Capacity-Achieving Distributions for the Discrete-Time Poisson Channel–Part I: General Properties and Numerical Techniques

Jihai Cao, Steve Hranilovic, *Senior Member, IEEE*, and Jun Chen, *Member, IEEE*

***Abstract*—Despite being an accepted model for a wide variety of optical channels, few general results on optimal signalling for discrete-time Poisson (DTP) channels are known. Among the most significant is that under simultaneous peak and average constraints, the capacity-achieving distributions are discrete with a finite number of mass points.**

In this paper, several fundamental properties of capacity-achieving distributions for DTP channels are established. In particular, we demonstrate that all capacity-achieving distributions of the DTP channel have zero as a mass point. In the case of only a peak constraint, it is further shown that the optimal distribution always has a mass point at the maximum amplitude. Finally, under solely an average power constraint, it is shown that a finite number of mass points are insufficient to achieve the capacity. In addition to these analytical results, a numerical algorithm based on deterministic annealing is presented which can efficiently compute both the channel capacity and the associated optimal input distribution under peak and average power constraints. Numerical lower bounds based on the envelope of information rates induced by the maxentropic distributions are also shown to be extremely close to the capacity, especially in the low power regime.

***Index Terms*—Discrete-time Poisson channel, capacity-achieving distributions, deterministic annealing.**

I. Introduction

THE discrete-time Poisson (DTP) channel is commonly used to model low intensity, direct detection optical communication channels. For this channel model, the intensity of the input signal is allowed to vary between discrete time slots while remaining fixed inside each interval, and the channel output is a statistic on the number of received photons in each time interval [1]. Constraints are placed on both the peak power, A and the average optical power, ε.

Although currently no analytic expression is known for the capacity of the DTP channel, many capacity bounds have been developed. In [2], [3], McEliece derived several upper bounds on the capacity of the DTP channel with zero photodetector dark current under peak and average power constraints. In [4], Shamai established lower and upper bounds on the capacity of the DTP channel for binary inputs that are located at $\{0, A\}$ with average power constraint only. Lapidoth and Moser [5] derived analytical lower and upper bounds on the capacity of the DTP channel with dark current. These bounds are asymptotically tight as the peak and average powers tend to infinity, although they are often loose in the lower power regime. In [6] lower and upper bounds on DTP channel capacity are given asymptotically as the average power tends to zero with a fixed peak power. In [7], Martinez derived a class of tight lower bounds based on the Gamma distribution and non-asymptotic upper bounds via duality for the DTP channel with no dark current under only an average power constraint.

Relatively little work has been done on capacity-achieving distributions for the DTP channel. In [8], Smith provided a general characterization of the capacity-achieving distributions for peak amplitude and average power constrained additive Gaussian channels and established the Karush-Kuhn-Tucker (KKT) conditions on the optimality of an input distribution. Following Smith's approach, Shamai [9] proved that the capacity-achieving distributions for the DTP channel with peak and average constraints consist of a finite number of mass points. Extensions of Smith's seminal work to different but related channel models can be found in [10], [11], [12].

In contrast to previous work on the DTP channel which concentrates on bounding capacity, this paper investigates the capacity-achieving distributions of the DTP channel. This approach not only provides insight on the channel capacity but is also a useful tool to guide signalling design. It is shown that the capacity-achieving distributions always have a mass point at zero and, in the case with only a peak power constraint, also a point at the maximum amplitude. We also prove that distributions that consist of a finite number of mass points are not capacity-achieving for the DTP channel with only average power constraint.

In addition to the analytical results, this paper presents a computationally-efficient numerical algorithm based on the deterministic annealing (DA) framework of Rose [14] to find the capacity and the associated optimal input distributions. A somewhat surprising observation on the capacity achieving distributions for the DTP channel is that for a fixed average power constraint and increasing A, the peak power constraint can alternate between being active and inactive. We also obtain numerical capacity lower bounds using maxentropic source distributions. Simulation results show that the envelope of these bounds is very close to the channel capacity, especially when the received power is smaller or at most on the same order as the the dark current.

In the second part of this paper [20] the special case of

Manuscript received February 19, 2013; revised August 2 and October 31, 2013. The editor coordinating the review of this paper and approving it for publication was E. Agrell.

The authors are with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, Canada (e-mail: caojih@grads.ece.mcmaster.com, {hranilovic, junchen}@mcmaster.ca).

Digital Object Identifier 10.1109/TCOMM.2013.112513.130142

binary capacity-achieving distributions for the DTP channel are considered under a number of operating regimes.

The balance of this paper is organized as follows. Section II presents the channel model and Section III contains the developed general properties of the capacity-achieving distribution. A numerical algorithm to compute the capacity and the associated optimal input distribution, along with a class of tight lower bounds based on the maxentropic distributions, is presented in Section IV. The paper concludes in Section V.

II. Channel Model

In the DTP channel, data are transmitted by sending pulse amplitude modulation (PAM) intensity signals which are fixed in discrete time intervals. The receiver is modelled as a photon counter which generates an integer representing the number of received photons. Specifically, in each time interval, ΔT, a channel input intensity x [photons/second] is corrupted by the combined impact of background radiation and photodetector dark current at a rate of λ [photons/second]. The channel output y [photons], is a random value related to the the number of received photons in ΔT and obeys a Poisson distribution with mean $(x+\lambda)\Delta T$. Without loss of generality, assume $\Delta T = 1$ and accordingly

$$P_{Y|X}(y|x) = \frac{(x+\lambda)^y}{y!} e^{-(x+\lambda)}, \quad x \in \mathbb{R}^+,\ y \in \mathbb{Z}^+. \tag{1}$$

Furthermore, due to device constraints and limited energy storage at the transmitter (e.g., a satellite in deep-space laser communications), there is an average power constraint,

$$\mathbb{E}(X) \leq \varepsilon, \qquad \text{[Average Power Constraint]}. \tag{2}$$

Also, due to the dynamic range limitation of transmitter, the peak intensity must also be constrained,

$$0 \leq X \leq A, \qquad \text{[Peak Power Constraint]}. \tag{3}$$

Without loss of generality, it is assumed that $0 \leq \varepsilon \leq A$. Notice that the constraints in this model can also be relaxed to yield DTP channels with only average power constraint (i.e., $A \to \infty$) or only peak power constraint (i.e., $\varepsilon = A$). Unless otherwise noted, in this paper, A is finite.

The channel capacity, C, of a DTP channel is the maximum mutual information over input distributions satisfying average and peak power constraints. It is well known that the capacity-achieving distribution is unique and discrete with a finite number of mass points for finite A and ε [9]. Therefore, there is no loss of generality in considering an input distribution defined over constellation $\psi_x \triangleq \{x_1, x_2, \ldots, x_n\}$, $0 \leq x_1 < x_2 < \ldots < x_n \leq A$, with corresponding probability masses $\psi_p \triangleq \{p_1, p_2, \cdots, p_n\}$. Let F_x denote the cdf of the input, that is

$$dF_x = p_1\delta(x-x_1) + p_2\delta(x-x_2) + \ldots + p_n\delta(x-x_n),$$

where $\delta(\cdot)$ denotes the Dirac impulse functional. Thus,

$$\begin{aligned} C &\triangleq \max_{F_x \in \mathcal{F}} I(X;Y) \\ &= \max_{F_x \in \mathcal{F}} \int_x \left[\sum_y P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{P_Y(y)} \right] dF_x, \end{aligned} \tag{4}$$

where

$$\mathcal{F} \triangleq \left\{ F_x(x) : \int_0^A dF_x = 1, \mathbb{E}_{F_x}\{X\} \leq \varepsilon \right\},$$

$\mathbb{E}$ is the expectation operator, $P_Y(\cdot)$ denotes the corresponding distribution on the channel output and $P_{Y|X}(\cdot|\cdot)$ denotes the channel law.

Finally, define $\psi_x^*(A,\varepsilon)$, $\psi_p^*(A,\varepsilon)$, $F_x^*(A,\varepsilon)$ to be the corresponding optimal values under constraints A and ε.

III. General Results for Capacity-Achieving Distributions

In this section, three general properties of the capacity-achieving distribution are demonstrated under different constraints.

A. Properties of the capacity-achieving distribution

Lemma 1 (Shifting downward increases mutual information)**.** *Let X denote a random variable defined over constellation $\psi_x = \{x_1, x_2, \ldots, x_n\}$, $0 < x_1 < x_2 < \ldots < x_n \leq A$, with corresponding probability masses $\psi_p = \{p_1, p_2, \cdots, p_n\}$, $\forall i$ $p_i \neq 0$. Let Y denote the output of a DTP channel generated by X. Define another input $X_\Delta = X - \Delta$ with shifted constellation $\psi_{x_\Delta} = \{x_1 - \Delta, x_2 - \Delta, \cdots, x_n - \Delta\}$, and let Y_Δ denote the corresponding output. For any $\Delta \in (0, x_1]$,*

$$I(X;Y) \leq I(X_\Delta; Y_\Delta), \tag{5}$$

with equality if and only if $|\psi_x| = 1$.

Proof: A detailed proof can be found in Appendix A. ■

Corollary 2 (Mass point at zero)**.** *The capacity-achieving distribution for the DTP channel under average and peak power constraints always contains a mass point located at 0. That is, $0 \in \psi_x^*(A,\varepsilon)$ for any constraints A and ε.*

Proof: This is a direct consequence of Lemma 1. ■

Lemma 3 (Squeezing decreases mutual information)**.** *Let X denote a random variable defined over constellation $\psi_x = \{x_1, x_2, \ldots, x_n\}$, $0 \leq x_1 < x_2 < \ldots < x_n \leq A$, with corresponding probability masses $\psi_p = \{p_1, p_2, \cdots, p_n\}$, $\forall i$ $p_i \neq 0$. Let Y denote the output of a DTP channel generated by X. Define another input $X_\alpha = \alpha X$ with squeezed constellation $\{\alpha x_1, \alpha x_2, \cdots, \alpha x_n\}$, and let Y_α denote the corresponding output. For $\alpha \in [0,1)$,*

$$I(X;Y) \geq I(X_\alpha; Y_\alpha), \tag{6}$$

with equality if and only if $|\psi_x| = 1$.

Proof: This result can be proved by invoking [13, Theorem 2]. A more elementary proof is provided in Appendix B. ■

Corollary 4 (Point at peak amplitude)**.** *For the DTP channel, when only the peak power constraint is imposed, the capacity-achieving distribution always contains a mass point located at A, i.e., $A \in \psi_x^*(A,A)$.*

Proof: This follows from Lemma 3 directly. ■

Another interpretation of Lemma 1 is that the mutual information is monotonically decreasing with λ. By Proposition 8

in Appendix B, increasing λ is equivalent to shifting the constellation to the right. This interpretation can also be shown via [13, Theorem 1].

B. On the capacity-achieving distribution under only average power constraint

As in [8], define

$$\begin{aligned} i(x, F_x) &\triangleq -\sum_{y=0}^{\infty} P_{Y|X}(y|x) \log \frac{P_Y(y)}{P_{Y|X}(y|x)} \\ &= (x+\lambda)\log(x+\lambda) - x - \sum_{y=0}^{\infty} e^{-(x+\lambda)} \frac{(x+\lambda)^y}{y!} \\ &\quad \times \log\left(\int_0^A e^{-x}(x+\lambda)^y dF_x\right). \quad (7) \end{aligned}$$

The mutual information induced by F_x can be written as

$$I(F_x) = \int_0^A i(x, F_x) dF_x. \quad (8)$$

Finally, define the *multiplier function* with Lagrange multiplier $\mu \geq 0$ as

$$M(\mu, x, F_x) = I(F_x) + \mu(x - \varepsilon) - i(x, F_x). \quad (9)$$

The following theorem from [9] is of fundamental importance for this paper.

Theorem 5. *[KKT conditions [9, Eq. (24),(25)]] $F_x(A, \varepsilon)$ is capacity-achieving iff the following conditions are satisfied for some $\mu \geq 0$,*

$$M(\mu, x, F_x(A, \varepsilon)) \geq 0, \quad x \in [0, A], \quad (10)$$
$$M(\mu, x, F_x(A, \varepsilon)) = 0, \quad x \in \psi(F_x(A, \varepsilon)), \quad (11)$$

where $\psi(F_x(A, \varepsilon))$ is the set of points of increase of $F_x(A, \varepsilon)$.

The following result provides a partial characterization of the capacity-achieving distribution for the DTP channel when the peak power constraint is relaxed (i.e., $A \to \infty$ and thus only an average power constraint is imposed).

Theorem 6 (Insufficiency of distributions with bounded support under average power constraint)**.** *Distributions with bounded support are not capacity-achieving for the DTP channel under only an average power constraint.*

Proof: Suppose instead that the capacity-achieving distribution dF_x^* under average power constraint ε has a bounded support $\Omega \subseteq [0, A^*]$. It follows from Theorem 5 that

$$M(\mu, x, F_x^*) \geq 0, \quad x \in [0, A], \quad (12)$$
$$M(\mu, x, F_x^*) = 0, \quad x \in \psi(F_x^*),$$

for any $A \geq A^*$.

Suppose x_1 and x_2 are two arbitrary points of increase of F_x^* with $x_1 < x_2$. In view of the fact that $M(\mu, x_1, F_x^*) = M(\mu, x_2, F_x^*) = 0$, we can rewrite the multiplier function as

$$\begin{aligned} &M(\mu, x, F_x^*) \\ &= \frac{x - x_1}{x_2 - x_1}\left(i(x_2, F_x^*) - i(x_1, F_x^*)\right) + i(x_1, F_x^*) - i(x, F_x^*). \end{aligned}$$

Note that

$$\begin{aligned} &M(\mu, A, F_x^*) \\ &= \frac{A - x_1}{x_2 - x_1}\left(i(x_2, F_x^*) - i(x_1, F_x^*)\right) + i(x_1, F_x^*) - i(A, F_x^*) \\ &= \frac{A - x_1}{x_2 - x_1}\left(i(x_2, F_x^*) - i(x_1, F_x^*)\right) + i(x_1, F_x^*) \\ &= -(A+\lambda)\log(A+\lambda) + A \\ &\quad + \sum_{y=0}^{\infty} e^{-(A+\lambda)} \frac{(A+\lambda)^y}{y!} \log\left(\int_0^A e^{-x}(x+\lambda)^y dF_x^*\right) \\ &\leq \frac{A - x_1}{x_2 - x_1}\left(i(x_2, F_x^*) - i(x_1, F_x^*)\right) + i(x_1, F_x^*) \\ &\quad - (A+\lambda)\log(A+\lambda) + A \\ &\quad + \sum_{y=0}^{\infty} e^{-(A+\lambda)} \frac{(A+\lambda)^y}{y!} \log\left((A^*+\lambda)^y\right) \quad (13) \\ &= \frac{A - x_1}{x_2 - x_1}\left(i(x_2, F_x^*) - i(x_1, F_x^*)\right) + i(x_1, F_x^*) \\ &\quad - (A+\lambda)\log(A+\lambda) + A + (A+\lambda)\log(A^*+\lambda), \quad (14) \end{aligned}$$

where (13) is due to the fact that the support of dF_x^* is contained in $[0, A^*]$. It can be seen from (14) that $M(\mu, A, F_x^*) < 0$ when A is large enough since the term $-A\log A$ prevails ($i(x_2, F_x^*)$ and $i(x_1, F_x^*)$ can be viewed as constants), which is contradictory with (12). Thus, under solely an average power constraint, distributions with bounded support are not capacity-achieving. ■

It was shown by Shamai in [9] that, with peak power constraint and with or without average power constraint, the capacity-achieving distribution for the DTP channel must consist of a finite number of mass points. Theorem 6 indicates that this conclusion does not hold if the peak power constraint is relaxed.

IV. Numerical Results for Capacity-Achieving and Capacity-Approaching Distributions

Although some general properties of capacity-achieving distributions for the DTP are known, closed-form analytical expressions remain unknown in general. Therefore, it is of practical importance to develop numerical methods to compute such distributions. The numerical computation of both the channel capacity and the optimal distribution are useful for system implementation such as code design. These numerical results, as well as tight lower and upper bounds, serve as guidelines for system design.

A. Deterministic annealing algorithm for DTP capacity computation

A particle method was developed in [16] to compute the capacity of the DTP channel under average and peak power constraints as well as the associated optimal input distribution. This method is computationally intensive as it requires an initial discrete distribution with large enough cardinality (usually hundreds) to ensure convergence. In this paper, we propose a more computationally efficient algorithm based on the *deterministic annealing* (DA) method. The DA method was originally developed to compute the rate-distortion function

Input: A, λ
Output: A discretized segment of the capacity curve
Initialization (typical values): $s = 1$, $n = 2$, $x_1 = 0$, $x_2 = 0.01$, $p(x_1) = p(x_2) = 0.5$, $\omega = 10^{-5}$, $\delta = 10^{-2}$, $N_{BA} = N_{GD} = 20$;
repeat
  Initialize: $k = 1$, $\varepsilon^{(0)} = 0$;
  repeat
    $k = k + 1$;
    [Blahut-Arimoto] **for** N_{BA} *iterations* **do**
      For every mass point, update $p(x_i)$ according to rules:

$$Q^{(k)}(x_i|y_j) = \frac{p(x_i^{(k-1)})P(y_j|x_i^{(k-1)})}{\sum_{i'} p(x_{i'}^{(k-1)})P(y_j|x_{i'}^{(k-1)})}, \tag{15}$$

$$p^{(k)}(x_i) = \frac{\exp\left(\sum_j P(y_j|x_i^{(k-1)}) \log Q(x_i^{(k-1)}|y_j) - s x_i^{(k-1)}\right)}{\sum_i \exp\left(\sum_j P(y_j|x_i^{(k-1)}) \log Q(x_i^{(k-1)}|y_j) - s x_i^{(k-1)}\right)}, \tag{16}$$

$$p^{(k)}(y_j) = \sum_i p(x_i^{(k-1)})P(y_j|x_i^{(k-1)}), \tag{17}$$

    end
    [Gradient Descent] **for** N_{GD} *iterations* **do**
      Update all x_i with appropriate choice for step-size θ_k

$$x_i^{(k)} = x_i^{(k-1)} + \theta_k \frac{\partial}{\partial x_i} \left(\sum_y P_{Y|X}(y|x_i) \log Q(x_i|y) - s x_i \right) \Bigg|_{x_i = x_i^{(k-1)}} \tag{18}$$

    end
    Compute $\varepsilon^{(k)} = \sum_i x_i^{(k)} p(x_i^{(k)})$;
  until $|\varepsilon^{(k)} - \varepsilon^{(k-1)}| \le \omega \vee k \ge 100$;

$$C = \sum_i p(x_i) \sum_j P(y_j|x_i) \log \left(\frac{P(y_j|x_i)}{p(y_j)} \right).$$

  Apply the "two-symbols-one-location strategy" ;
  Re-initialize: $s = (1 - \delta)s$;
until $s \le 10^{-5}$;

Algorithm 1: The deterministic annealing algorithm.

[14], which provides a strong motivation for our work due to the similarity between the computation of channel capacity and rate-distortion function. Unlike the particle method which only generates a single point on the capacity curve, the DA method results in a discretized segment of the capacity curve.

1) The deterministic annealing & gradient decent algorithm: According to [17, Corollary 9], a parametric expression of channel capacity in terms of s for a given A is given by

$$C(\varepsilon) = s\varepsilon + \max_{Q(x|y)} \sum_x \exp\left(\sum_y P_{Y|X}(y|x) \log Q(x|y) - s c(x) \right), \tag{19}$$

where $C(\varepsilon)$ is some point on the capacity curve parameterized by s, $Q(x|y)$ is the conditional distribution of $X = x$ given $Y = y$, with $c(x)$ being the cost associated with symbol x. For the DTP channel, we have $c(x) = x$. The parameter s can be interpreted as the slope of the capacity curve at a given average power, ε and a fixed peak amplitude and has a similar physical meaning as β in [14].

Define the cost function κ as

$$\kappa \triangleq \max_{Q(x|y)} \sum_x \exp\left(\sum_y P_{Y|X}(y|x) \log Q(x|y) - s x \right). \tag{20}$$

In contrast with the case of computing the rate-distortion function where the cost function is minimized [14], the cost function (20) is maximized in the computation of channel capacity.

Algorithm 1 presents a sketch of the DA algorithm. The inputs to the algorithm are A and λ. The algorithm initializes with a binary distribution ($n = 2$) and s set to some positive value. Through the annealing process, the value of s decreases to 0 gradually. For every fixed s, the cost function (20) is deterministically maximized through updating the probability masses and positions of the input symbols recursively. The input probability is updated based on the Blahut-Arimoto rule (15)-(17), while the positions of mass points are updated via a gradient descent technique (18). A similar gradient descent

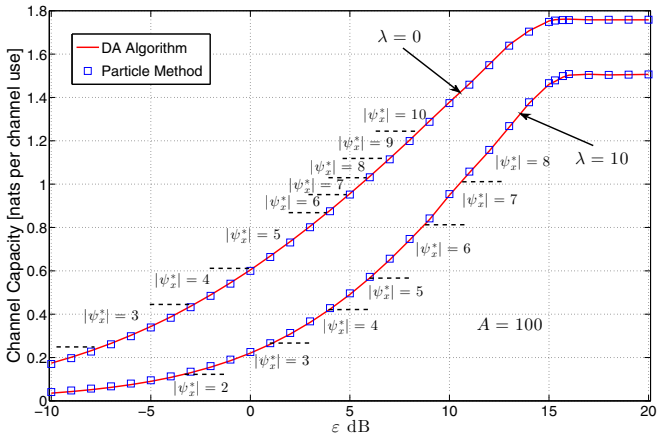


Fig. 1. Channel capacity when $A = 100$, $\lambda = 0$ and 10 with various ε. The dash lines indicate the transition points of $|\psi_x^*|$, the number of mass points in the capacity-achieving distribution.

method has been used to improve clustering performance through deterministic annealing [18] and in our previous work [16]. It should be mentioned that, although simple to implement, the gradient descent technique is not guaranteed to converge to the global optimal solution.

More mass points may be needed during the annealing process as s decreases. To change the value of n, a *two-symbols-one-location strategy* is adapted from [14, Sec. VI]. In this technique two mass points are assigned to each location, each carrying half the probability mass. During the iteration of algorithm, the points may stay merged at the same location or diverge depending on the phase transition condition. Constellation points that approach to within a small distance of each other (e.g. 10^{-3}) are merged into a two co-located mass points with the same total probability. Notice that some of the duplicate mass points may also diverge away from each other during the gradient descent phase. For example, assume for a given s, $n = 2$. However, if the corresponding optimal distribution is ternary, then one of the mass points will diverge away from its pair during step (18). The algorithm will then store 6 points before next iteration. For further details on this technique, the reader is referred to [14].

The result of the DA algorithm has not been proven to necessarily converge to the global optimum (i.e., capacity-achieving) distribution. Convergence of the Blahut-Arimoto algorithm is guaranteed and the convergence of the gradient descent algorithm to a local optimum can be realized through the proper selection of the step size θ_k in (18) (e.g., according to the Armijo rule [15]). Although not necessarily optimum, the output of the DA algorithm can be tested for optimality against the KKT conditions in Theorem 5. In the numerical results that follow, in practice all of the outputs of the DA algorithm satisfy Theorem 5 are thus *capacity-achieving*.

B. Simulation results using DA

Figure 1 plots the channel capacity curves for increasing ε, with the peak power constraint $A = 100$ and for $\lambda = 0$ and $\lambda = 10$ with $\delta = 0.01$. For comparison, the results obtained via the particle method [16] are also presented. From the simulation results, it is apparent that both of these algorithms yield the channel capacity as well as the capacity-achieving distribution. Notice, however, that the DA algorithm generates a segment of the capacity curve rather than discrete points.

Consider the case of $A = 100$ and $\lambda = 10$ in Fig. 1 where $|\psi_x^*|$ denotes the total number of the mass points in the capacity-achieving distribution. For $\varepsilon < -4$ dB, the optimal input distribution is binary and as ε increases, so too does $|\psi_x^*|$. However, for $\varepsilon > 11$ dB, the capacity-achieving distribution has $|\psi_x^*| = 8$ points. Notice also that for $\varepsilon > 16$ dB, both of the capacity-achieving distribution and the channel capacity do not change with increasing ε. Neither the capacity-achieving distribution nor the channel capacity changes with increasing ε. This saturation in capacity with increasing ε indicates that the average power constraint becomes ineffective and the capacity is limited by the peak power constraint. Notice that this phenomenon occurs for $\varepsilon >\approx A/2 = 17$ dB.

The case of peak power constraint only and $\lambda = 0$ was treated by Shamai in [9, Eq. (14)], where it is claimed that the capacity-achieving distribution is

$$dF_x^* = (1 - \beta_1 - \beta_2)\delta(x) + \beta_1\delta(x - 0.3839A) + \beta_2\delta(x - A) \tag{21}$$

in the region $3.3679 \leq A < \phi$, for some $\phi > 3.3679$. In other words, the constellation of the capacity-achieving distribution scales linearly with A. However, after extensive simulation, we have found that for ternary capacity-achieving distributions, (21) is only capacity-achieving with $A = 3.3679$ but not for larger values of A. In particular, for $\lambda = 0$, $A = 4$ and only peak power constraint, the capacity-achieving distribution is

$$dF_x^* = 0.4927\delta(x) + 0.0768\delta(x - 1.4033) + 0.4305\delta(x - 4),$$

where the middle point is not located at $x = 4 \times 0.3839 = 1.5356$.

C. Example: Inactive peak power constraint

It is easy to see that the average power constraint is inactive if it is greater than the mean of the capacity-achieving distribution under the peak power constraint only. However, it is natural to expect that the peak power constraint is always active since separating constellation points maximally should improve performance. Somewhat surprisingly, we shall show that the peak power constraint can be inactive (i.e., $A \notin \psi_x^*$) in some cases. It should be pointed out that $A \notin \psi_x^*$ does not mean the peak power constraint is superfluous. Indeed, according to Theorem 6, distributions with bounded support are not capacity-achieving if the peak power constraint is removed.

Figure 2 plots the capacity-achieving distributions for $\lambda = 0$, $\varepsilon = 0.0594439$ and various A. In both sub-figures, the dashed curve represents the peak power constraint A while the dots represent the positions of mass points in the optimal input distributions. All simulations are carried out using the numerical technique developed in Sec. IV-A. In Fig. 2(a) the capacity-achieving distributions are plotted for different peak-to-average ratios, A/ε with ε and λ fixed. Fig. 2(b) highlights how the positions of the mass points in F_x^* evolve with increasing A. Notice that, when $A \leq 1$, the peak power constraint is active and F_x^* has a mass point at A. When

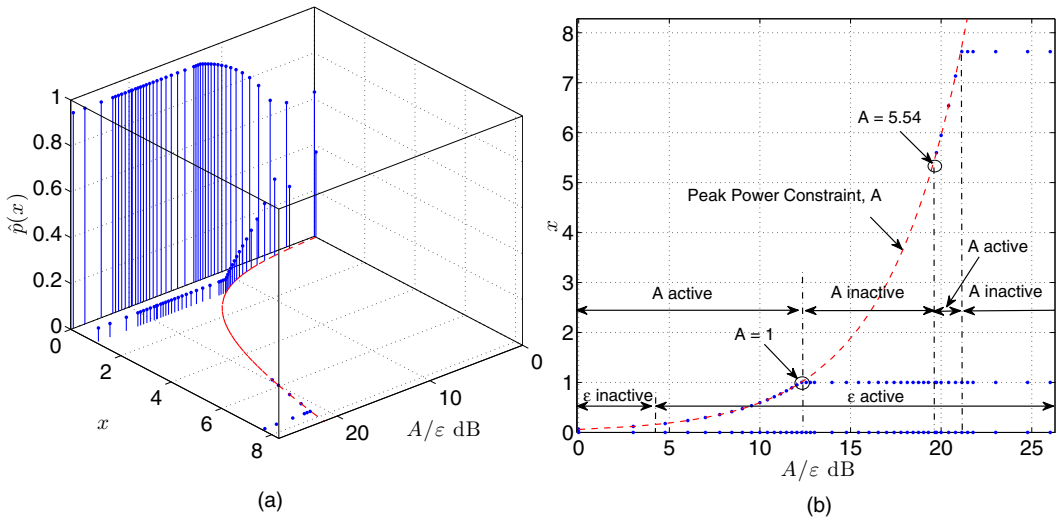


Fig. 2. Capacity-achieving distributions when $\lambda = 0$ and $\varepsilon = 0.0594439$ with various A: (a) distributions and (b) positions of mass points.

$1 < A \leq 5.54$, as noted earlier, there is slack in the peak power constraint and $\psi_x^* = \{0, 1\}$. Once $A > 5.54$, the capacity achieving distribution is ternary and again the peak power constraint is active. Thus, in this example, we observe that for a fixed ε and λ the peak power constraint alternates between being active and inactive as A increases.

To investigate further, consider the specific case for the same λ and ε as above and $A = 50\varepsilon = 2.972195$. Using DA from Sec. IV-A, the capacity-achieving distribution is found to be binary and of the form

$$dF_x^* = (1-\varepsilon)\delta(x) + \varepsilon\delta(x-1). \tag{22}$$

Thus, the average constraint is active while there is slack in the peak constraint. Further, the multiplier function in (9) is computed to be

$$M(\mu, x, F_x^*) = 3.7844e^{-x} - x\log x + 2.392x - 3.7844, \tag{23}$$

and plotted in Fig. 3. Considering Theorem 5, it is clear from Fig. 3 that dF_x^* in (22) is in fact capacity-achieving for *any* $A \in (1, 5.54]$ and the peak power constraint is inactive in this range.

D. Maxentropic capacity-approaching distributions

Although there are some algorithms to compute both the capacity and optimal input distributions for DTP channels (e.g., Section IV-A), often it is instructive to develop simple closed-form input distributions which give a tight lower bounds on the capacity.

Define a family of distributions, termed *maxentropic*, which have equally spaced mass points in $[0, A]$ and which maximize entropy subject to average and peak power constraints. In particular, the maxentropic distributions are given as [12]

$$dF_x^\dagger(K) = \sum_{k=0}^{K} \bar{p}_k \delta\left(x - k\frac{A}{K}\right),$$

where

$$\bar{p}_k = \frac{1}{K+1}, \qquad A \leq 2\varepsilon, \tag{24}$$

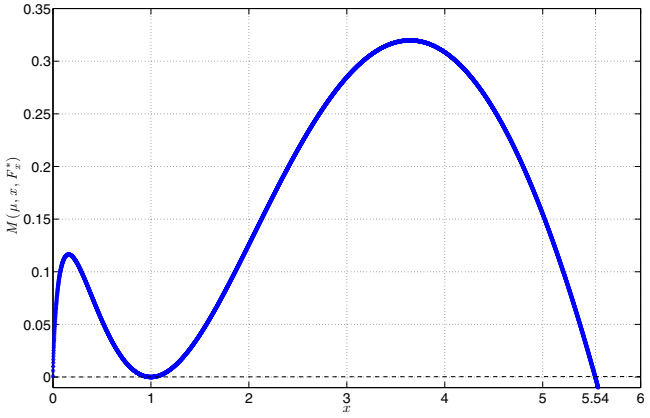


Fig. 3. Multiplier function (23) for $\varepsilon = 0.0594439$ and $\lambda = 0$.

$$\bar{p}_k = \frac{t^k}{1 + t + t^2 + \ldots + t^K}, \qquad A \geq 2\varepsilon, \tag{25}$$

and t is some number between 0 and 1. These distributions are in fact *Boltzmann distributions* which are widely applied in statistical physics and clustering [19] and have also been applied in Gaussian noise-corrupted optical channels [12].

E. Performance of Maxentropic Distributions on DTP Channels

The information rates induced by the maxentropic distributions are shown in Figs. 4 and 5 for different A/ε and $\lambda = 3$. The channel capacity in both cases are included for comparison and computed through the deterministic annealing method of Section IV-A.

It can seen from the figures that the performance of the maxentropic distributions is very close to the channel capacity and even capacity-achieving in some cases. At low ε, the binary maxentropic distribution

$$dF_x^\dagger(1) = \left(1 - \frac{\varepsilon}{A}\right)\delta(x) + \frac{\varepsilon}{A}\delta(x - A) \tag{26}$$

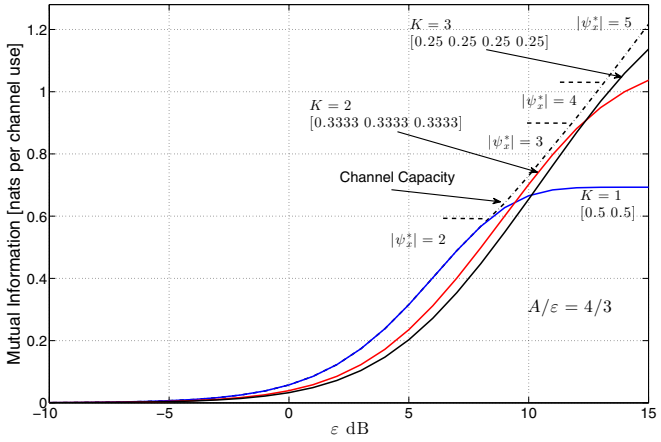


Fig. 4. Mutual information induced by the maxentropic input distributions and the number of mass points versus ε when $A/\varepsilon = 4/3$ and $\lambda = 3$. For comparison, the channel capacity and the number of mass points in the capacity-achieving distribution, $|\psi_x^*|$, are also provided.

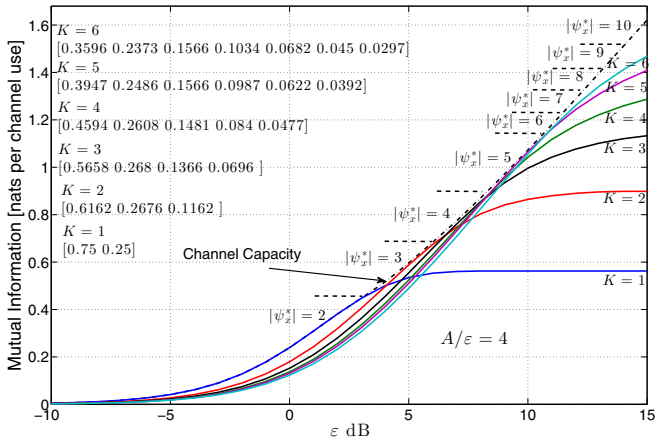


Fig. 5. Mutual information induced by the maxentropic input distributions and the number of mass points versus ε when $A/\varepsilon = 4$ and $\lambda = 3$. For comparison, the channel capacity and the number of mass points in the capacity-achieving distribution, $|\psi_x^*|$, are also provided.

is capacity achieving in both figures. As ε increases, the number of mass points, K, required to approach the capacity also increases. In fact, the mutual information induced by the maxentropic distributions is often negligibly far from the capacity for high ε with less mass points than the optimal case. For example, in Fig. 4 when $\varepsilon > 12.5$ dB, a $K = 3$ maxentropic distribution is close to the channel capacity while $|\psi_x^*| = 5$ points are in the optimal distribution. Similarly, in Fig. 5, a $K = 6$ maxentropic distribution approaches the capacity where up to 8 mass points are needed when $\varepsilon > 12.5$ dB.

Therefore, the envelope of the information rates of maxentropic distributions forms a close approximation to the channel capacity of DTP channels. The simple forms of the maxentropic distributions also make them a useful first step in non-uniform signalling design for DTP channels and often has the practical benefit of having near-capacity performance with less mass points.

V. Conclusions

The DTP channel is a good model for many interesting channels, including long range space optical systems. This paper provides insight into the capacity-achieving and approaching distributions for DTP channels, including some fundamental properties of the capacity-achieving distributions. In addition, a numerical algorithm is presented based on deterministic annealing which generates a segment of the capacity curve and the associated optimal input distributions. A simple family of maxentropic input distributions is defined and used to develop tight lower bounds on the channel capacity through the evaluation of the envelope of information rates for different K.

An interesting insight of this work is that at low input powers, when the received power is smaller or at most on the same order as the the dark current, that binary inputs are often optimal. Indeed, our simulations have shown that at low input powers the binary maxentropic distribution in (26) is optimal over a wide range of A. In the companion to this paper [20], we expand our analytical results to develop necessary and sufficient conditions for binary inputs being optimal as well as presenting a closed-form expression for the capacity-achieving distribution for large λ under both peak and average constraints.

Appendix A
Proof of Lemma 1

Proposition 7. *If both $X \to Y \to Z$ and $X \to Z \to Y$ form Markov chains, then $P_{Y|X}(y|x)/P_{Z|X}(z|x)$ does not depend on x.*

Proof: If both $X \to Y \to Z$ and $X \to Z \to Y$ form Markov chains, then

$$\begin{aligned} P_{Y|Z}(y|z) &= P_{Y|Z,X}(y|z,x) \\ &= \frac{P_{Y,Z|X}(y,z|x)}{P_{Z|X}(z|x)} \\ &= \frac{P_{Y|X}(y|x)P_{Z|Y}(z|y)}{P_{Z|X}(z|x)}, \end{aligned}$$

where $P_{Z|X,Y} = P_{Z|Y}$ by Markov chain definition. This implies that $P_{Y|X}(y|x)/P_{Z|X}(z|x)$ does not depend on x. ■

Now we proceed to prove Lemma 1. Let $\mathrm{Pois}(\Delta)$ denote a Poisson distributed random variable with mean Δ. Let $W \sim \mathrm{Pois}(\Delta)$ be independent of X_Δ and Y_Δ. By the data processing inequality, we have

$$I(X_\Delta; Y_\Delta + W) \le I(X_\Delta; Y_\Delta). \tag{27}$$

The conditional distribution of $Y_\Delta + W$ given $X_\Delta = x - \Delta$ is the same as that of Y given $X = x$, which implies $H(Y_\Delta + W|X_\Delta = x - \Delta) = H(Y|X = x)$ for any $x \ge \Delta$. Moreover, $Y_\Delta + W$ and Y are identically distributed; as a consequence, we have $H(Y_\Delta + W) = H(Y)$. Now one can readily show that

$$I(X_\Delta; Y_\Delta + W) = I(X; Y), \tag{28}$$

which, together with (27), yields the desired inequality.

Note that the equality in (27) holds if and only if $X_\Delta \to Y_\Delta + W \to Y_\Delta$ form a Markov chain. Since $X_\Delta \to Y_\Delta \to$

$Y_\Delta + W$ form a Markov chain, it can be shown by leveraging Proposition 7 that $X_\Delta \to Y_\Delta + W \to Y_\Delta$ also form a Markov chain if and only if $|\psi_x| = 1$. This completes the proof of Lemma 1.

APPENDIX B
PROOF OF LEMMA 3

For a DTP channel with dark current of rate λ and an input distribution specified by constellation ψ_x and probability masses ψ_p, let $I_{\lambda,\psi_x,\psi_p}$ denote the resulting mutual information between the channel input and the channel output. One can prove the following proposition by following the derivation of (28).

Proposition 8. $I_{\lambda,\psi_x,\psi_p} = I_{0,\psi_x+\lambda,\psi_p}$

Now we proceed to prove Lemma 3. It suffices to consider the case $\alpha \in (0,1)$ since the degenerate case $\alpha = 0$ is trivially true. Define $\mathrm{Binom}(y,\alpha)$ as a Binomial distribution with $y \in \mathbb{Z}^+$ trials and with success probability α in each trial.

First consider the special case $\lambda = 0$. Introduce a random variable Z such that $X \to Y \to Z$ form a Markov chain, where the conditional distribution of Z given $Y = y$ is $\mathrm{Binom}(y,\alpha)$ for all y. By the data processing inequality, we have

$$I(X;Z) \le I(X;Y). \tag{29}$$

Note that the conditional distribution of Z given $X = x$ is the same as that of Y_α given $X_\alpha = \alpha x$, which implies $H(Z|X = x) = H(Y_\alpha|X_\alpha = \alpha x)$ for any $x \ge 0$. Moreover, Z and Y_α are identically distributed; as a consequence, we have $H(Z) = H(Y_\alpha)$. Now one can readily show that

$$I(X;Z) = I(X_\alpha;Y_\alpha). \tag{30}$$

which, together with (29), implies

$$I_{0,\alpha\psi_x,\psi_p} \le I_{0,\psi_x,\psi_p}. \tag{31}$$

Note that the equality in (29) holds if and only if $X \to Z \to Y$ form a Markov chain. Since $X \to Y \to Z$ form a Markov chain, it can be shown by leveraging Proposition 7 that $X \to Z \to Y$ also form a Markov chain if and only if $|\psi_x| = 1$. Therefore, the equality in (31) holds if and only if $|\psi_x| = 1$.

For the general case $\lambda \ge 0$, it can be verified that

$$\begin{aligned} I_{\lambda,\psi_x,\psi_p} &= I_{0,\psi_x+\lambda,\psi_p} && (32)\\ &\ge I_{0,\psi_x+\frac{\lambda}{\alpha},\psi_p} && (33)\\ &\ge I_{0,\alpha\psi_x+\lambda,\psi_p}, && (34)\\ &= I_{\lambda,\alpha\psi_x,\psi_p}, && (35) \end{aligned}$$

where (32) and (35) are due to Proposition 8, (33) is due to Lemma 1, and (34) is due to (31). Clearly, $I_{\lambda,\psi_x,\psi_p} \ge I_{\lambda,\alpha\psi_x,\psi_p}$ is equivalent to the desired inequality

$$I(X;Y) \ge I(X_\alpha,Y_\alpha). \tag{36}$$

To complete the proof of Lemma 3, it suffices to show that the equality in (36) holds if and only if $|\psi_x| = 1$. The "if" part is trivially true. The "only if" part is a simple consequence of the fact that the equality in (31) holds only if $|\psi_x| = 1$.

REFERENCES

[1] R. M. Gagliardi and S. Karp, *Optical Communications*. John Wiley & Sons, 1976.

[2] R. J. McEliece, E. R. Rodemich, and A. L. Rubin, "The practical limits of photon communication," *Jet Propulsion Laboratory Deep Space Network Progress Reports*, Pasadena, CA 91103, vol. 42–55, pp. 63–67, 1979.

[3] R. J. McEliece, "Practical codes for photon communication," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 393–397, Jul. 1981.

[4] S. Shamai (Shitz), "On the capacity of a direct-detection photon channel with intertransition-constrained binary input," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1540–1550, Nov. 1991.

[5] A. Lapidoth and S. M. Moser, "On the capacity of the discrete-time Poisson channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 303–322, Jan. 2009.

[6] A. Lapidoth, J. H. Shapiro, V. Venkatesan, and L. Wang, "The discrete-time Poisson channel at low input powers," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3260–3272, Jun. 2011.

[7] A. Martinez, "Spectral efficiency of optical direct detection," *J. Opt. Soc. America B*, vol. 24, no. 4, pp. 739–749, Apr. 2007.

[8] J. G. Smith, "The information capacity of amplitude and variance-constrained scalar Gaussian channels," *Inf. Contr.*, vol. 18, pp. 203–219, 1971.

[9] S. Shamai (Shitz), "Capacity of a pulse amplitude modulated direct detection photon channel," *Proc. Inst. Elec. Eng.*, vol. 137, no. 6, pp. 424–430, Dec. 1990.

[10] S. Shamai (Shitz) and I. Bar-David, "The capacity of average and peak-power-limited quadrature Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1060–1071, Jul. 1995.

[11] 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.

[12] 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.

[13] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1837–1849, May 2008.

[14] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1939–1952, Nov. 1994.

[15] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.

[16] J. Cao, S. Hranilovic, and J. Chen, "Channel capacity and non-uniform signalling for discrete-time Poisson channels," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 5, no. 4, pp. 329–337, Apr. 2013.

[17] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[18] G. Qiu, M. R. Varley, and T. J. Terrell, "Improved clustering using deterministic annealing with a gradient descent technique," *Pattern Recognition Lett.*, vol. 15, pp. 607–610, 1994.

[19] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Phys. Rev. Lett.*, vol. 65, pp. 945–948, 1990.

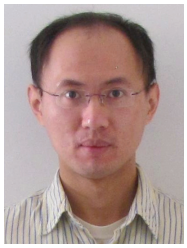
[20] J. Cao, S. Hranilovic, and J. Chen, "Capacity-achieving distributions for the discrete-time Poisson channel–part II: binary inputs," *IEEE Trans. Commun.* (TCOM-TPS-13-0143).

Jihai Cao received the B.A.Sc. and M.A.Sc. degrees with honours in electrical engineering from Harbin Institute of Technology, China in 2006 and 2008 respectively and Ph.D. from McMaster University, Canada in 2013. His field of interest includes wireless optical communication, information theory, and coding.

Steve Hranilovic (S'94-M'03-SM'07) received the B.A.Sc. degree with honours in electrical engineering from the University of Waterloo, Canada in 1997 and M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Toronto, Canada in 1999 and 2003 respectively.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he also serves as Associate Chair for undergraduate studies. During 2010-2011 he spent his research leave as Senior Member, Technical Staff in Advanced Technology for Research in Motion, Waterloo, Canada. His research interests are in the areas of free-space and wired optical communications, digital communication algorithms, and electronic and photonic implementation of coding and communication systems. He is the author of the book *Wireless Optical Communications Systems* (New York:Springer, 2004).

Dr. Hranilovic is a licensed Professional Engineer in the Province of Ontario and was awarded the Government of Ontario Early Researcher Award in 2006. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS in the area of Optical Wireless Communications.



Jun Chen (S'03-M'06) received the B.E.degree with honors in communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001 and the M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University, Ithaca, NY, in 2004 and 2006, respectively.

He was a Postdoctoral Research Associate in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign, Urbana, IL, from 2005 to 2006, and a Postdoctoral Fellow at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, from 2006 to 2007. He is currently an Associate Professor of Electrical and Computer Engineering at McMaster University, Hamilton, ON, Canada. His research interests include information theory, wireless communications, and signal processing.

He received several awards for his research, including the Josef Raviv Memorial Postdoctoral Fellowship in 2006, the Early Researcher Award from the Province of Ontario in 2010, and the IBM Faculty Award in 2010.