Probabilistic Shaping for Thermal Noise-Limited Direct-Detection Channels With Peak and Average Power Constraints

Ethan M. Liang, *Member, IEEE*, Shannon Fan, and Joseph M. Kahn, *Life Fellow, IEEE*

***Abstract*—We study probabilistic shaping for thermal noise-limited direct-detection channels under both peak and average power constraints. We use the high-rate continuous approximation (HCA) to derive induced marginal densities and shaping gains under varying peak power constraints. The HCA analysis finds that the exact induced marginal density is well approximated by a truncated beta distribution or a truncated exponential distribution for sufficiently long shaping block length N. We derive closed-form expressions for the shaping gain for finite N and finite peak-to-average (power) ratio (PAR) and show that shaping gains within 0.1 dB of the ultimate shaping gain can be achieved with $N = 100$ or a PAR as low as 4. We study shaping using finite, discrete constellations and find that a scaled geometric distribution obtained by sampling a truncated exponential distribution achieves a shaping gain within 0.22 dB of a numerically optimized distribution for modulation orders $M = 4, 8,$ and 16 and varying constraints on the PAR. We also consider shaping for peak- and average-power-constrained direct-detection channels with limited extinction ratios. We find that the induced marginal density is still well approximated by a truncated beta distribution or a truncated exponential distribution and derive closed-form expressions for the shaping gain.**

***Index Terms*—Direct detection, fiber optic communication, probabilistic shaping.**

I. Introduction

DRAMATIC growth in cloud computing and Internet traffic places increasing demands on the throughput and density of intra-data-center interconnects and free-space optical links [1], [2]. Satisfying these demands requires improving the optical and electrical energy efficiency of these systems. These systems typically use standard direct detection, in which information is encoded by modulating the transmitted optical intensity and recovered by detecting the received intensity. We use the term "standard direct detection" (SDD) to distinguish it from other direct detection-based methods, which can encode and recover information using the absolute or differential phase in the transmitted electric field [3].

SDD links are subject to constraints on average and peak transmitted optical power. Average transmitted power constraints arise because of nonlinear effects in optical fibers [4], [5] and because of eye safety requirements [6], [7], [8]. Peak power constraints arise from limited laser power, limited power-handling capabilities of modulators and fibers, and eye safety requirements [7], [8].

Probabilistic shaping (PS) offers one approach to improve the optical energy efficiency of optical links. While PS has been widely employed in commercial long-haul coherent links [9], it is not yet commonly used in short-reach SDD links [10]. Several key challenges have impeded the adoption of PS in SDD links. SDD links are deployed in massive volumes, placing strict cost and complexity constraints on these systems [1]. PS requires an expansion of the peak-to-average (power) ratio (PAR), complicating its implementation in these peak power-limited systems. PS also requires an expansion of the constituent constellation relative to a system without PS to achieve the same net bit rate [11]. In addition, many SDD systems are subject to strict latency requirements, reducing the applicability of PS, which can require encoding over hundreds or thousands of symbols [12].

There has been considerable previous theoretical work on the capacity and capacity-achieving input distribution for the thermal noise-limited SDD channel. Asymptotically tight upper and lower bounds were derived for the SDD channel capacity under average power constraints only [13], [14] and under both peak and average power constraints [15], [16], [17]. The capacity-achieving input distribution for the SDD channel under peak and average power constraints was proven to be finite and discrete [18]. An exhaustive search algorithm to compute the optimal discrete input distribution was presented in [7].

There has also been considerable previous work on designing low-complexity discrete input distributions for the thermal noise-limited SDD channel with only an average power constraint, only a peak power constraint, or both peak and average power constraints. The Maxwell-Boltzmann (MB) distribution was shown to outperform the geometric distribution under constraints on the average detected electrical power when using low-order pulse amplitude modulation (PAM) constellations [19]. Shaping for the peak power-limited SDD channel was studied

Received 19 February 2025; revised 21 July 2025; accepted 7 September 2025. Date of publication 16 September 2025; date of current version 17 November 2025. This work was supported in part by Meta Under the Stanford Photonics Research Center, in part by Stanford System X Alliance, and in part by the National Science Foundation under Grant DGE-1656518. *(Corresponding author: Shannon Fan.)*

The authors are with the E. L. Ginzton Laboratory, Department of Electrical Engineering, Stanford University Ginzton Laboratory, Stanford, CA 94305 USA (e-mail: emliang@stanford.edu; sgfan@stanford.edu; jmk@ee.stanford.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JLT.2025.3610894.

Digital Object Identifier 10.1109/JLT.2025.3610894

for the case of a modulation order $M = 4$ [20]. A review paper studying PS for the thermal noise-limited SDD channel using the MB distribution can be found in [21]. Most recent works on shaping for the thermal noise-limited SDD channel explicitly use the MB distribution or only consider peak power constraints [22], [23], [24], [25], [26], [27].

We highlight two particularly notable and relevant contributions toward shaping for the thermal noise-limited peak- and average-power-constrained (PAPC) SDD channel. One identifies the exponential distribution as the maximum-entropy distribution under these constraints and finds that the performance of the numerically optimized input distribution can be approached using a scaled geometric distribution with a properly chosen modulation order [28]. The other uses the high-rate continuous approximation (HCA) to study shaping for the PAPC SDD channel, derives the optimal bounding region for this channel, and presents closed-form expressions for the volume and shaping gain in the high-PAR regime [29].

In this paper, we use HCA [30] to study PS for thermal noise-limited PAPC SDD links. HCA was first applied to thermal noise-limited SDD channels under average optical power constraints, which to the authors' knowledge was the first work on shaping for the SDD channel [31]. Our analysis extends the results originally presented in [29] to arbitrary PAR. We find that the induced marginal density is closely approximated by a truncated beta distribution or truncated exponential distribution in the transmitted intensity. We derive closed-form expressions for the shaping gain for finite PS symbol encoding length (or number of dimensions) N and PAR. Our HCA analysis shows that shaping gains within 0.1 dB of the ultimate shaping gain can be achieved while constraining the PAR and block length N to not exceed 4 and 100, respectively. We also study shaping for finite, discrete constellations, showing that a scaled geometric distribution obtained by sampling a truncated exponential distribution achieves an average shaping gain within 0.22 dB of that achieved by the numerically optimized input distribution.

The remainder of this paper is organized as follows: Section II introduces the thermal noise-limited SDD channel and defines the average power and peak power constraints. Section III presents the HCA analysis for the thermal noise-limited PAPC SDD channel. Section IV presents shaping results for finite, discrete input distributions. Section V concludes the paper. Appendix A establishes that Jordan measurability of the shaping region is a sufficient condition for ensuring the validity of the shaping gain expressions obtained using HCA. Appendix B studies HCA for the thermal noise-limited PAPC SDD channel under finite extinction-ratio constraints.

II. Preliminaries

In SDD, information is modulated onto and detected from the intensity of the optical electric field. In the thermal noise-limited regime, the dominant noise is signal-independent additive white Gaussian noise (AWGN) that is added to the detected electrical current [3]. The thermal noise-limited SDD channel is described by

$$Y = I + N, \quad I \geq 0. \tag{1}$$

The detected electrical current Y is proportional to the sum of the received signal intensity I and AWGN $N \sim \mathcal{N}(0, \sigma_{\text{th}}^2)$, which is independent of the transmitted intensity I. We parameterize the average optical power by a thermal noise-limited signal-to-noise ratio (SNR), $\text{SNR}^{(\text{th})}$, which is defined as

$$\text{SNR}^{(\text{th})} = (\mathbb{E}[I])^2 / \sigma_{\text{th}}^2. \tag{2}$$

Let $\bar{P}$ and A be upper bounds on the average and peak optical powers, respectively. The average optical power constraint is equivalent to a constraint on the expected value of I and is expressed as by $\mathbb{E}[I] \leq \bar{P}$. The peak optical power constraint is given by $\sup\{\mathcal{I}\} \leq A$, where $\mathcal{I}$ denotes the support of I. Note that the nonnegativity constraint on I requires that every element in $\mathcal{I}$ be nonnegative.

We note that the discrete-time channel model in (1) is tacitly derived from an underlying continuous-time channel using a specified set of basis functions. Owing to the non-existence of nonnegative minimum-bandwidth root-Nyquist pulses [29], the design of the basis function must strike a balance between bandwidth efficiency, transmission rate, and intersymbol interference. In practice, a common choice for the basis functions are nonoverlapping rectangular pulses, which are not band-limited but are free from intersymbol interference. The discrete-time channel model in (1) assumes any residual intersymbol interference is negligible.

A. High-Rate Continuous Approximation

HCA is an approach to PS that approximates the input constellation by a dense high-dimensional coset code [30]. We refer the reader to [3] for a comprehensive tutorial on the subject. In this paper, we provide only a very brief overview of the essential concepts of HCA.

A coset code is constructed by intersecting a translate of a lattice with a high-dimensional bounding region. HCA presumes that sufficiently dense high-dimensional coset codes can be well-approximated by continua of uniform density over bounding regions [3], [30], [32]. HCA transforms the problem of computing the optimal input distribution to designing bounding regions with certain specified properties in a set of natural coordinates for the channel. Some examples of specified properties include minimum average power or minimum peak power. The induced marginal density is obtained by marginalizing the uniform density over the bounding region onto one or several dimensions.

The thermal noise-limited SDD channel under average optical power constraints was first studied in [31]. The natural coordinates for the thermal noise-limited SDD channel are electric field intensities. An *optimal* bounding region minimizes the average power of the interior points subject to a constraint on the volume and is given by an N-simplex with maximum power P. The resulting induced marginal density using the minimum average power bounding region is an exponential distribution in the transmitted intensity.

A *reference* bounding region is used to compute the shaping gain compared to uniform signaling [31]. The reference bounding region is defined as a hypercube in the natural coordinates with maximum L_∞-norm A. However, the reference bounding

region can also be viewed as a bounding region that minimizes the peak power subject to constraints on the bounding region volume.

In HCA, the shaping gain is defined as the reduction factor of the average optical power of the optimal bounding region relative to the reference bounding region [3]. In Section III, we derive analytical expressions for the shaping gain as functions of N and PAR. The analytical expressions yield estimates of the shaping gain in the limit of dense constellations and high SNR. In Section IV, we present shaping gains using numerical simulations and finite constellations. For the numerical simulations, the shaping gain (in dB) at a target mutual information (MI) is half the difference in $\mathrm{SNR}^{(\mathrm{th})}$ (in dB) required by a uniform input distribution and an optimized input distribution. This factor of $1/2$ is required because shaping gains are stated in terms of average optical power.

The shaping gain expressions obtained using HCA in Section III require that sufficiently dense coset codes obtained using an optimized bounding region and a reference bounding region have approximately equal average error probabilities [30]. In [32], it was shown that this property holds when the optimized bounding region is spherical. In Appendix A, we show that for channels with only AWGN, the average error probabilities for sufficiently dense coset codes obtained by the intersection of a fixed coset with Jordan-measurable bounding regions are approximately equal. All bounding regions considered in this paper are Jordan-measurable, ensuring the validity of our shaping gain expressions.

III. Shaping With Peak and Average Power Constraints

A bounding region can be constructed to minimize the average optical power of the contained signal points subject to an additional peak power constraint. In particular, the intersection of the minimum average optical power bounding region and the minimum peak optical power bounding region encloses the set of N-D signal points with minimum average optical power subject to a peak power constraint. The PAPC bounding region is optimal in the sense that no other bounding region of the same volume can have a lower average power while also satisfying the peak symbol power constraint A. This construction was first proposed for the thermal noise-limited SDD channel in [29, Sec. IV], where closed-form expressions for the volume, average power, and shaping gain are provided in the regime in which $A \in [P/2, P]$.

In this section, we generalize the results from [29]. Specifically, we compute closed-form expressions for the volume of the PAPC bounding region for arbitrary P and A. We subsequently use the resulting volume expression to derive the induced marginal density and shaping gain for arbitrary N and PAR. We separately consider shaping for PAPC SDD channel with an additional finite extinction ratio constraint in Appendix B.

1) Bounding Region: Let $\mathbf{I} = [I_1, I_2, \ldots, I_N]^T \in \mathbb{R}_+^N$ be a real nonnegative N-D vector of signal intensities in N disjoint symbol intervals. The total optical power of $\mathbf{I}$ is $\|\mathbf{I}\|$ and the optical power of the ith subsymbol is I_i. Let P be the an upper bound on the total optical power and A be the peak power constraint on all subsymbols. The PAPC bounding region with constraints A and P is given by

$$\mathcal{R}(N, P, A) = A \cdot [0,1]^N \cap P \cdot \mathbb{S}_N, \tag{3}$$

where $[0,1]^N$ is the hypercube with unit side length and $\mathbb{S}_N = \{(x_1, \ldots, x_n) \in \mathbb{R}_+^N : x_1 + \cdots + x_N \leq 1\}$ is the N-simplex. The bounding regions $A \cdot [0,1]^N$ and $P \cdot \mathbb{S}_N$ are the reference and optimal bounding regions for the average-power-constrained SDD channel [31], respectively. The corresponding volumes for the reference and optimal bounding regions are

$$\mathcal{V}\left(P \cdot \mathbb{S}_N\right) = \frac{P^N}{N!} \tag{4}$$

and

$$\mathcal{V}\left(A \cdot [0,1]^N\right) = A^N, \tag{5}$$

respectively. $\mathcal{R}(N, P, A)$ can be interpreted as the set of points that are common to both $A \cdot [0,1]^N$ and $P \cdot \mathbb{S}_N$.

2) Volume Derivation: The volume of the PAPC bounding region can be derived from the Irwin-Hall distribution [33]. Let $\{X_1, X_2, \ldots, X_N\}$ be independent uniform random variables with support over $[0, A]$ and let $Z = X_1 + \cdots + X_N$. The CDF of Z is an Irwin-Hall distribution and is given by

$$\Pr[Z \leq P] = \sum_{k=0}^{N} (-1)^k \binom{N}{k} \left(\frac{P}{A} - k\right)_+^N \tag{6}$$

$$= \frac{P^N/N!}{A^N} \sum_{k=0}^{N} (-1)^k \binom{N}{k} \left(1 - k\frac{A}{P}\right)_+^N,$$

$$P \in [0, NA]. \tag{7}$$

The fraction of the volume of the hypercube $[0, A]^N$ with L_1-norm at most P is exactly equal to $\Pr[Z \leq P]$. We scale $\Pr[Z \leq P]$ by A^N to achieve the correct volume, yielding

$$\mathcal{V}(N, P, A) = \frac{P^N}{N!} \kappa(N, A/P), \quad A \in [P/N, P], \tag{8}$$

where

$$\kappa(N, A/P) = \left(\sum_{k=0}^{N} (-1)^k \binom{N}{k} \left(1 - k\frac{A}{P}\right)_+^N \right). \tag{9}$$

Note from (4) that $P^N/N!$ is the volume of $P \cdot \mathbb{S}_N$, indicating that $\kappa(N, A/P)$ can be interpreted as the factor of volume decrease due to the peak power constraint. An alternative derivation directly computing the volume from the intersection between an N-cube and N-simplex is given in [34].

3) Induced Marginal Density: To compute the induced marginal density, we first introduce the following alternative parameterization for the ratio A/P:

$$\frac{A}{P} = \frac{\alpha}{N}, \quad \alpha \in [1, N]. \tag{10}$$

The restriction $\alpha \in [1, N]$ ensures that $\mathcal{R}(N, P, A)$ is a hypercube or N-simplex in the limits when $\alpha = 1$ or $\alpha = N$, respectively. Examining the effect of (10) on (9), we see that

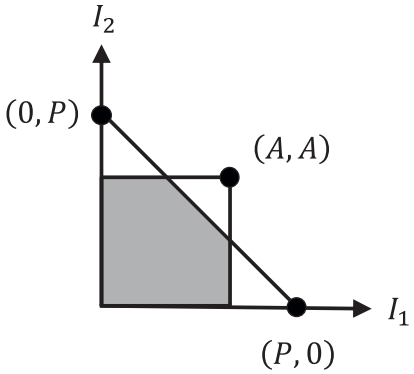

Fig. 1. Optimal bounding region for the thermal noise-limited PAPC SDD channel with constraints P and A and $N = 2$. $\mathcal{R}(N, P, A)$ from (3) is depicted by the shaded region.

α controls the number of nonzero terms in the summation. In Section III-3 below, we show that α is closely related to the PAR.

The induced marginal density can be computed by marginalizing over $\mathcal{R}(N, A, P)$ on the basic dimension I. The induced marginal density on I is proportional to

$$\begin{aligned} \mathrm{f}_I(i; N, P, A) &\propto \mathcal{V}(N-1, P-i, A) \\ &\propto \frac{\mathcal{V}(N-1, P-i, A)}{P^{N-1}/(N-1)!} \\ &= \left(1 - \frac{i}{P}\right)^{N-1} \kappa\left(N-1, \frac{A}{P-i}\right), \\ & i \in [0, A]. \end{aligned} \tag{11}$$

Using (10), we can remove the explicit dependence of I on P. Substituting (10) into (11), we get

$$f_I(i; N, A, \alpha) \propto \left(1 - i\frac{\alpha}{NA}\right)^{N-1} \kappa\left(N-1, \frac{\alpha}{N - i\alpha/A}\right) \tag{12}$$

$$\approx \exp\left(-\alpha\frac{i}{A}\right) \kappa\left(N-1, \frac{\alpha}{N - i\alpha/A}\right), \quad i \in [0, A], \tag{13}$$

where the approximation holds for large N. (12) is a product of two factors. The first factor $(1 - i\frac{\alpha}{NA})^{N-1}$ is proportional to a truncated beta distribution. The second factor $\kappa(N-1, \frac{\alpha}{N - i/A})$ is a correction factor that modifies the truncated beta distribution. The first factor in (13) is proportional to a truncated exponential distribution.

Despite the additional correction factor, (12) is well approximated by the truncated beta distribution and truncated exponential distribution for sufficiently high N. Fig. 2 shows the exact marginal density vs. $i/\bar{P}$ for (a) $N = 10$ and (b) $N = 50$ for varying PAR constraints, where $\bar{P}$ is the average optical power of the exact induced marginal distribution. The dashed lines and dotted lines indicate, respectively, the probability densities for truncated beta distribution and exponential distribution with the same PAR as the identically colored exact marginal density.

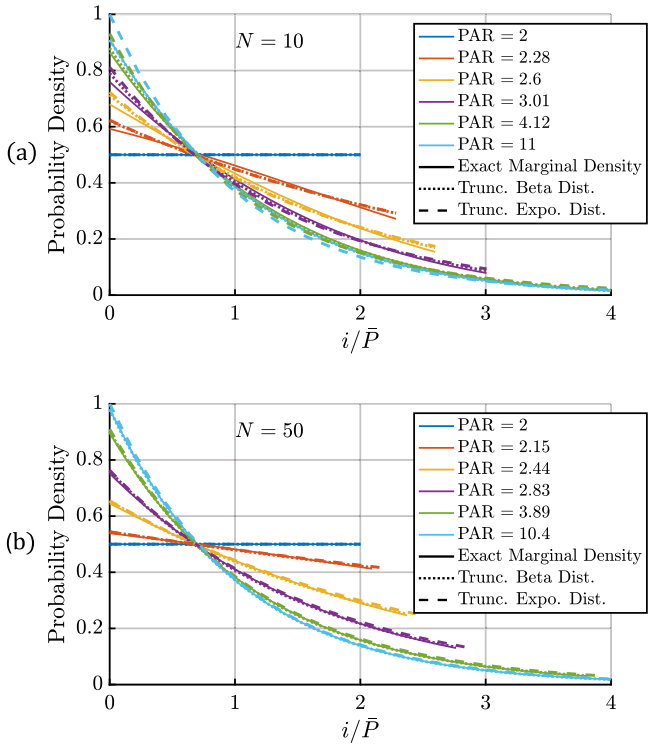

Fig. 2. The exact induced marginal density for the optimal bounding region vs. $i/\bar{P}$ for (a) $N = 10$ and (b) $N = 50$ for varying PAR constraints. The PAR can be varied by adjusting the parameter α. The dotted and dashed lines are, respectively, probability density functions for a truncated (trunc.) beta distribution (dist.) and a truncated exponential distribution with the same PAR as the identically colored exact induced marginal density.

In Fig. 2(a), there is a visible difference between the exact marginal distribution and both the truncated beta and exponential distributions for all PAR > 2. However, the truncated beta distribution provides a closer fit than the truncated exponential distribution for all PAR > 2. Fig. 2(b) shows the truncated beta and truncated exponential distributions provide closer fits to the exact marginal density for $N = 50$ than at $N = 10$, and the closeness of the fits improves as N increases. Again, the truncated beta distribution provides a better fit than the truncated exponential distribution for all PAR > 2.

4) Average Power Computation: Note that (12) contains two instances of i/A and has support over $[0, A]$. It can be useful to transform the induced marginal density so that it has support over [0,1]. Let $U = I/A$. The induced marginal density in terms of U is

$$\begin{aligned} f_U(u; N, \alpha) &\propto A\left(1 - u\frac{\alpha}{N}\right)^{N-1} \kappa\left(N-1, \frac{\alpha}{N - \alpha u}\right) \\ &\propto \left(1 - u\frac{\alpha}{N}\right)^{N-1} \kappa\left(N-1, \frac{\alpha}{N - \alpha u}\right) \end{aligned} \tag{14}$$

$$\approx \exp(-\alpha u)\, \kappa\left(N-1, \frac{\alpha}{N - \alpha u}\right), \; u \in [0, 1]. \tag{15}$$

The approximation in (15) holds for large N. Note that after the linear transformation from I to U, the PDF of U is only

a function of N and α and has support over [0,1]. Therefore, in contrast with the marginal density in (12), the normalized marginal density in (14) is not a function of P or A directly but depends on P or A only through α.

We can use (14) to compute the average power of the induced marginal density. Let $\bar{P}(N, A, \alpha)$ be the average power of the induced 1-D distribution. Note that the average power of I is related to U by $\mathbb{E}[I] = A \cdot \mathbb{E}[U]$, yielding

$$\bar{P}\,(N, A, \alpha) = A \cdot \mathbb{E}[U]. \tag{16}$$

$\mathbb{E}[U]$ can be computed numerically and only depends on N and α. It follows directly from (16) that the PAR of the exact marginal distribution is given by $1/\mathbb{E}[U]$ and is therefore only a function of N and α.

We use the simple analytic form of the truncated beta distribution to obtain an approximation to the average optical power of the exact induced marginal distribution in (16). Let $X \sim \text{Beta}(1, N)$ be a Beta distribution whose support has been modified from [0,1] to $[0, P]$. Suppose X_1 is obtained by truncating X to have support over $[0, A]$. Substituting in $P = \frac{AN}{\alpha}$, the PDF of X_1 is

$$f_{X_1}(x; N, A, \alpha) = \frac{\alpha}{A} \frac{\left(1 - x\frac{\alpha}{AN}\right)^{N-1}}{1 - \left(1 - \frac{\alpha}{N}\right)^{N}}, \; x \in [0, A]. \tag{17}$$

The average optical power of X_1 is

$$\bar{P}(X_1) = \mathbb{E}[X_1] = \frac{A}{\alpha} \frac{N}{N+1} \frac{1 - (1+\alpha)\left(1 - \frac{\alpha}{N}\right)^{N}}{1 - \left(1 - \frac{\alpha}{N}\right)^{N}}. \tag{18}$$

Notice from (18) that the PAR of X_1, $\text{PAR}(X_1) = A/\bar{P}(X_1)$, only depends on N and α. Similarly, recall from (16) that the PAR of the exact marginal distribution is also only a function of N and A. This enables us to study the PAR of both the exact marginal distribution and truncated beta distribution only as functions of N and α.

Fig. 3 shows (a) $\bar{P}/A$ vs. α and (b) PAR vs. α for the exact marginal density from (12) and for the truncated beta distribution from (17) for varying values of $\alpha \in [1, 10]$ and $N \in \{10, 25, 50\}$. The exact power is computed using (16) and the power of the truncated beta distribution is computed using (18). A reference dashed line showing $\alpha = \text{PAR}$ is also included. For values of $\alpha < 2$, PAR$\to 2$ as N increases. For values of $\alpha > 2$, PAR$\to \alpha$ as N increases. We can see from the numerical simulation that for large N,

$$\lim_{N\to\infty} \text{PAR}(N, \alpha) \approx \max\{2, \alpha\}, \; \alpha \geq 1. \tag{19}$$

In Fig. 2, the truncated beta distribution and the truncated exponential distributions provide relatively close fits to the exact marginal distribution for all PARs studied. In Fig. 3, however, the PAR for the exact marginal distribution differs from the truncated beta distribution for low values of α. This seeming contradiction is resolved by noting the exact marginal and truncated beta distributions are constructed to achieve the same PAR in Fig. 2. In Fig. 3, by contrast, the exact marginal and truncated beta distributions are constructed according to the same geometric parameter α, which may not necessarily result in the same PAR.

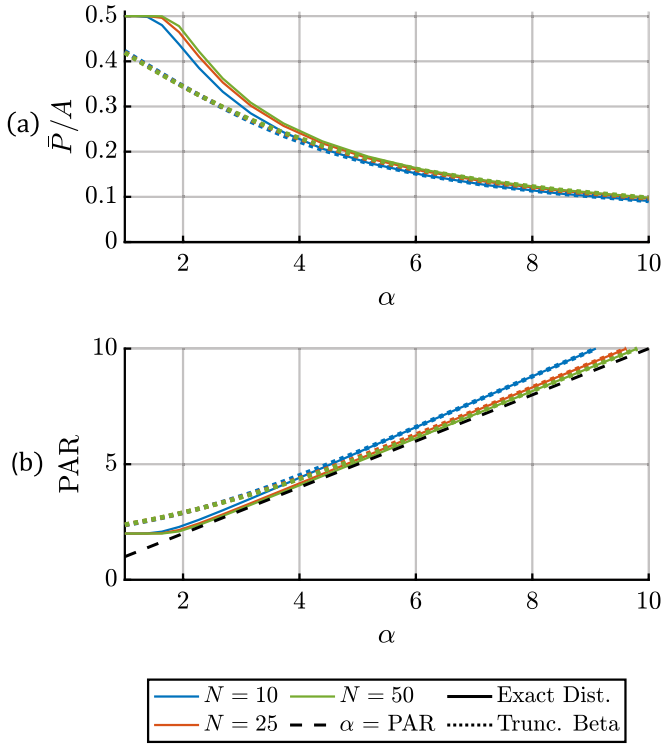

Fig. 3. $\bar{P}/A$ (a) and peak-to-average (power) ratio (PAR) (b) vs. α for various values of N. The ordinates of (a) and (b) are inverses of each other. The exact marginal distribution (dist.) and truncated (trunc.) beta distribution are specified by the solid and dotted lines, respectively. The dashed line depicts PAR$=\alpha$ and is provided as a reference.

For values of α less than approximately 4, the truncated beta distribution has a noticeably lower average power than the exact marginal distribution. This difference becomes increasingly small for $\alpha > 4$.

5) Shaping Gains: We use HCA to a compute closed-form expression for the shaping gain as a function of N and PAR. This follows directly from analytical expressions for rate loss and average powers as functions of N and PAR. Using the volumes in (5) and (8) and the average power in (16), we obtain the following expression for the shaping gain in terms of N and α:

$$\gamma(N, \alpha) = \frac{N}{2\alpha\mathbb{E}[U]} \left(\frac{\kappa(N, \alpha/N)}{N!}\right)^{1/N}. \tag{20}$$

Equation (20) enables computation of the shaping gain in terms of only N and α, as U only depends on N and α. Note that by sweeping α, we can indirectly determine the PAR as described in (19). In all of the following results, we use the exact PAR.

Fig. 4 shows the shaping gain vs. PAR for various values of N using (20). The top solid line is the shaping gain computed by solving the maximum-entropy optimization problem given in [31]. The top dashed line is the ultimate shaping gain of $10\log_{10}(e/2)$, which was originally computed in [31] and included as a reference. In contrast with the solution given by solving the maximum-entropy optimization problem, HCA yields a semi-analytic derivation for the shaping gain that is parameterized by both N and α . For example, a shaping gain

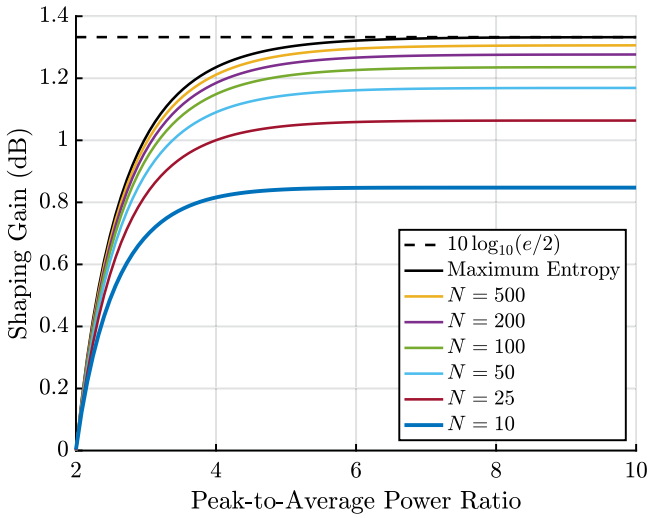

Fig. 4. Shaping gain vs. peak-to-average (power) ratio for various values of N. Maximum entropy is obtained by solving the optimization problem in [31]. The ultimate shaping gain of $10\log_{10}(e/2) \approx 1.33$ dB is provided as a reference.

of approximately 1 dB can be achieved using $N = 25$ and PAR $= 4$.

In general, the shaping gain increases with both N and PAR. Shaping gains within 0.10 dB and 0.02 dB of the ultimate shaping gain are achieved with PAR values of 4 and 6, respectively. In addition, shaping gains within 0.49 dB and 0.10 dB of the ultimate shaping gain can be achieved using $N = 10$ and $N = 100$, respectively, with no restriction on PAR.

IV. Coded Modulation With Finite Constellations

HCA can be used to derive closed-form continuous approximations to the optimal input distribution and the shaping gain at high SNR. However, finite and discrete constellations must be used in practice. In this section, we compute the coded modulation (CM) capacity for the thermal noise-limited PAPC SDD channel. We first discuss the methodology for computing the optimal discrete input distribution and truncated exponential distribution subject to constraints on the peak power and average power. We then compare the numerically optimized input distribution to the scaled geometric distribution for PAR ≥ 2 and show that the scaled geometric distribution achieves shaping gains within 0.22 dB of the numerically optimized shaping gains for modulation orders $M = 4, 8, 16$.

A. Coded Modulation Capacity

The coded modulation capacity is the applicable definition for the channel capacity when finite, discrete, and uniformly spaced[1] constellations are used [35]. The coded modulation capacity for the PAPC thermal noise-limited SDD channel is

$$C_{\text{CM}}^{(\text{SD})}\left(\text{SNR}^{(\text{th})}, \text{PAR}, M\right) \triangleq \max_{\substack{\sigma_{\text{th}}, P_I(i): \\ \frac{(\mathbb{E}[I])^2}{\sigma_{\text{th}}^2} \leq \text{SNR}^{(\text{th})} \\ \frac{\max(\mathcal{I})}{\mathbb{E}[I]} \leq \text{PAR}}} \mathbb{I}(I, Y), \tag{21}$$

where M is the modulation order, $\mathbb{I}(I, Y)$ is the MI, $I > 0$ is the transmitted intensity, $\mathcal{I}$ is a given set of input intensities, and $P_I(i)$ is the probability mass function (PMF) over $\mathcal{I}$. An equivalent description of coded modulation capacity optimizes over the spacing between signal points in $\mathcal{I}$ instead of the noise variance σ_{th}^2 [35].

The PAR constraint in (21) requires that $\mathbb{E}[I] \geq \max\{\mathcal{I}\}/\text{PAR}$, which places a lower bound on $\mathbb{E}[I]$ that is independent of σ_{th}. The noise standard deviation σ_{th} effectively parameterizes $E[I]$, so the PAR constraint places a lower bound on σ_{th}, i.e., $\sigma_{\text{th}}^2 \geq \frac{(\max\{\mathcal{I}\})^2}{(\text{PAR})^2 \cdot \text{SNR}^{(\text{th})}}$.

We numerically estimate MI for three different input distributions: the optimal input distribution computed using the Blahut-Arimoto (BA) algorithm [36], [37], the scaled geometric distribution, and the discrete uniform distribution. We note that the discrete uniform distribution is a special case of the scaled geometric distribution with a shape parameter equal to zero.

The optimal input PMF can be computed using the BA algorithm by optimizing over $\sigma_{\text{th}}^2 \geq \frac{\max\{(\mathcal{I})\}^2}{(\text{PAR})^2 \cdot \text{SNR}^{(\text{th})}}$. For lower values of $\text{SNR}^{(\text{th})}$ and PAR > 2, the optimized value of σ_{th}^2 is very close to the lower bound specified above. For sufficiently high $\text{SNR}^{(\text{th})}$ and PAR > 2, the lower bound is loose.

The probability density function (PDF) for the truncated exponential distribution with maximum power constraint A and shaping parameter β is

$$f_I(i; \beta, A) = \frac{\beta}{1 - \exp(-\beta A)} \exp(-\beta i), \; i \in [0, A]. \tag{22}$$

The average power for the continuous truncated exponential distribution is $\frac{1}{\beta} - A\frac{\exp(-\beta A)}{1-\exp(-\beta A)}$. For a fixed peak power A, the average power of the continuous truncated exponential distribution can be controlled using β.

The scaled geometric distribution is the discrete version of the exponential distribution. The scaled geometric distribution with modulation order M and peak power A is obtained by sampling and normalizing M points from (22) uniformly over $[0, A]$. Due to Jensen's inequality, the average power of the scaled geometric distribution is lower-bounded by the average power of the exponential distribution. In practice, the average power desired for scaled geometric distribution can be obtained by choosing a slightly smaller value of β than for the exponential distribution.

We note that $f_I(i; \beta, A) \propto f_I(i; \beta, \infty)$ over $[0, A]$. Therefore, sampling and normalizing from an exponential distribution with support from $[0, \infty]$ will result in the same PMF as sampling from (22). To obtain a scaled geometric PMF with the desired PAR, we first fix A and then compute the β required to achieve the target PAR. We then scale σ_{th} to achieve the $\text{SNR}^{(\text{th})}$ constraint.

[1]For simplicity, we follow the convention from [35] where uniformly spaced constellations are assumed. It is possible to use a more general notion of coded modulation capacity which does not assume uniformly spaced constellations.

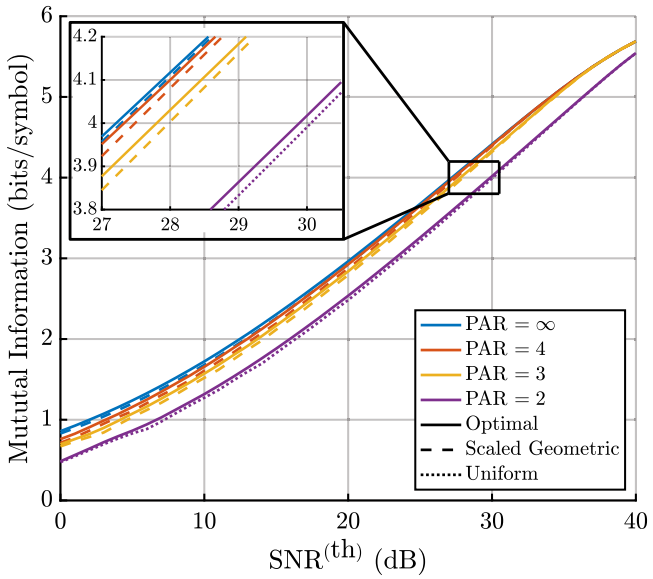

Fig. 5. Mutual information (bits/symbol) vs. SNR$^{(\mathrm{th})}$ for $M = 64$ and varying values for the peak-to-average (power) ratio (PAR). The optimal input distribution is computed using the Blahut-Arimoto method and is jointly optimized with the thermal noise variance. The scale and shape parameters of the scaled geometric distribution are jointly optimized to maximize mutual information subject to the PAR constraint. For PAR $= 2$, the scaled geometric distribution is identical to the uniform distribution, so only a single dotted line is included.

The PAR for the discrete uniform distribution is 2, which is the minimum PAR that we consider in this paper. Therefore, the PAR constraint does not affect the uniform distribution. Because $\mathcal{I}$ and M are fixed, we can estimate the achievable rate at SNR$^{(\mathrm{th})}$ under uniform signaling by choosing the σ_{th} that satisfies $\sigma_{\mathrm{th}}^2 = \frac{(\mathbb{E}[I])^2}{\mathrm{SNR}^{(\mathrm{th})}}$. We note that the uniform distribution can be seen as a special case of the truncated exponential distribution with $\beta = 0$.

B. Varying Peak-to-Average (Power) Ratio Constraints

The HCA derivation for the thermal noise-limited PAPC SDD channel is approximated by the truncated exponential distribution. In this section, we study shaping using the scaled geometric distribution for PAR $= 3, 4, \infty$. We first study the shaping gain using $M = 64$ and show that the closed-form expression for shaping gain closely predicts the empirical shaping gain under various PAR constraints. We then study shaping for more practical modulation orders of $M = 4, 8, 16$ and show that the scaled geometric distribution achieves shaping gains within 0.22 dB of the optimal shaping gains across all M and PAR values studied.

1) Shaping for Higher Modulation Order: In this section, we provide numerical analysis to verify the analytical results derived in Section III. Modulation orders of $M = 64$ or greater are generally sufficient to verify the shaping gains provided by HCA for direct-detection systems under average power constraints [3].

Fig. 5 shows MI vs. SNR$^{(\mathrm{th})}$ for PAR $= 2, 3, 4,$ and ∞ and modulation order $M = 64$. The solid lines indicate that the input distribution was designed using the Blahut-Arimoto algorithm and is jointly optimized with σ_{th}^2 as described in Section IV-A. The dashed lines indicate that the input distribution is a scaled geometric distribution with optimized values for β and σ_{th}. For PAR $= 2$, the scaled geometric distribution reduces to the discrete uniform distribution and is indicated by the dotted line. The inset plot provides a closer view of the shaping gains near MI $= 4$ bits/symbol.

TABLE I
SHAPING GAINS AT MUTUAL INFORMATION $= 4$ BITS/SYMBOL FROM FIG. 5

M	PAR	Shaping Gain: Blahut-Arimoto	Shaping Gain: Geometric	Shaping Gain: Analytical	Difference: Blahut-Arimoto	Difference: Geometric
64	∞	1.42 dB	1.39 dB	1.33 dB	0.09 dB	0.06 dB
	4	1.37 dB	1.29 dB	1.24 dB	0.13 dB	0.05 dB
	3	1.13 dB	1.04 dB	1.01 dB	0.12 dB	0.03 dB
	2	0.09 dB	0 dB	0 dB	0.09 dB	0 dB

Table I lists the shaping gains at 4 bits/symbol for the optimal distribution and the scaled geometric distribution, and the analytical shaping gains for $M = 64$ and PAR $= 2, 3, 4,$ and ∞. The shaping gains correspond to the differences in the average optical power required to achieve MI $= 4$ bits/symbol in the inset plot of Fig. 5. The analytical shaping gain is computed from (20) and is depicted in Fig. 4.

The empirical shaping gains obtained by the optimal input distributions for PAR $= 2, 3, 4,$ and ∞ are 0.09 dB to 0.13 dB larger than the analytically derived shaping gains. Similarly, the empirical shaping gains obtained by the scaled geometric distribution PAR $= 3, 4,$ and ∞ are 0.03 dB to 0.06 dB larger than the analytically derived shaping gains.

The shaping gain relative to uniform signaling generally decreases as SNR$^{(\mathrm{th})}$ increases over the range of SNR$^{(\mathrm{th})}$ studied. This is a feature generally observed among many direct-detection receivers in optical communications [3]. The decrease in shaping gain is in contrast with standard coherent detection, where the shaping gain increases with both modulation order and SNR. Therefore, for larger M and higher SNR$^{(\mathrm{th})}$, we expect the shaping gain differences observed in Table I to decrease, as is the case in the average power-constrained thermal noise-limited SDD channel [3].

2) Shaping for Lower Modulation Order: We now consider shaping for modulation orders $M = 4, 8,$ and 16. This study is motivated, in part, because lower-order modulation orders are commonly used in short-reach optical links [10], [38].

Fig. 6 shows MI vs. SNR$^{(\mathrm{th})}$ for $M = 4, 8,$ and 16. For each modulation order and from top to bottom, PAR is varied from ∞ to 2 in descending order. Solid lines and dashed lines indicate the optimal input distribution and scaled geometric input distribution, respectively, which are computed using the methods described in Section IV-A. The dotted line indicates the uniform distribution, which is a special case of the scaled geometric distribution with PAR $= 2$. (a), (b), and (c) use the optimal M from the sets $\{2, 3, 4\}$, $\{2, 3, 4, 6, 8\}$, and $\{2, 3, 4, 6, 8, 16\}$, respectively.

Table II shows the average shaping gain for $M = 4, 8,$ and 16 and PAR $= 2, 3, 4,$ and ∞ for the data shown in Fig. 6. The average shaping gain for each input distribution type, M, and PAR collection is defined as half the average difference in SNR$^{(\mathrm{th})}$ between the optimized input distribution and uniform

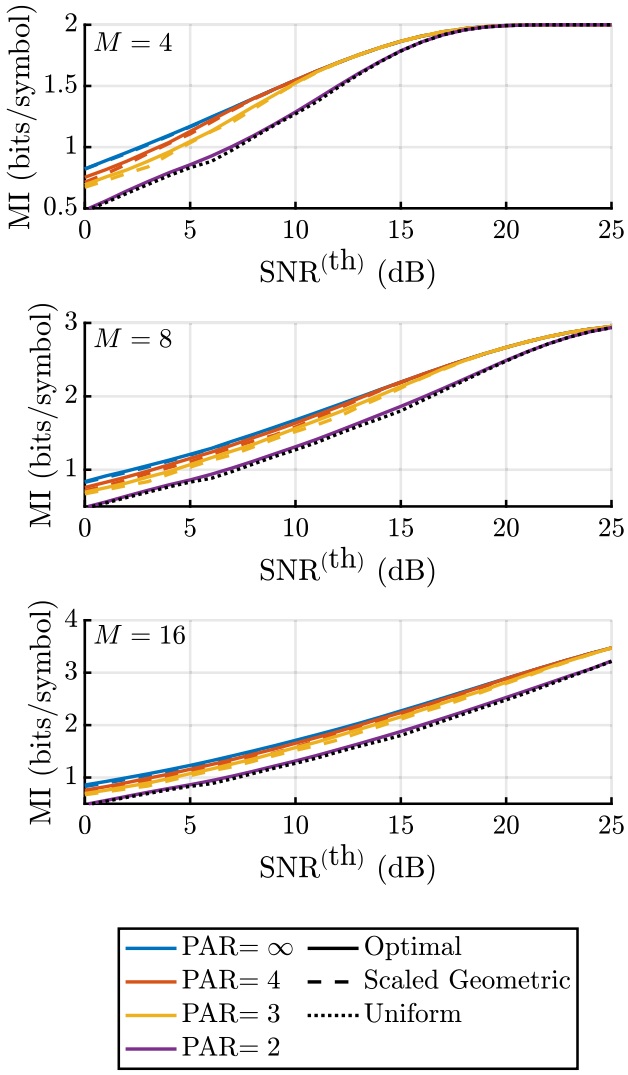

Fig. 6. Mutual information (MI) vs $\mathrm{SNR}^{(\mathrm{th})}$ for (a) $M = 4$, (b) $M = 8$, and (c) $M = 16$. From top to bottom, the color indicates the peak-to-average (power) ratio (PAR) constraint. The solid lines indicate that the input distribution is computed using the Blahut-Arimoto algorithm and jointly optimized with the noise variance. The dash-dotted lines indicate that a scaled geometric input distribution with an optimized shape parameter is used. The dotted line indicates that the uniform input distribution is used.

signaling over a specific $\mathrm{SNR}^{(\mathrm{th})}$ interval. For a fixed input distribution type, M, and PAR, the average difference is computed over the interval $\mathrm{SNR}^{(\mathrm{th})} \in [0, K]$ dB, where K is the $\mathrm{SNR}^{(\mathrm{th})}$ required to achieve $\mathrm{MI} = 2/3 \log_2(M)$. In the third and fourth columns, the optimal input distribution and scaled geometric input distribution are used, respectively. The fifth column is the difference in average shaping gain between the optimal input distribution and scaled geometric input distribution.

The scaled geometric distribution achieves an average shaping gain within 0.22 dB of the highest value obtained for all values of M and PAR values studied. This corresponds to shaping gains within 0.10 dB, 0.21 dB, and 0.22 dB of the numerically optimized average shaping gain for $M = 4$, $M = 8$, and $M = 16$, respectively.

TABLE II
AVERAGE SHAPING GAINS FOR $M = 4, 8,$ AND 16 FROM FIG. 6

M	PAR	Shaping Gain: Blahut-Arimoto	Shaping Gain: Geometric	Difference
4	∞	2.12 dB	2.09 dB	0.03 dB
	4	1.81 dB	1.72 dB	0.09 dB
	3	1.34 dB	1.24 dB	0.10 dB
	2	0.12 dB	0 dB	0.12 dB
8	∞	2.10 dB	2.04 dB	0.06 dB
	4	1.82 dB	1.61 dB	0.21 dB
	3	1.40 dB	1.19 dB	0.21 dB
	2	0.22 dB	0 dB	0.22 dB
16	∞	2.03 dB	1.90 dB	0.13 dB
	4	1.75 dB	1.54 dB	0.21 dB
	3	1.38 dB	1.16 dB	0.22 dB
	2	0.22 dB	0 dB	0.22 dB

For each fixed M, we generally observe that the absolute difference in shaping gain increases as PAR decreases from ∞ to 2. The absolute difference in shaping gain is largest for PAR $= 2$ for each M studied. We note that when PAR $= 2$, the scaled geometric distribution is equivalent to the uniform distribution. We also note that for a fixed PAR, the absolute difference in shaping gain tends to increase as M increases.

V. CONCLUSION

We studied PS for the thermal noise-limited PAPC SDD channel using HCA. We derived closed-form expressions for the shaping gain at arbitrary N and PAR and showed that the induced marginal density is well approximated by the truncated beta distribution or the truncated exponential distribution for sufficiently high N . We also considered shaping using finite, discrete constellations for the thermal noise-limited PAPC SD channel. We showed that the scaled geometric distribution achieves an average shaping gain within 0.22 dB of the numerically optimized distribution over a wide range of PAR and M. We used HCA to derive closed-form expressions for the induced marginal density and shaping gain for the thermal noise-limited PAPC SD channel with finite extinction ratio.

APPENDIX A
THE CONTINUOUS APPROXIMATION AND JORDAN MEASURE

Given a fixed dimension N, scale factor c, a coset Λ, and a Jordan-measurable support region $\mathbb{D}$, we are concerned with computing the average error probability of a coset code $\mathbb{C}(c\Lambda, \mathbb{D}) := c\Lambda \cap \mathbb{D}$. For sake of simplicity, we further assume Λ is a lattice with unit covering radius so that $c\Lambda$ has covering radius c. Note that the error probability is independent of constant shifts of the lattice points under signal-independent additive white Gaussian noise (AWGN). Let y be a signal point from $\mathbb{C}(c\Lambda, \mathbb{D})$ and let $\Pi(y) = \{x : \arg\min_{z \in \mathbb{C}(c\Lambda, \mathbb{D})} \|z - x\|_2 = y\}$ be the Voronoi region for y. Under AWGN, the conditional error

probabilities of two signal points from $\mathbb{C}(c\Lambda, \mathbb{D})$ are equal if their Voronoi regions are congruent.

Tarokh et al. [32] showed that for a spherical bounding region, the fraction of signal points from $\mathbb{C}(c\Lambda, \mathbb{D})$ whose Voronoi regions are not mutually congruent approaches zero as $c \to 0$. The authors further remark that this property also holds for squares and any "sufficiently well-behaved support region". In this Appendix, we further quantify the concept of a "sufficiently well-behaved" support region. Specifically, we prove that the desired property holds for any bounded Jordan-measurable support region $\mathbb{D}$.

We first establish some definitions. Let $\partial\mathbb{D}$ be the boundary of $\mathbb{D}$. Let us define $\partial\mathbb{D}_\epsilon$ as the set of points that is ϵ near $\partial\mathbb{D}$. Analytically, $\partial\mathbb{D}_\epsilon$ is given by

$$\partial\mathbb{D}_\epsilon := \{x : \|x - v\|_2 \leq \epsilon \text{ for some } v \in \partial\mathbb{D}\}. \tag{23}$$

We note that $\partial\mathbb{D}_\epsilon \not\subset \mathbb{D}$, i.e., that $\partial\mathbb{D}_\epsilon$ contains points not in $\mathbb{D}$. We define the near-boundary points as $\partial\mathbb{D}_{2c} \cap \mathbb{C}(c\Lambda, \mathbb{D})$. The interior points are similarly given by $(\mathbb{D}\backslash\partial\mathbb{D}_{2c}) \cap \mathbb{C}(c\Lambda, \mathbb{D})$. For concreteness, we assume that the zero point is an interior point with respect to $\mathbb{D}$ with Voronoi region Π. Under this assumption, all interior points have Voronoi regions that are congruent to Π [32]. The Voronoi regions for the near-boundary points may not be congruent to Π. Let M_1 and M_2 denote the number of interior points and near-boundary points, respectively.

In the following, we show that for a fixed dimension N, a Jordan-measurable set $\mathbb{D}$, and scaled lattice $c\Lambda$, it holds that

$$\frac{|\{y \in \mathbb{C}(c\Lambda, \mathbb{D}) : \Pi(y) \not\subset \mathbb{D}\}|}{|\mathbb{C}(c\Lambda, \mathbb{D})|} \to 0 \text{ as } c \to 0. \tag{24}$$

Note that increasing the rate is equivalent to decreasing the covering radius c [32].

First, observe that $\mathcal{V}(\partial\mathbb{D}_\epsilon) \to 0$ as $\epsilon \to 0$. This holds because $\partial\mathbb{D}$ is the limit of the sequence $\{\partial\mathbb{D}_{\frac{1}{n}}\}, n = 1, 2, \ldots$. Since $\mathbb{D}$ is Jordan-measurable, it has a measure-zero boundary ($\mathcal{V}(\partial\mathbb{D}) = 0$) [39]. Thus

$$\begin{aligned}\lim_{\epsilon\to 0} \mathcal{V}(\partial\mathbb{D}_\epsilon) &= \mathcal{V}\left(\lim_{\epsilon\to 0} \partial\mathbb{D}_\epsilon\right) \\ &= \mathcal{V}(\partial\mathbb{D}) \\ &= 0.\end{aligned}$$

Next, we bound the ratio of the number near-boundary points to total number of lattice points:

$$M_2 \leq \frac{\mathcal{V}((\partial\mathbb{D}_{3c} \cap \mathbb{D}) \cup \partial\mathbb{D}_c)}{\mathcal{V}(\Pi)},$$

$$M_1 + M_2 \geq \frac{\mathcal{V}(\mathbb{D}) - \mathcal{V}(\partial\mathbb{D}_c \cap \mathbb{D})}{\mathcal{V}(\Pi)}.$$

The first inequality follows from the fact that $\bigcup_{y\in\partial\mathbb{D}_{2c}\cap\mathbb{C}(c\Lambda,\mathbb{D})}(y + \Pi) \subseteq (\partial\mathbb{D}_{3c} \cap \mathbb{D}) \cup \partial\mathbb{D}_c$. The second inequality follows similarly from the fact that $\mathbb{D}\backslash\partial\mathbb{D}_c \subseteq \bigcup_{y\in\mathbb{C}(c\Lambda,\mathbb{D})}(y + \Pi)$. Combining these two inequalities, we get

$$\frac{M_2}{M_1 + M_2} \leq \frac{\mathcal{V}((\partial\mathbb{D}_{3c} \cap \mathbb{D}) \cup \partial\mathbb{D}_c)}{\mathcal{V}(\mathbb{D}) - \mathcal{V}(\partial\mathbb{D}_c \cap \mathbb{D})}.$$

With the construction above, we obtain

$$\begin{aligned}\frac{M_2}{M_1 + M_2} &\leq \frac{\mathcal{V}((\partial\mathbb{D}_{3c} \cap \mathbb{D}) \cup \partial\mathbb{D}_c)}{\mathcal{V}(\mathbb{D}) - \mathcal{V}(\partial\mathbb{D}_c \cap \mathbb{D})} \\ &\leq \frac{\mathcal{V}(\partial\mathbb{D}_{3c})}{\mathcal{V}(\mathbb{D}) - \mathcal{V}(\partial\mathbb{D}_c \cap \mathbb{D})} \\ &\to 0 \text{ as } c \to 0 \qquad \left(\lim_{c\to 0} \mathcal{V}(\partial\mathbb{D}_{3c}) = 0\right).\end{aligned}$$

Hence, as desired we have

$$\lim_{c\to 0} \frac{M_2}{M_1 + M_2} = 0. \tag{25}$$

Appendix B
Shaping With Finite Extinction Ratio

The HCA derivation presented in Section III uses the bounding region in (3). This bounding region presumes that constellation points with one or more coordinates with exactly zero power can be transmitted. In practice, however, most direct-detection optical links have limited extinction ratio, precluding the modulation of constellation points with coordinates that have precisely or nearly zero power.

In this appendix, we use HCA to analyze the thermal noise-limited PAPC SD channel with additional finite extinction-ratio constraints. We derive the optimal bounding region that minimizes the average optical power subject to the PAPC and finite extinction-ratio constraints. We then derive closed-form expressions for the resulting induced marginal density and shaping gain. Note that a tilde symbol is used to distinguish the notation in Appendix B from that in Section III.

1) Optimal Bounding Region: The extinction ratio $r_{\text{ex}} \geq 1$ is defined as the ratio of the maximum transmitted optical power and the minimum transmitted optical power. The minimum transmitted power in any time interval is A/r_{ex}. The optimal bounding region for the average power-constrained thermal noise-limited SDD channel subject to both peak-power and extinction-ratio constraints is given by

$$\tilde{\mathcal{R}}\left(N, P, A, r_{\text{ex}}\right) = A \cdot [1/r_{ex}, 1] \cap P \cdot \mathbb{S}_N. \tag{26}$$

The bounding region in (26) can be written in terms of a shifted version of the PAPC bounding region in (3):

$$\begin{aligned}&\tilde{\mathcal{R}}\left(N, P, A, r_{\text{ex}}\right) \\ &\quad = \mathcal{R}\left(N, P - N\frac{A}{r_{\text{ex}}}, A\left(1 - \frac{1}{r_{\text{ex}}}\right)\right) + \mathbf{1}\frac{A}{r_{\text{ex}}}.\end{aligned} \tag{27}$$

The bounding region from (26) is illustrated by the shaded region in Fig. 7.

2) Induced Marginal Density: From (27), we can see that the bounding region $\tilde{\mathcal{R}}(N, P, A, r_{\text{ex}})$ is nearly equivalent to the bounding region without an extinction-ratio constraint, but with three differences. The total optical power constraint P is decreased to $P - N\frac{A}{r_{\text{ex}}}$ and peak-power constraint A is decreased to $A(1 - \frac{1}{r_{\text{ex}}})$. Lastly, the bounding region is shifted by the vector $\mathbf{1}\frac{A}{r_{\text{ex}}}$. Because shifting by $\mathbf{1}\frac{A}{r_{\text{ex}}}$ does not change the enclosed

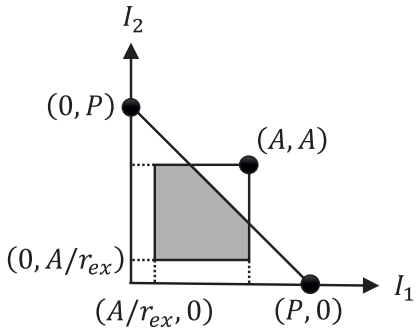

Fig. 7. Optimal bounding region for the thermal noise-limited PAPC SDD channel with constraints P, A, and r_{ex} and $N = 2$. $\tilde{\mathcal{R}}(N, P, A, r_{ex})$ from (26) is depicted by the shaded region.

volume, we have

$$\tilde{\mathcal{V}}\left(N, P, A, r_{\text{ex}}\right) = \mathcal{V}\left(N, P - N\frac{A}{r_{\text{ex}}}, A\left(1 - \frac{1}{r_{\text{ex}}}\right)\right). \quad (28)$$

The exact induced marginal density with limited r_{ex} can be computed using (28) in a similar method to the derivation in Section III-3. The resulting induced marginal distribution is

$$\begin{aligned} &f_{\tilde{I}}(i; N, A, \alpha, r_{\text{ex}}) \\ &\propto \left(1 - \frac{\alpha}{r_{\text{ex}}} - \frac{\alpha}{N}\left(\frac{i}{A} - \frac{1}{r_{\text{ex}}}\right)\right)^{N-1} \\ &\quad \cdot \kappa\left(N - 1, \frac{\alpha\left(1 - \frac{1}{r_{\text{ex}}}\right)}{N\left(1 - \frac{\alpha}{r_{\text{ex}}}\right) - \alpha\left(\frac{i}{A} - \frac{1}{r_{\text{ex}}}\right)}\right), \\ &\quad i \in \left[\frac{A}{r_{\text{ex}}}, A\right]. \end{aligned} \quad (29)$$

Similar to the derivation for (14), we can compute a normalized input distribution using $\tilde{U} = \tilde{I}/A$. The induced marginal density is given by

$$\begin{aligned} &f_{\tilde{U}}(u; N, \alpha, r_{\text{ex}}) \\ &\propto \left(1 - \frac{\alpha}{r_{\text{ex}}} - \frac{\alpha}{N}\left(u - \frac{1}{r_{\text{ex}}}\right)\right)^{N-1} \\ &\quad \cdot \kappa\left(N - 1, \frac{\alpha\left(1 - \frac{1}{r_{\text{ex}}}\right)}{N\left(1 - \frac{\alpha}{r_{\text{ex}}}\right) - \alpha\left(u - \frac{1}{r_{\text{ex}}}\right)}\right), \\ &\quad u \in \left[\frac{1}{r_{\text{ex}}}, 1\right]. \end{aligned} \quad (30)$$

Equation (30) does not depend on A or P directly, but only indirectly through α. The average power of the induced marginal density is given by

$$\bar{\tilde{P}}\left(N, A, \alpha, r_{\text{ex}}\right) = \mathbb{E}[\tilde{I}] = A \cdot \mathbb{E}[\tilde{U}], \quad (31)$$

where the PDFs of $\tilde{I}$ and $\tilde{U}$ are defined in (29) and (30), respectively. (31) can be computed by integrating (30).

3) Shaping Gain: Using the volumes in (5) and (28) and the average power in (31), we obtain the following expression for the shaping gain:

$$\tilde{\gamma}(N, \alpha, r_{\text{ex}}) = \frac{N\left(1 - \frac{\alpha}{r_{\text{ex}}}\right)}{2\alpha\mathbb{E}[\tilde{U}]}\left(\frac{\kappa\left(N, \frac{\alpha}{N}\left(\frac{1 - \frac{1}{r_{\text{ex}}}}{1 - \frac{\alpha}{r_{\text{ex}}}}\right)\right)}{N!}\right)^{1/N}. \quad (32)$$

Note that (32) reduces to (20) as $r_{\text{ex}} \to \infty$.

There are several differences between (20) and (32). One difference arises from different definitions for U and $\tilde{U}$, although both may be approximated by a truncated beta distribution. (20) also contains the term $(1 - \frac{\alpha}{r_{\text{ex}}})$, which arises from the unmodulated carrier corresponding to the limited extinction ratio. Lastly, the second argument in κ is modified due to the inclusion of the extinction ratio.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: A review," *Optica*, vol. 5, no. 11, pp. 1354–1370, Nov. 2018.

[2] A. Jahid, M. H. Alsharif, and T. J. Hall, "A contemporary survey on free space optical communication: Potentials, technical challenges, recent advances and research direction," *J. Netw. Comput. Appl.*, vol. 200, 2022, Art. no. 103311.

[3] E. M. Liang and J. M. Kahn, "Probabilistic shaping distributions for optical communications," *J. Lightw. Technol.*, vol. 43, no. 4, pp. 1501–1524, Feb. 2025.

[4] A. D. Ellis, M. E. McCarthy, M. A. Z. A. Khateeb, M. Sorokina, and N. J. Doran, "Performance limits in optical communications due to fiber nonlinearity," *Adv. Opt. Photon.*, vol. 9, no. 3, pp. 429–503, Sep. 2017.

[5] X. Zhou and C. Lam, "Four-wave mixing penalty for WDM-based ethernet PMDs in O-band," presented at the *IEEE 802.3df Electron. Plenary Meeting*, May 2022.

[6] H. Srinivas et al., "Modeling and experimental measurement of power efficiency for power-limited SDM submarine transmission systems," *J. Lightw. Technol.*, vol. 39, no. 8, pp. 2376–2386, Apr. 2021.

[7] A. Chaaban, Z. Rezki, and M.-S. Alouini, "On the capacity of intensity-modulation direct-detection Gaussian optical wireless communication channels: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 455–491, Firstquarter 2022.

[8] *Safety of Laser Products*, Standard IEC 60825:2025, International Electrotechnical Commission, Geneva, Switzerland, Jan. 2025.

[9] G. Böcherer, P. Schulte, and F. Steiner, "Probabilistic shaping and forward error correction for fiber-optic communication systems," *J. Lightw. Technol.*, vol. 37, no. 2, pp. 230–244, Jan. 2019.

[10] *IEEE Standard for Ethernet - Amendment 10: Media Access Control Parameters, Physical Layers, and Management Parameters for 200 Gb/s and 400 Gb/s Operation*, IEEE Standard 802.3bs-2017, Dec. 2017.

[11] F. R. Kschischang and S. Pasupathy, "Optimal shaping properties of the truncated polydisc," *IEEE Trans. Inf. Theory*, vol. 40, no. 3, pp. 892–903, May 1994.

[12] A. Amari et al., "Introducing enumerative sphere shaping for optical communication systems with short blocklengths," *J. Lightw. Technol.*, vol. 37, no. 23, pp. 5926–5936, 2019.

[13] S. Hranilovic and F. R. Kschischang, "Capacity bounds for power- and band-limited optical intensity channels corrupted by Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 784–795, May 2004.

[14] A. A. Farid and S. Hranilovic, "Capacity bounds for wireless optical intensity channels with Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6066–6077, Dec. 2010.

[15] A. Lapidoth, S. M. Moser, and M. A. Wigger, "On the capacity of free-space optical intensity channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4449–4461, Oct. 2009.
[16] A. A. Farid and S. Hranilovic, "Channel capacity and non-uniform signalling for free-space optical intensity channels," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 9, pp. 1553–1563, Dec. 2009.
[17] S. M. Moser, "Capacity results of an optical intensity channel with input-dependent Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 207–223, Jan. 2012.
[18] T. H. Chan, S. Hranilovic, and F. R. Kschischang, "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2073–2088, Jun. 2005.
[19] Z. He, T. Bo, and H. Kim, "Probabilistically shaped coded modulation for IM/DD system," *Opt. Exp.*, vol. 27, no. 9, pp. 12126–12136, Apr. 2019.
[20] T. Wiegart, F. Da Ros, M. P. Yankov, F. Steiner, S. Gaiarin, and R. D. Wesel, "Probabilistically shaped 4-PAM for short-reach IM/DD links with a peak power constraint," *J. Lightw. Technol.*, vol. 39, no. 2, pp. 400–405, Jan. 2021.
[21] D. Che, J. Cho, and X. Chen, "Does probabilistic constellation shaping benefit IM-DD systems without optical amplifiers?," *J. Lightw. Technol.*, vol. 39, no. 15, pp. 4997–5007, 2021.
[22] T. Prinz et al., "PAM-6 coded modulation for IM/DD channels with a peak-power constraint," in *Proc. 11th Int. Symp. Topics Coding*, 2021, pp. 1–5.
[23] M. Liu, M. Gao, and J. Ke, "Multi-distributed probabilistically shaped PAM-4 system for intra-data-center networks," *Chin. Opt. Lett.*, vol. 19, no. 11, Nov. 2021, Art. no. 110604.
[24] Z. Sun, D. Tang, W. Zuo, H. Cui, Z. Wu, and Y. Qiao, "Triple-convex probabilistic constellation shaping PAM8 in IM/DD system," *IEEE Photon. Technol. Lett.*, vol. 35, no. 15, pp. 846–849, Aug. 2023.
[25] M. S.-B. Hossain et al., "Probabilistic shaping for high-speed unamplified IM/DD systems with an O-band EML," *J. Lightw. Technol.*, vol. 41, no. 16, pp. 5373–5382, Aug. 2023.
[26] D. Kim and H. Kim, "Optimum symbol distribution of probabilistically shaped PAM signals in amplifier-less IM-DD systems," presented at the *26th Optoelectron. Commun. Conf.*, Hong Kong, China, Jul. 3–7, 2021, Paper M3A.5.
[27] D. Kim, Z. He, T. Bo, Y. Yu, and H. Kim, "Transmission of 36-Gbaud PAM-8 signal in IM/DD system using pairwise-distributed probabilistic amplitude shaping," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, 2020, pp. 1–3.
[28] D. Kim and H. Kim, "Capacity-achieving symbol distributions for directly modulated laser and direct detection systems," *Opt. Exp.*, vol. 31, no. 8, pp. 12609–12623, Apr. 2023.
[29] S. Hranilovic and F. Kschischang, "Optical intensity-modulated direct detection channels: Signal space and lattice codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1385–1399, Jun. 2003.
[30] G. D. Forney and L.-F. Wei, "Multidimensional constellations. I. Introduction, figures of merit, and generalized cross constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 877–892, Aug. 1989.
[31] D.-S. Shiu and J. M. Kahn, "Shaping and nonequiprobable signaling for intensity-modulated signals," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2661–2668, Nov. 1999.
[32] V. Tarokh, A. Vardy, and K. Zeger, "Universal bound on the performance of lattice codes," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 670–681, Mar. 1999.
[33] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2. Hoboken, NJ, USA: Wiley, 1995.
[34] Ethan M. Liang, "Shaping for optical direct-detection links," Ph.D. dissertation, Leland Stanford Junior Univ., Stanford, CA, USA, Mar. 2025.
[35] G. Böcherer, "Probabilistic signal shaping for bit-metric decoding," in *Proc. IEEE Int. Symp. Inf. Theory*, 2014, pp. 431–435.
[36] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.
[37] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.
[38] X. Zhou, R. Urata, and H. Liu, "Beyond 1 Tb/s intra-data center interconnect technology: IM-DD or coherent?," *J. Lightw. Technol.*, vol. 38, no. 2, pp. 475–484, Jan. 2020.
[39] T. Tao, *An Introduction to Measure Theory*, vol. 126. Providence, RI, USA: Amer. Math. Soc., 2011.

Ethan M. Liang (Graduate Student Member, IEEE) received the A.S. degree in computer science from El Camino College, Torrance, CA, USA, in 2016, the B.S. degree in electrical engineering from the University of California, Los Angeles, CA, in 2019, and the M.S. degree in electrical engineering in 2022 from Stanford University, Stanford, CA, where he is currently working toward the Ph.D. degree with the Department of Electrical Engineering. His research interests include optical communications and coding theory. He is a National Science Foundation Graduate Research Fellow.

Shannon Fan is currently working toward the dual B.S. degrees in electrical engineering and in mathematics with Stanford University, Stanford, CA, USA. Her research interests include optical communications and coding theory.

Joseph M. Kahn (Life Fellow, IEEE) received the A.B., M.A., and Ph.D. degrees in physics from the University of California, Berkeley, CA, USA, in 1981, 1983, and 1986, respectively. From 1987 to 1990, he was with AT&T Bell Laboratories. In 1989, he demonstrated the first successful synchronous (i.e., coherent) detection using semiconductor lasers, achieving record receiver sensitivity. From 1990 to 2003, he was with Electrical Engineering and Computer Sciences Faculty, Berkeley. He demonstrated coherent detection of QPSK in 1992. In 1999, D. S. Shiu and Kahn published the first work on probabilistic shaping for optical communications. In the 1990 s and early 2000 s, Kahn and collaborators performed seminal work on indoor and outdoor free-space optical communications and multi-input multi-output wireless communications. In 2000, Kahn and K. P. Ho founded StrataLight Communications, whose 40 Gb/s-per-wavelength long-haul fiber transmission systems were deployed widely by AT&T, Deutsche Telekom, and other carriers. In 2002, Ho and Kahn applied to patent the first electronic compensation of fiber Kerr nonlinearity. StrataLight was acquired by Opnext in 2009. In 2003, he became a Professor of electrical engineering with E. L. Ginzton Laboratory, Stanford University. Kahn and collaborators have extensively studied rate-adaptive coding and modulation, and digital signal processing for mitigating linear and nonlinear impairments in coherent systems. In 2008, E. Ip and Kahn (and G. Li independently) invented simplified digital backpropagation for compensating fiber Kerr nonlinearity and dispersion. Since 2004, Kahn and collaborators have been studying propagation, modal statistics, spatial multiplexing and imaging in multimode fibers, elucidating principal modes and demonstrating transmission beyond the traditional bandwidth-distance limit in 2005, deriving the statistics of coupled modal group delays and gains in 2011, and deriving resolution limits for imaging in 2013. His research interests include optical frequency comb generators, coherent data center links, rate-adaptive access networks, fiber Kerr nonlinearity mitigation, ultra-long-haul submarine links, and optimal free-space transmission through atmospheric turbulence. Kahn was the recipient of the National Science Foundation Presidential Young Investigator Award in 1991.