over all $a_0, a_1, \ldots, a_{M-1}$ satisfying

$$\sum_i a_i^2 = \text{const} \; .$$

We can write the constrained maximization using Lagrange multipliers as the maximization of

$$J = I_A(X;Y) + \lambda \sum_i a_i^2 \; . \tag{8}$$

Differentiating with respect to a_i and setting it to zero, we obtain a necessary condition for the optimum signal set

$$\begin{aligned} \frac{\partial J}{\partial a_i} &= \frac{\partial I_A(X;Y)}{\partial a_i} + 2a_i\lambda \\ &= 0 \end{aligned} \tag{9}$$

for all $a_i \in A$. Note that

$$\lambda = -\frac{1}{2a_i} \cdot \frac{\partial I_A(X;Y)}{\partial a_i} \; . \tag{10}$$

Then the necessary condition turns out to be

$$\left\{ \begin{array}{l} \frac{1}{a_0}\frac{\partial I_A(X;Y)}{\partial a_0} = \frac{1}{a_1}\frac{\partial I_A(X;Y)}{\partial a_1} = \cdots = \frac{1}{a_{M-1}}\frac{\partial I_A(X;Y)}{\partial a_{M-1}} \\ \sum\limits_i a_i^2 = \text{const} \; . \end{array} \right. \tag{11}$$

As an analytical formula for computing optimum equiprobable signal sets using (11) is difficult to obtain, we resort to the numerical optimization methods described in [25], [26], which involve only Monte-Carlo evaluations of $I_A(X;Y)$ via (5). Some optimized signal sets will be reported later.

As is well known, various shaping techniques are capable of achieving the 1.53 dB gap asymptotically and thus of approaching the Shannon capacity over an AWGN channel. It is interesting to see whether the use of an optimum, nonuniformly spaced signal constellation will be able to achieve the Shannon capacity. This is answered by the following proposition.

Proposition 2: The channel capacity for an equiprobable M-ary input (memoryless) AWGN channel, namely $C^*_{\text{equ}} = I_{A^*}(X;Y)$, is equal to the Shannon capacity $C = 1/2\log_2(1+\text{SNR})$ as $M \to \infty$.

Proof: First we show that $C^*_{\text{equ}} \leq C$. This is trivial by recognizing that the input process, i.e., the equally likely signal set A^*, can be visualized as a subset in the set of input process of an AWGN channel.

To show $C^*_{\text{equ}} \geq C$ as $M \to \infty$, we consider a specific signal set, namely, the geometrically Gaussian-like signal set A^{geo} defined as follows: Take $M+1$ points on the real line

$$-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_{M-1} < \alpha_M = \infty \tag{12}$$

such that for $0 \leq i \leq M-1$

$$\frac{1}{\sqrt{2\pi P}} \int_{\alpha_i}^{\alpha_{i+1}} e^{-x^2/2P} dx = \frac{1}{M} \; . \tag{13}$$

Then the i-th element a_i^{geo} of A^{geo} is defined by

$$a_i^{\text{geo}} = \frac{M}{\sqrt{2\pi P}} \int_{\alpha_i}^{\alpha_{i+1}} x e^{-x^2/2P} dx. \tag{14}$$

Put simply, a_i^{geo} is the centroid of the equiprobable intervals (α_i, α_{i+1}) with respect to a Gaussian distribution of variance P, P denotes the average signal power. This signal set may not necessarily be optimum for any finite alphabet size M, but it is asymptotically optimum over an AWGN channel in the sense that the mutual information $I_{A^{\text{geo}}}(X;Y) = C = \frac{1}{2}\log_2(1 + \frac{P}{\sigma^2})$ as $M \rightarrow \infty$. The asymptotic optimum property of the Gaussian-like signal set has been proved in [9, see also Appendix A]. Then it follows

$$\begin{aligned} C_{\text{equ}}^{*} &= I_{A^*}(X;Y) \\ &= \max_{A} I_A(X;Y) \\ &\geq I_{A^{\text{geo}}}(X;Y) \\ &= C \qquad \text{as } M \rightarrow \infty. \end{aligned} \tag{15}$$

■

Remarks: Proposition 2 implies that the Shannon capacity of a continuous-input and -output AWGN channel can be attained asymptotically by an optimum equiprobable signal constellation without requiring any shaping techniques. That is to say, the shaping gain arising from the use of (uniformly-spaced) nonequiprobable signal sets can be alternatively obtained by its duality, i.e., by the use of nonuniformly-spaced (equiprobable) signal sets. As the input to the channel is in general equally likely, the gain promised by nonuniformly-spaced equiprobable signal sets incurs no additional complexity, whereas the shaping gain usually requires additional coding/decoding complexity to realize nonequiprobable use of signal points in the constellation.

Note that Proposition 1 is widely applicable to all finite discrete-input memoryless channels and not necessarily limited to the AWGN channel. For instance, the same philosophy extends to the frequency-nonselective Rayleigh slow-fading channel, which can be modeled by

$$y_j = \theta_j * x_j + w_j \; , \tag{16}$$

where θ is the normalized Rayleigh fading factor with $E[\theta^2] = 1$ and density function $p(\theta) = 2\theta \exp(-\theta^2)$, $\theta \geq 0$. Assume the perfect channel state information (CSI) is available at the receiver, i.e., the receiver knows the exact value of θ_k at any time instant, then the mutual information I_A^{CSI} becomes

$$\begin{aligned} I_A^{\text{CSI}}(X;Y) &= I_A(X;Y,\theta) \\ &= I_A(X;\theta) + I_A(X;Y|\theta) \quad \text{(chain rule)} \\ &= I_A(X;Y|\theta) \quad \text{(because } X \text{ and } \theta \text{ are independent)} \\ &= \log_2 M - E_{X,Y,\theta}\left[\log_2 \frac{\sum_{a_i:i=0}^{M-1} p_\theta(y|a_i)}{p_\theta(y|x)} \right] \quad \text{(assuming equiprobable input)} \; , \end{aligned} \tag{17}$$

which, in principle, can be evaluated numerically; then an optimum signal constellation at a specific SNR maximizing the mutual information over a Rayleigh fading channel can be obtained by solving the underlying nonlinear optimization problem.

Often the computational effort of solving the optimization problem is dominated by the cost of evaluating the cost function $I_A(X,Y)$. To evaluate (17) efficiently, we can translate (16) into another form in the presence of CSI, namely,

$$\tilde{y}_j = x_j + v_j \ , \tag{18}$$

where v_j is an *instant* Gaussian-distributed noise sample w_j/θ_j with zero mean and variance σ^2/θ_j^2. Following a similar procedure as in (5), we obtain

$$\begin{aligned} I_A^{\mathrm{CSI}}(X;Y) &= I_A(X;Y|\theta) \\ &= \log_2 M - \frac{1}{M}\sum_{k=0}^{M-1} E_v\left\{\log_2 \sum_{i=0}^{M-1} \exp\left[-\frac{|a^k+v-a^i|^2-|v|^2}{2\sigma^2/\theta_j^2}\right]\right\} \ , \end{aligned} \tag{19}$$

where E_v denotes expectation over the random variable v that has a Cauchy-like probability density function given by [see Appendix B]

$$p(v) = \frac{\sigma^2}{(2\sigma^2+v^2)^{3/2}} \ . \tag{20}$$

Equation (19) can be much more easily evaluated by Monte Carlo method than (17) as it involves only a one-dimensional integral. To evaluate $I_A^{\mathrm{CSI}}(X;Y)$ using the Monte-Carlo method, the random samples of v should be generated in two steps: first the Rayleigh distribution for θ_j, and then a Gaussian distribution with zero mean and variance σ^2/θ_j^2.

It can be verified that the information rate for a given signal constellation in (5) as well as in (19), depends solely on the channel SNR. Consequently, the optimum signal constellation is also a function of the SNR, and independent of the specific signal power or noise variance values.

III. Numerical Results

In this section we compute the capacities and optimum signal constellations with equiprobable input at specific channel SNRs for the two typical channels we discussed above: AWGN channels and Rayleigh fading channels with perfect knowledge of CSI.

A. AWGN Channels

Whereas the geometrically Gaussian-like signal set A^{geo} has been known to be asymptotically optimum for an AWGN channel as M tends to infinity, it is clear that A^{geo} is not the optimum signal set for maximizing the mutual information with equiprobable input, as A^{geo} is associated with the signal power whereas the optimum signal set depends on the channel SNR.

Because the conventional uniformly spaced signal constellation is defined as $\pm1, \pm3, \cdots, \pm(M-1)$, whose average power equals $(M^2-1)/3$, we henceforth consider A^{geo} with the same average power constraint,[1] i.e., $P = (M^2-1)/3$. Table I shows the resulting geometrically Gaussian-like signal sets for 4-PAM, 8-PAM, and 16-PAM, obtained by solving Eqs. (13) and (14).

Fig. 1 depicts information rates of equiprobable-input AWGN channels with conventional uniformly spaced signal constellations and with geometrically Gaussian-like signal sets in Table I under 4-PAM, 8-PAM, and 16-PAM. One can see that at low SNRs the geometrically Gaussian-like signal sets clearly

[1] Unless explicitly stated, all M-ary signal sets reported in this paper have a normalized power constraint of $(M^2-1)/3$.

TABLE I

THE GEOMETRICALLY GAUSSIAN-LIKE SIGNAL SETS WITH POWER $(M^2-1)/3$. NOTE THAT FOR 16-PAM, ONLY THE POSITIVE PART OF THE SIGNAL SET IS SHOWN; THE NEGATIVE PART CAN BE OBTAINED BY SYMMETRY.

	$a_0^{\rm geo}$	$a_1^{\rm geo}$	$a_2^{\rm geo}$	$a_3^{\rm geo}$	$a_4^{\rm geo}$	$a_5^{\rm geo}$	$a_6^{\rm geo}$	$a_7^{\rm geo}$
$A_{4-\rm PAM}^{\rm geo}$	0.782578	3.06391	−0.782578	−3.06391	—	—	—	—
$A_{8-\rm PAM}^{\rm geo}$	0.74469	2.3162	4.22081	7.76308	−0.74469	−2.3162	−4.22081	−7.76308
$A_{16-\rm PAM}^{\rm geo}$	0.731853	2.21401	3.75499	5.40744	7.25275	9.44391	12.3625	18.3467

outperform their uniformly spaced counterparts, closely approaching the Shannon limit. For instance, under 16-PAM, the Gaussian-like signal set achieves an information rate of 2 bit/T at 11.944 dB, which is only 0.183 dB away from the Shannon capacity, whereas for the uniformly spaced signal set the needed SNR is 12.527 dB, which is 0.583 dB worse than the Gaussian-like signal set and 0.766 dB away from the Shannon capacity. Moreover, as the channel SNR tends to be even lower, the difference between the information rate of the Gaussian-like signal set and the Shannon capacity is negligible; this also holds true for fixed channel SNR but increased alphabet size M.

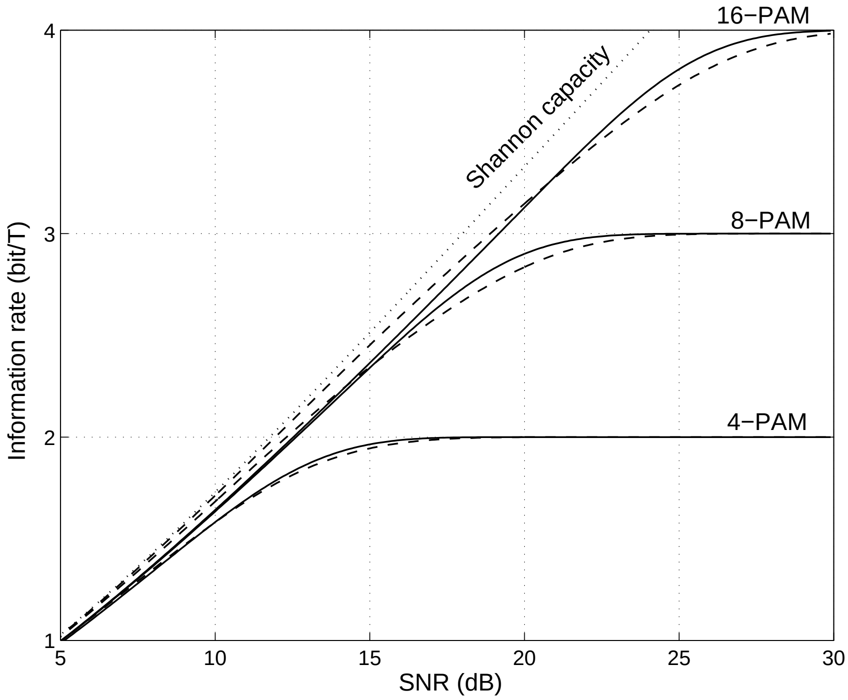

Fig. 1. Information rates of the equiprobable-input AWGN channel with uniformly spaced signal constellations (solid lines) and with geometrically Gaussian-like signal sets (dashed lines). The Shannon capacity limit (dotted line) of the continuous input–output AWGN channel is also plotted.

From Fig. 1 we can also observe that, at high SNRs, the geometrically Gaussian-like signal sets perform worse than their uniformly spaced counterparts. Specifically, at a SNR of 25 dB and for 16-PAM, the uniformly spaced constellation achieves an information rate of 3.808 bit/T whereas the Gaussian-like one achieves only 3.731 bit/T. This suggests that the Gaussian-like signal set is advantageous only at relatively low SNRs, and that at high SNRs, a signal set optimized for the specific SNR should be utilized instead.

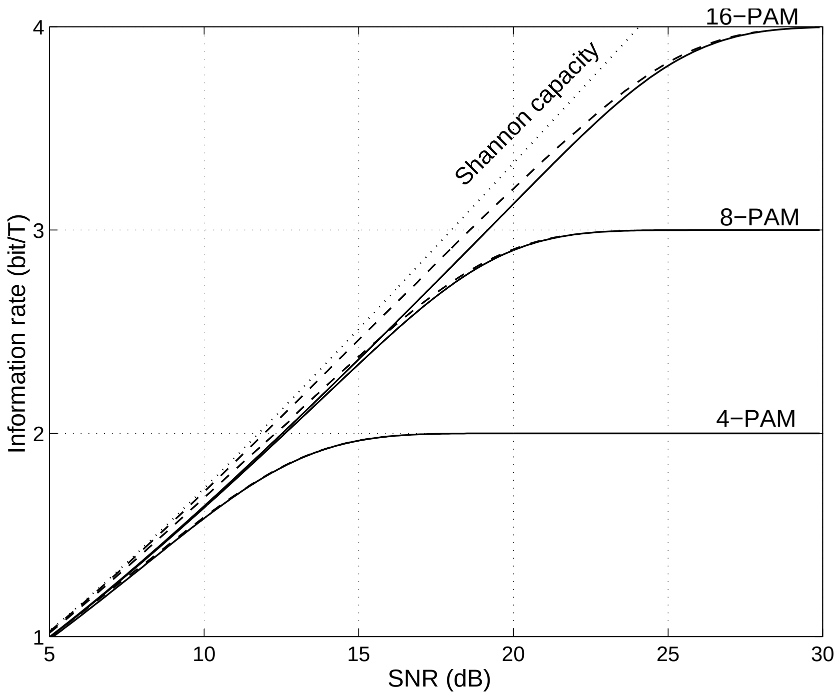

Fig. 2. Information rates of the equiprobable input AWGN channel with uniformly spaced signal constellations (solid lines) and with nonuniformly signal sets (dashed lines) optimized for individual channel SNRs.

Fig. 2 shows information rates of equiprobable-input AWGN channels with conventional uniformly spaced signal constellations and with nonuniformly spaced signal sets optimized at individual SNRs under 4-PAM, 8-PAM, and 16-PAM: in the very low SNR region, the optimized signal sets exhibit essentially the same performance as the geometrically Gaussian-like does, and in the very high SNR region the optimized signal sets converge to the uniformly spaced signal sets. Unlike the geometrically Gaussian-like signal sets, the optimum signal sets consistently outperform the uniformly spaced ones at any SNR.

We tabulate two optimized signal sets for 16-PAM over AWGN channels in Table II: the first is optimized at an SNR of 19.0 dB, the second at 12.5 dB. When solving the optimization problem of finding the optimum signal set, a symmetrical condition is imposed on the signal set, i.e., the signal points in the negative part are defined to be the signal points in the positive part multiplied by -1. In this way, the M-dimensional optimization problem reduces to a $M/2$-dimensional one. Interestingly enough, it is found empirically that this constraint induces no noticeable performance loss, leading to the conjecture that the optimum signal set holds the symmetry property over an AWGN channel.

TABLE II

TWO 16-PAM SIGNAL SETS OPTIMIZED FOR THE EQUIPROBABLE-INPUT AWGN CHANNEL AT AN SNR OF 19.0 DB AND 12.5 DB. THE NEGATIVE PART CAN BE OBTAINED BY SYMMETRY.

	a_0^*	a_1^*	a_2^*	a_3^*	a_4^*	a_5^*	a_6^*	a_7^*
$A^*_{19\text{ dB}}$	0.781318	2.3634	4.01108	5.78051	7.77234	10.1671	13.1507	16.9582
$A^*_{12.5\text{ dB}}$	0.675268	2.64975	3.26062	5.90869	6.96619	9.96857	12.3329	18.0825

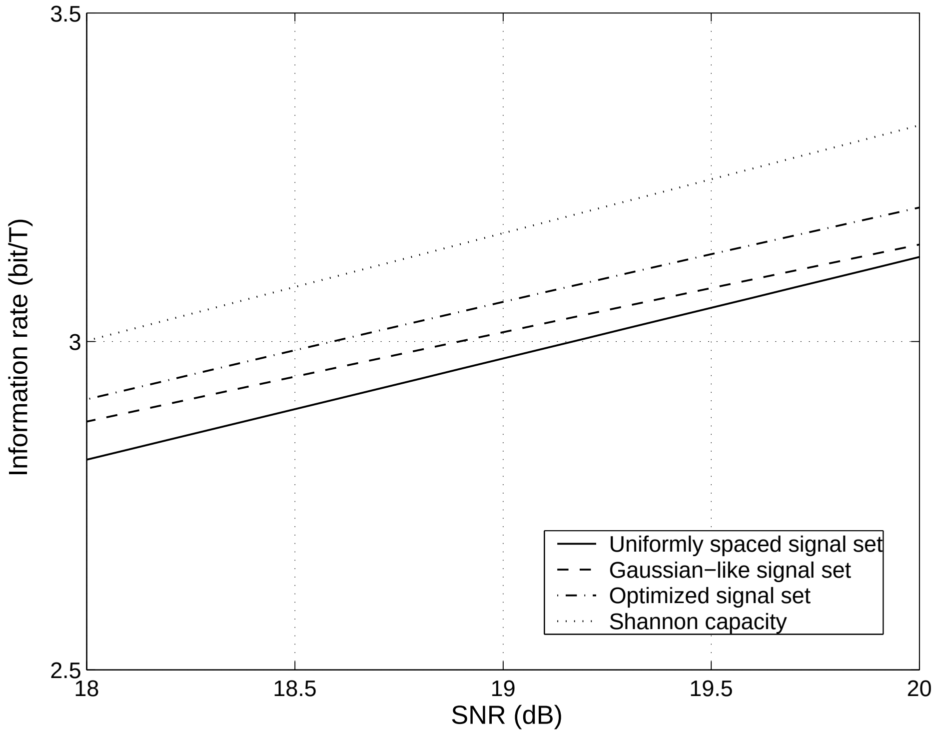

Fig. 3. Information rates of the equiprobable-input AWGN channel with uniformly spaced signal constellation, geometrically Gaussian-like signal set, and optimized signal set $A^*_{19\ \mathrm{dB}}$ under 16-PAM.

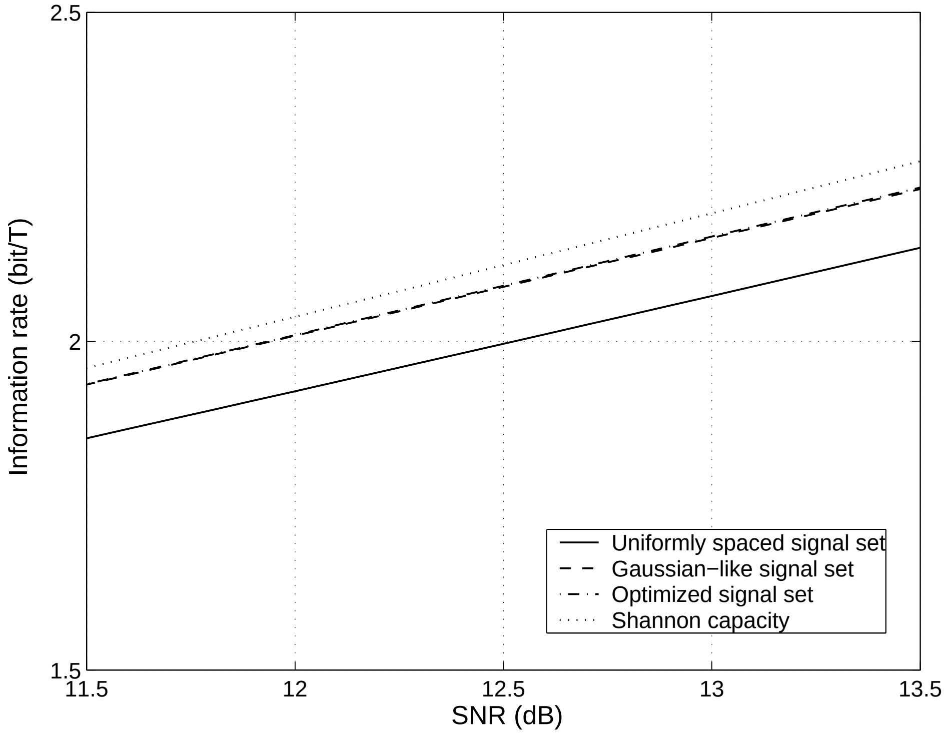

Fig. 4. Information rates of the equiprobable-input AWGN channel with uniformly spaced signal constellation, geometrically Gaussian-like signal set, and optimized signal set $A^*_{12.5\ \mathrm{dB}}$ under 16-PAM.

Figs. 3 and 4 show performance comparisons in terms of achievable information rates for three equiprobable signal sets: the uniformly spaced, the geometrically Gaussian-like, and the optimized one, in the SNR range of 19 dB and 12.5 dB, respectively. At an information rate of 3 bit/T, the optimized signal set $A^*_{19\ \mathrm{dB}}$ outperforms the geometrically Gaussian-like signal set by 0.304 dB and the uniformly spaced signal constellation by 0.577 dB. The optimized signal set $A^*_{12.5\ \mathrm{dB}}$ exhibits essentially the same performance as the geometrically Gaussian-like signal set in the SNR range of 12.5 dB, and both outperform the uniformly spaced signal constellation by approx. 0.58 dB.

With higher-order modulation schemes, the performance gain of the optimum signal set over the uniformly spaced one can be even larger. Table III summarizes empirical results of such performance gains when the alphabet size M is increasing up to 256. This observation is in good agreement with Proposition 2, asserting that the 1.53 dB gap between equiprobable and Gaussian input can be recovered by the use of an equiprobable but optimum signal constellation.

TABLE III

PERFORMANCE GAIN OF THE OPTIMUM SIGNAL SET.

Modulation	Transmit rate	Performance gain
32-PAM	2.5 bit/T	0.82 dB
64-PAM	3.0 bit/T	1.01 dB
128-PAM	3.5 bit/T	1.15 dB
256-PAM	4.0 bit/T	1.26 dB

B. Rayleigh Fading Channels

Here we report optimized signal constellations for Rayleigh fading channels with ideal CSI. We summarize two optimized signal sets for 16-PAM in Table IV: the first is optimized at an SNR of 22 dB, the second at 15 dB. Again, a symmetrical constraint is imposed on the signal constellation during the optimization procedure. Figs. 5 and 6 show the performance comparison in terms of achievable information rates under the equiprobable-input constraint between the uniformly spaced signal constellation and the optimized one in the SNR range of 22 dB and 15 dB, respectively. We observe that the optimized signal set $A^*_{22\ \mathrm{dB}}$ outperforms the uniformly spaced signal constellation by 0.38 dB at an information rate of 3 bit/T, and $A^*_{15\ \mathrm{dB}}$ outperforms the uniformly spaced signal constellation by 0.5 dB at an information rate of 2 bit/T.

TABLE IV

TWO 16-PAM SIGNAL SETS OPTIMIZED FOR THE EQUIPROBABLE- INPUT RAYLEIGH FADING CHANNEL WITH CSI AT AN SNR OF 22 DB, AND 15 DB. THE NEGATIVE PART CAN BE OBTAINED BY SYMMETRY.

	a^*_0	a^*_1	a^*_2	a^*_3	a^*_4	a^*_5	a^*_6	a^*_7
$A^*_{22\ \mathrm{dB}}$	0.870211	2.57815	4.31533	6.1972	8.16027	10.3523	12.9902	16.5246
$A^*_{15\ \mathrm{dB}}$	0.909399	2.07285	3.84487	5.55386	7.42087	9.86724	13.0089	17.5381

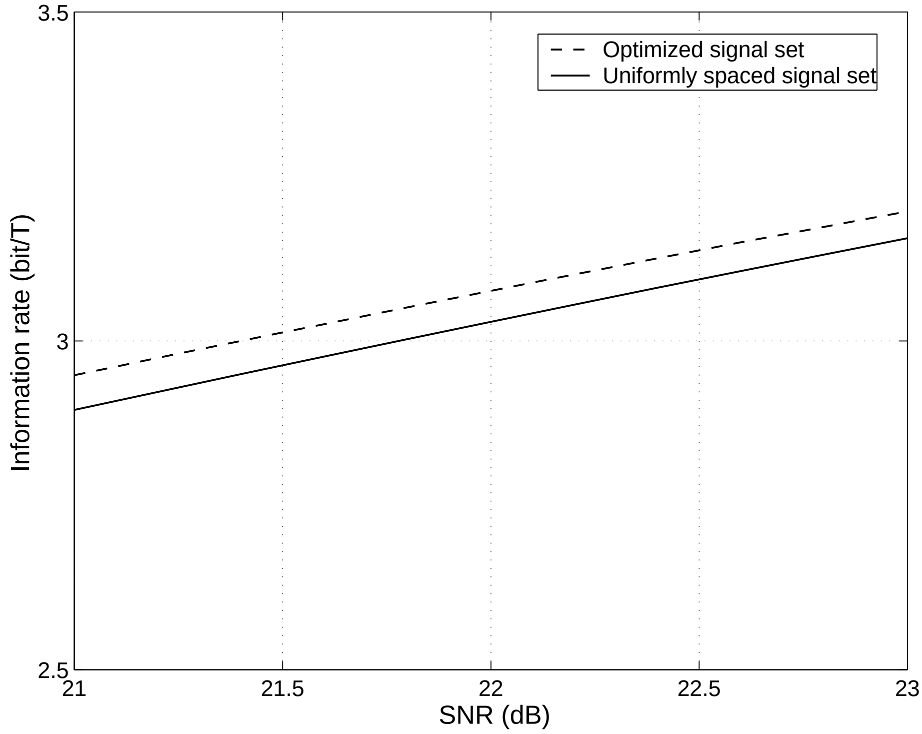

Fig. 5. Information rates of the equiprobable-input Rayleigh fading channel (CSI) with the uniformly spaced signal constellation and optimized signal set $A^*_{22\ \mathrm{dB}}$ under 16-PAM.

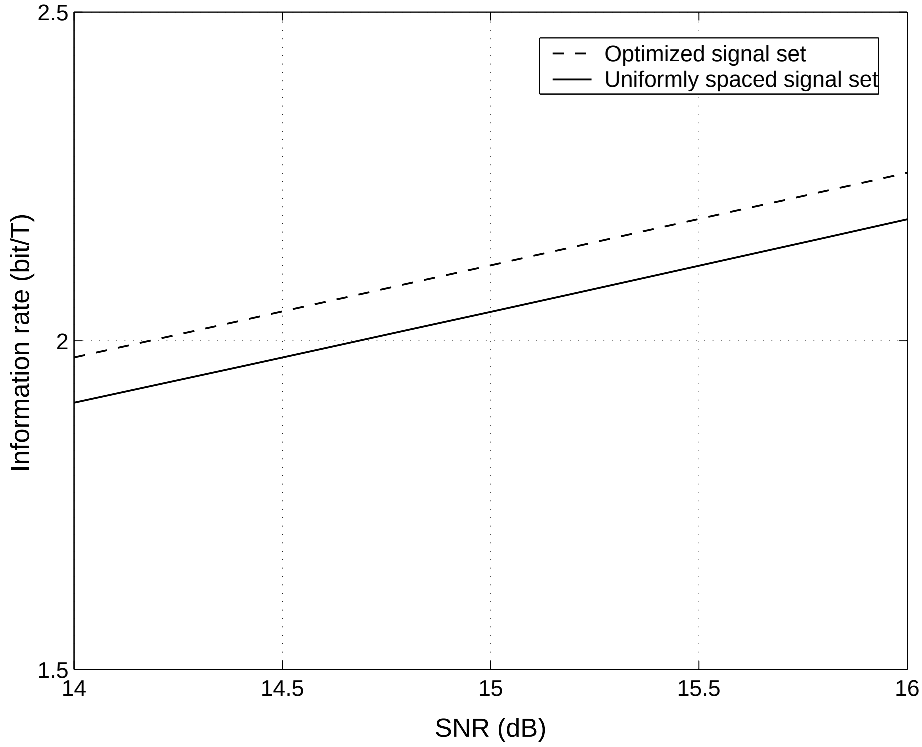

Fig. 6. Information rates of the equiprobable-input Rayleigh fading channel (CSI) with the uniformly spaced signal constellation and optimized signal set $A^*_{15\ \mathrm{dB}}$ under 16-PAM.

IV. CODING AND DECODING ISSUES

As mentioned, classical coded modulations using the set-partitioning concept or the multilevel-coding approach use a trellis code to maximize the minimum Euclidean distance between code sequences. Clearly, the optimum signal set, which in general is nonuniformly spaced, is not in harmony with this design philosophy because the minimum Euclidean distance between signal points has been weakened. To fully exploit the performance gain promised by the information-theoretic arguments in Section III, we need more powerful coding schemes.

We consider a class of error-correcting codes, the so-called LDPC codes, discovered by Gallager [16] in 1962. LDPC codes are defined in terms of a sparse parity-check matrix and are known to be asymptotically good for all channels with symmetric stationary ergodic noise [17]. To comply with 2^b-ary modulation, we consider the generalization of the conventional binary LDPC codes to finite fields GF(2^b) such that each point in the signal set corresponds to an element in GF(2^b). Such a code is defined in terms of a LDPC matrix H with m rows and n columns. The nonzero positions in H are established by the progressive edge-growth (PEG) algorithm [19] to maximize the girth of the underlying bipartite graph and thus facilitate iterative decoding. After the nonzero positions of H have been determined, we fill the nonzero entries in H with the nonzero elements from the finite field GF(2^b) according to a uniform probability distribution. Assume the resulting matrix H is full rank, we can encode every $n-m$ information symbols (each symbol carries b bits) into a codeword X that meets $HX = 0$. Note that each element of X is an element of the finite field GF(2^b), and is mapped to some point in the signal constellation prior to transmission. To minimize the bit-error performance, a Gray mapper is adopted in our experiments.

We transmit codeword X, which is received as $Y = X + W$, where W is a noise vector sampled from the underlying channel noise distribution. An instance of the decoding problem requires finding the most probable vector $\hat{X}$, given Y, which can be approximately solved by the sum-product algorithm (SPA). The SPA can be viewed as a message-passing algorithm on a bipartite graph defined by the parity-check matrix H, which contains n symbol and m check nodes. Each symbol node corresponds to a noisy sample received, each check node represents a check equation that should be fulfilled by its associated symbol nodes. Let edges $e_{i,j}$ connect check node i with noisy symbol node j. For each edge $e_{i,j}$ in the graph, the quantities $q^a_{i,j}$ and $r^a_{i,j}$ are iteratively updated, in which $a \in$ GF(2^b) and a also corresponds to some point in the signal constellation. The quantity $q^a_{i,j}$ denotes the probability that the j-th symbol of $\hat{X}$ is a, given the information obtained via connected check nodes other than check node i. The quantity $r^a_{i,j}$ represents the probability of check i being satisfied if the j-th symbol of $\hat{X}$ is considered fixed at a, given the information stemming from connected symbol nodes other than symbol node j. The complexity of decoding scales as $nt2^{2b}$ per iteration, with t denoting the average number of edges incident to a symbol node. For detailed description of the decoding algorithm for LDPC codes over GF(2^b), we refer the reader to [20].

We are particularly interested in whether the performance gain of an optimum signal constellation promised by the information-theoretic argument can be realized by LDPC codes over GF(2^b). We consider two LDPC codes. Both codes are rate-3/4 defined over GF(2^4), and constructed using the PEG algorithm in [19]. Code 1 has a block length of 4096 symbols (in binary, of length 4096x4), and it is a regular LDPC code with symbol-node degree 3. Code 2 has a larger block length, namely 20,000 symbols, which is an irregular LDPC code with symbol-node degree distribution $\lambda(x) = 0.643772x^2 + 0.149719^3 + 0.193001x^4 + 0.013508^5$. In constructing both codes, the degree sequence of check node has been made as uniform as possible. The codes are iteratively decoded with the sum-product algorithm up to 80 iterations. Note that the overall data transmission rate is 3 bit/T, and thus matched to an AWGN channel of approx. 19.0 dB, at which $A^*_{19\text{ dB}}$ is the optimum signal constellation. Fig. 7 compares the performance of the optimized signal constellation $A^*_{19\text{ dB}}$, the geometrically Gaussian-like signal set $A^{\text{geo}}_{16-\text{PAM}}$, and the uniformly spaced signal set over an AWGN channel. It reveals that a performance gain of approx. 0.5 dB is achieved at a bit error rate (BER) of 10^{-5} by using the optimum nonuniformly spaced constellation instead of using the uniformly spaced one. Recall that in this case a gain of 0.577 dB is expected by the information-theoretic argument in Section III-A. Interestingly enough, the performance gain predicted by the information-theoretic argument has been almost fully exploited by a very simple modulation scheme combined with LDPC codes over GF(2^b).

We also apply the same LDPC codes to a slowing-fading Rayleigh channel with perfect knowledge of CSI at the receiver. Fig. 8 compares the performance of the optimized signal constellation $A^*_{22\text{ dB}}$ with that of the uniformly spaced one using the two LDPC codes. Again a performance gain of nearly 0.35 dB is observed at a BER of 10^{-5} when using the optimum signal set. This gain is in fairly good agreement with the information-theoretic claim in Section III-B.

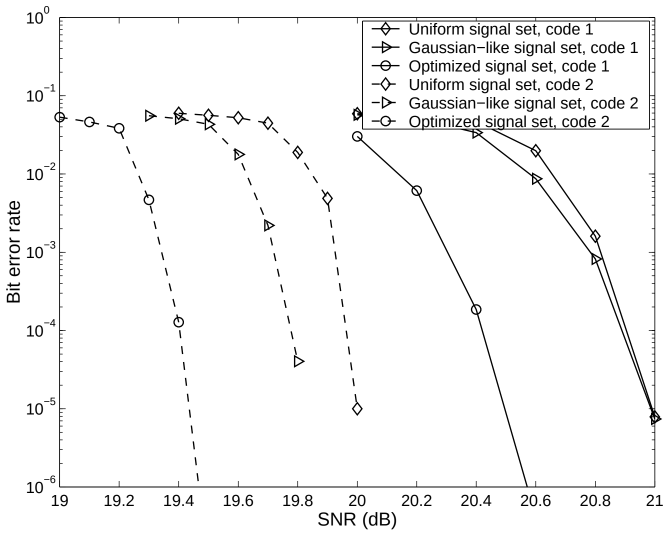

Fig. 7. Performance comparison of the optimized signal set ($A^*_{19\text{ dB}}$), the Gaussian-like signal set $A^{\text{geo}}_{16-\text{PAM}}$, and the uniformly spaced signal set using 16-PAM and GF(2^4) LDPC codes over an AWGN channel.

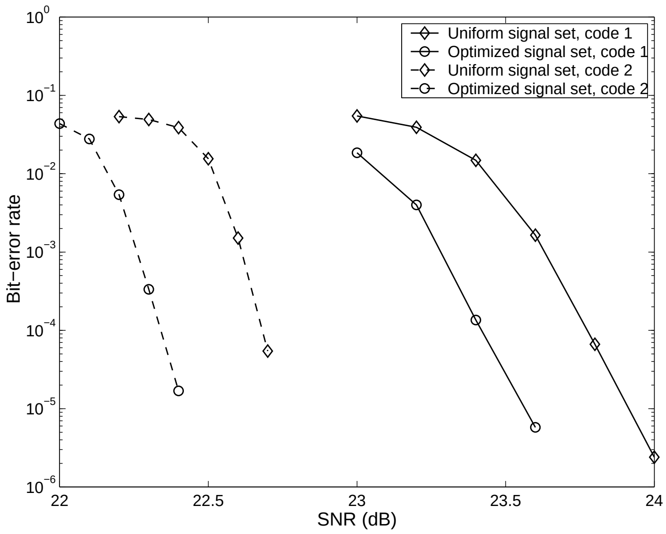

Fig. 8. Performance comparison between the optimized signal set ($A^*_{22\text{ dB}}$) and the uniformly spaced signal set using 16-PAM and GF(2^4) LDPC codes over a Rayleigh fading channel with CSI.

V. Conclusions

We have presented an optimization problem to design nonuniformly spaced constellations for equiprobable M-ary coded modulation schemes. Given the channel SNR constraint, optimum (equiprobable) signal sets are obtained in an attempt to maximize the mutual information between channel input and output. It has been shown that the optimum signal set can achieve the ultimate Shannon capacity over an AWGN channel asymptotically without any shaping technique. Extensive comparisons between an optimum signal set, a geometrically Gaussian-like signal set, and a uniformly spaced signal set have been provided. Extensions to Rayleigh fading channels have also been investigated.

Rather than using a trellis code to exploit the performance gain of a nonuniformly spaced signal constellation, we investigated the more powerful low-density parity-check codes over GF(2^b) in conjunction with 2^b-ary modulation schemes. Simulation results showed that such a coding scheme can faithfully exploit almost the entire performance gain promised by the information-theoretic argument.

To emphasize the advantage of the class of optimum, nonuniformly spaced constellations, we have deliberately neglected two important issues, namely, the peak-to-average ratio (PAR) and the constellation expand ratio (CER), both of which are of practical interest. Note that an optimum, nonuniformly spaced signal set often leads to a larger PAR, which for practical applications is undesirable. One possible remedy might be to consider the requirements of PAR and/or CER during the optimization procedure by incorporating them as additional constraints. Another potential approach to reduce PAR is by going to multidimensional constellations, which, of course, is an interesting topic for further research.

Appendix A

For self-completeness, we summarize the proof for which the equiprobable signaling specified by (13) and (14) satisfies $I_{A^{\rm geo}} = C = \frac{1}{2}\log_2(1+\frac{P}{\sigma^2})$ asymptotically in the sense of M, which was first proved in [9].

First we establish an upper bound on the average energy of $A^{\rm geo}$,

$$\begin{aligned} \frac{1}{M}\sum_{i=0}^{M-1}(a_i^{\rm geo})^2 &= \frac{1}{M}\sum_{i=0}^{M-1}\left[\frac{M}{\sqrt{2\pi P}}\int_{\alpha_i}^{\alpha_{i+1}} x e^{-x^2/2P}dx\right]^2 \\ &= M\sum_{i=0}^{M-1}[E(xI_i(x))]^2 \ , \end{aligned} \tag{21}$$

where $E(\cdot)$ is the expectation with respect to the probability density function $\frac{1}{\sqrt{2\pi P}}e^{-x^2/2P}$, and $I_i(x)$ is the indicator function

$$I_i(x) = \begin{cases} 1 & \text{if } x \in (\alpha_i, \alpha_{i+1}) \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Continuing with (21),

$$M\sum_{i=0}^{M-1}[E(xI_i(x))]^2 \le M\sum_{i=0}^{M-1} E[xI_i(x)]^2 E(I_i^2(x)) = E(x^2) = P \ . \tag{23}$$

Let X_M be the random variable uniformly distributed on $A^{\rm geo}$. Because the capacity-achieving output distribution is unique, it suffices to show that the density function of $X_M + W$ converges to that of $X + W$, where X and W are Gaussian distributions with variances of P and σ^2, respectively. The density function of $X_M + W$ is

$$\begin{aligned} p_M(x) &= \frac{1}{M}\sum_{i=0}^{M-1}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-a_i^{\rm geo})^2/2\sigma^2} \\ &= \frac{1}{2\pi\sigma\sqrt{P}}\sum_{i=0}^{M-1}\int_{\alpha_i}^{\alpha_{i+1}} e^{-(x-a_i^{\rm geo})^2/2\sigma^2-y^2/2P}dy \end{aligned} \tag{24}$$

and the density function of $X + W$ is

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi(P+\sigma^2)}}e^{-x^2/2(\sigma^2+P)} \\ &= \frac{1}{2\pi\sigma\sqrt{P}}\int_{-\infty}^{\infty} e^{-(x-y)^2/2\sigma^2-y^2/2P}dy\ . \end{aligned} \tag{25}$$

Now consider

$$\begin{aligned} |p_M(x)-p(x)| &\leq \frac{1}{2\pi\sigma\sqrt{P}}\sum_{i=0}^{M-1}\int_{\alpha_i}^{\alpha_{i+1}} e^{-y^2/2P}|e^{-(x-y)^2/2\sigma^2}-e^{-(x-a_i^{\rm geo})^2/2\sigma^2}|dy \\ &\leq \frac{1}{2\pi\sigma\sqrt{P}}\sum_{i;x\leq\alpha_i}\int_{\alpha_i}^{\alpha_{i+1}} e^{-y^2/2P}|e^{-(x-y)^2/2\sigma^2}-e^{-(x-a_i^{\rm geo})^2/2\sigma^2}|dy + \\ &\quad \frac{1}{2\pi\sigma\sqrt{P}}\sum_{i;x\geq\alpha_{i+1}}\int_{\alpha_i}^{\alpha_{i+1}} e^{-y^2/2P}|e^{-(x-y)^2/2\sigma^2}-e^{-(x-a_i^{\rm geo})^2/2\sigma^2}|dy + \frac{2}{\sqrt{2\pi}\sigma M} \\ &\leq \frac{1}{2\pi\sigma\sqrt{P}}\sum_{i;x\leq\alpha_i}\int_{\alpha_i}^{\alpha_{i+1}} e^{-y^2/2P}|e^{-(x-\alpha_i)^2/2\sigma^2}-e^{-(x-\alpha_{i+1})^2/2\sigma^2}|dy + \\ &\quad \frac{1}{2\pi\sigma\sqrt{P}}\sum_{i;x\geq\alpha_{i+1}}\int_{\alpha_i}^{\alpha_{i+1}} e^{-y^2/2P}|e^{-(x-\alpha_i)^2/2\sigma^2}-e^{-(x-\alpha_{i+1})^2/2\sigma^2}|dy + \frac{2}{\sqrt{2\pi}\sigma M} \\ &= \frac{1}{\sqrt{2\pi}\sigma M}\sum_{i;x\leq\alpha_i}\{e^{-(x-\alpha_i)^2}-e^{-(x-\alpha_{i+1})^2}\} + \frac{1}{\sqrt{2\pi}\sigma M}\sum_{i;x\geq\alpha_{i+1}}\{e^{-(x-\alpha_{i+1})^2}-e^{-(x-\alpha_i)^2}\} \\ &\quad +\frac{2}{\sqrt{2\pi}\sigma M} \\ &\leq \frac{4}{\sqrt{2\pi}\sigma M}\ . \end{aligned} \tag{26}$$

Thus $p_M(x)$ converges to $p(x)$ as M goes to infinity.

APPENDIX B

Suppose a random variable V is defined by

$$V = W/\Theta \ ,$$

where W is Gaussian distributed with zero mean and variance σ^2, namely,

$$p_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-w^2/2\sigma^2} \ ,$$

and Θ is normalized, Rayleigh-distributed:

$$p_\Theta(\theta) = 2\theta e^{-\theta^2} \ .$$

To evaluate the probability density function $p_V(v)$, we first calculate the probability that $V \leq v$:

$$\begin{aligned} \Pr(V \leq v) &= \Pr(W/\Theta \leq v) \\ &= \int_0^\infty \Pr(W \leq \theta v) p_\Theta(\theta) d\theta \\ &= \int_0^\infty \left[\int_{-\infty}^{\theta v} p_W(w) dw \right] p_\Theta(\theta) d\theta \\ &= \int_0^\infty \left[\int_{-\infty}^{\theta v} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-w^2/2\sigma^2} dw \right] 2\theta e^{-\theta^2} d\theta \ . \end{aligned} \tag{27}$$

Differentiating $\Pr(V \leq v)$ with respect to v gives rise to $p_V(v)$:

$$\begin{aligned} p_V(v) &= \frac{d[\Pr(V \leq v)]}{dv} \\ &= \int_0^\infty \frac{d}{dv} \left[\int_{-\infty}^{\theta v} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-w^2/2\sigma^2} dw \right] \cdot 2\theta e^{-\theta^2} d\theta \\ &= \frac{2}{\sqrt{2\pi}\sigma} \int_0^\infty \theta^2 e^{-\theta^2 - \frac{\theta^2 v^2}{2\sigma^2}} d\theta \\ &= \frac{1}{\sigma(2 + \frac{v^2}{\sigma^2})^{3/2}} \\ &= \frac{\sigma^2}{(2\sigma^2 + v^2)^{3/2}} \ . \end{aligned} \tag{28}$$

REFERENCES

[1] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. 28, pp. 55-67, Jan. 1982.

[2] H. Imai and S. Hirakawa, "A new multilevel coding method using error-correcting codes," *IEEE Trans. Inform. Theory*, vol. 23, pp. 371-377, May 1977.

[3] G. D. Forney, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff, and S. U. Qureshi, "Efficient modulation for band limited channels," *IEEE J. Select. Areas Commun.*, vol. 2, pp. 632-647, Sept. 1984.

[4] G. D. Forney, Jr., and L.-F. Wei, "Multidimensional constellations — Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 877-892, Aug. 1898.

[5] G. D. Forney, Jr., "Multidimensional constellations — Part II: Coronoi constellations," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 941-958, Aug. 1989.

[6] A. R. Calderbank and L. H. Ozarow, "Non-equiprobable signaling on the Gaussian channel," *IEEE Trans. Inform. Theory*, vol. 36, pp. 726-740, July 1990.

[7] G. D. Forney, Jr., "Trellis shaping," *IEEE Trans. Inform. Theory*, vol. 38, pp. 281-300, Mar. 1992.

[8] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 39, pp. 913-929, May 1993.

[9] Feng-Wen Sun and H. C. A. van Tilborg, "Approaching capacity by equiprobable signaling on the Gaussian channel," *IEEE Trans. Inform. Theory,* vol. 39, pp. 1714-1716, Sept. 1993.

[10] D. Divsalar, M. K. Simon, and J. H. Yuen, "Trellis coding with asymmetric modulations," *IEEE Trans. Commun.*, vol. 35, pp. 135-141, Feb. 1987.

[11] A. Chalid, I. Sasase, K. Yashima, and S. Mori, "Coded nonuniform phase/frequency modulation," *Proc. Int'l. Commun. Conf. (ICC'88)*, pp. 707-711, IEEE, 1988.

[12] T. Itoh, I. Sasse, and S. Mori, "Coded continuous phase modulation combined with nonuniform mapping and multi-H scheme," *Proc. Global Commun. Conf. (GLOBECOM'88)*, pp. 865-869, IEEE, 1988.

[13] S. C. Elmurr, "Multi-dimensional phase/frequency coded modulation with nonuniform signal constellation," *Proc. Southeastoncon'90*, pp. 1048-1053, IEEE, 1990.

[14] N. Seshadri and C.-E. W. Sundberg, "Coded modulation with time diversity, unequal error protection, and low delay for the Rayleigh fading channel," *Proc. Int'l Conf. Universal Personal Commun. (ICUPC'92)*, pp. 283-287, IEEE, 1992.

[15] W. Betts, A. R. Calderbank, and R. Laroia, "Performance of nonuniform constellations on the Gaussian channel," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1633-1638, Sept. 1994.

[16] R. G. Gallager, *Low Density Parity Check Codes*, MIT Press, Cambridge, MA, 1963.

[17] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, vol. 45, pp. 399-431, Mar. 1999.

[18] Special issue on codes and graphs and iterative algorithms, *IEEE Trans. Inform. Theory*, vol. 47, pp. 493-849, Feb. 2001.

[19] X.-Y. Hu, E. Eleftheriou, and D.-M. Arnold, "Progressive edge-growth (PEG) Tanner graphs," to be presented at *IEEE Global Commun. Conf. (GLOBECOM 2001): The Evolving Global Communications Network*, San Antonio, TX, Nov. 2001.

[20] M. C. Davey and D. MacKay, "Low-density parity-check codes over GF(q)," *IEEE Commun. Lett.*, vol. 2, no. 6, pp. 165-167, June 1998.

[21] D. Sommer and G. P. Fettweis, "Signal shaping by non-uniform QAM for AWGN channels and applications using Turbo coding," *Proc. 3rd ITG Conf. on Source and Channel Coding*, Munich, Germany, pp. 81-86, 2000.

[22] C. Fragouli, R. D. Wesel, D. Sommer, and G. P. Fettweis, "Turbo codes with non-uniform constellations," *Proc. Int'l. Commun. Conf. (ICC'01)*, pp. 70-73, 2001.

[23] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, vol. 44, pp. 927-945, May 1998.

[24] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147-1157, July 1994.

[25] S. L. S. Jacoby, J. S.Kowalik, J. T. Pizzo, and W. T. Veterling, *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[26] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Veterling, *Numerical Recipes in C*, Cambridge Univ. Press, UK, 1988.