Capacity Approximation of Memoryless Channels with Countable Output Alphabets

Tobias Sutter*, Peyman Mohajerin Esfahani*, David Sutter†, and John Lygeros*
*Automatic Control Laboratory, ETH Zurich, Switzerland
†Institute for Theoretical Physics, ETH Zurich, Switzerland
Email: {sutter, mohajerin, lygeros}@control.ee.ethz.ch, suttedav@phys.ethz.ch

***Abstract*—(To be considered for an IEEE Jack Keil Wolf ISIT Student Paper Award.) We present a new algorithm, based on duality of convex programming and the specific structure of the channel capacity problem, to iteratively construct upper and lower bounds for the capacity of memoryless channels having continuous input and countable output alphabets. Under a mild assumption on the decay rate of the channel's tail, explicit bounds for the approximation error are provided. We demonstrate the applicability of our result on the discrete-time Poisson channel having a peak-power input constraint.**

I. Introduction

Since Shannon's seminal 1948 paper [1], channel capacities have become a basic concept in information theory. In case of finite input and output alphabets several methods for computing the capacity of a memoryless channel have been studied, among those the well known *Blahut-Arimoto algorithm* [2], [3]. Mung and Boyd presented an efficient method to derive upper bounds for the channel capacity, based on geometric programming and duality of convex programming [4]. In a companion paper, we proposed an iterative first-order method for efficiently approximating the capacity using special smoothing techniques from convex optimization, leading to explicit error bounds with fast decay rate [5].

All the mentioned methods rely on finiteness of the input and output alphabets. For memoryless channels, with continuous alphabets, it is in general not clear how to compute or approximate the capacity. One approach toward this computation is described by Huang and Meyn in [6]. Their method is based on cutting planes, where the mutual information is iteratively approximated by linear functionals. In each iteration step, an infinite dimensional linear program has to be solved, which in general is NP-hard [7, p. 16]. It has been shown that their method converges to the optimal value, however no explicit error bound is available.

Capacity of Memoryless Channels.— The capacity of a memoryless channel consisting of an input and output alphabet $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ and a channel, described by the transition density $\mathsf{Pr}[Y \in \mathrm{d}y | X = x] = W(y|x)\,\mathrm{d}y$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, is given as

$$C(W) = \sup_{p \in \mathcal{P}(\mathcal{X})} I(p, W)\,,$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability distributions on $\mathcal{X}$ and the mutual information is defined as $I(p,W) := \int_{\mathcal{X}} D(W(\cdot|x) \| (pW)(\cdot))\, p(\,\mathrm{d}x)$. The probability distribution of the channel output induced by p and W is given by $(pW)(y) := \int_{\mathcal{X}} W(y|x) p(\,\mathrm{d}x)$. The relative entropy is defined as $D(W(\cdot|x) \| (pW)(\cdot)) := \int_{\mathcal{Y}} W(y|x) \log \frac{W(y|x)}{(pW)(y)}\,\mathrm{d}y$. In most cases it is essential to introduce additional constraints on the input distribution in order to obtain physically meaningful results; more details can be found in [8, Chapter 7]. Two common input constraints are an *average-power constraint* and a *peak-power constraint*. The average-power constraint requires that $\mathsf{E}[s(X)] \leq S$ for $S \geq 0$, where $s : \mathbb{R} \to \mathbb{R}_+$ denotes some cost function. The peak-power constraint demands that with probability one we have $X \in \mathbb{A}$ for some set $\mathbb{A} \subset \mathcal{X}$. For such a setup, i.e., having average and peak-power constraints, the channel capacity is given by

$$C_{\mathbb{A},S}(W) = \sup_{p \in \mathcal{P}(\mathbb{A})} \{ I(p, W) \; : \; \mathsf{E}[s(X)] \leq S \}. \qquad (1)$$

Contributions.— In this paper, we present a new approach to solve (1) for a countable output alphabet $\mathcal{Y} \subset \mathbb{N}$ and under a mild assumption on the tail of $W(\cdot|x)$. Our method exploits the fact that the dual problem of (1) has a particular structure that allows us to use Nesterov's smoothing method [9]. In case of only a peak-power constraint, this leads to an explicit (*a priori*) error bound. In addition, the novel method provides an *a posteriori* error. This is particularly useful as oftentimes explicit error bounds are conservative in practice. It is finally shown, that the class of discrete-time Poisson channels is an example of channels satisfying the required assumptions. We demonstrate the performance of the new method for a particular example of a discrete-time Poisson channel with a peak-power constraint.

Notation.— All the logarithms in this paper are with respect to the basis 2. The natural logarithm is denoted by $\ln(\cdot)$. We consider memoryless channels having a continuous input alphabet $\mathcal{X} \subset \mathbb{R}$ and a countable output alphabet $\mathcal{Y} \subset \mathbb{N}$. Let $\mathcal{P}(\mathcal{X})$ denote the space of probability distributions on $\mathcal{X}$ and $\mathcal{D}(\mathcal{X})$ all probability densities on $\mathcal{X}$. For a probability density $p \in \mathcal{D}(\mathbb{A})$ with support $\mathbb{A} \subset \mathcal{X}$ we denote the differential entropy by $h(p) := -\int_{\mathbb{A}} p(x) \log p(x)\,\mathrm{d}x$. We denote the maximum between a and b by $a \vee b$. Let $\mathbf{1}_{\mathbb{A}}(\cdot)$ be the standard indicator function of a set $\mathbb{A}$.

II. Capacity Approximation Scheme

We consider memoryless channels with continuous input and countable output alphabets. The class of discrete-time

Poisson channels is an example of such channels with particular interest in applications, for example to model direct detection optical communication systems [10], [11]. Consider $\mathcal{X} \subset \mathbb{R}$ as the input alphabet and $\mathcal{Y} = \mathbb{N}$ as the output alphabet. The channel is described by the transition density W.

Definition 1 (Polynomial Tail)**.** The channel W features a *k-ordered polynomial tail* if for $M \in \mathbb{N}$

$$R_k(M) := \sum_{i \geq M} \big(\sup_{x \in \mathcal{X}} W(i|x) \big)^k < \infty. \tag{2}$$

The following assumptions hold throughout this section.

Assumption 1.

(i) The channel W has a k-ordered polynomial tail for some $k \in (0,1)$ in the sense of Definition 1.

(ii) The mapping $x \mapsto W(y|x)$ is continuous for any $y \in \mathbb{N}$.

Given a channel W, we introduce an *M-truncated* version of the channel by

$$W_M(i|x) := \begin{cases} W(i|x) + \frac{1}{M} \sum\limits_{j \geq M} W(j|x), & i < M \\ 0, & i \geq M. \end{cases} \tag{3}$$

W_M can be seen as a channel with input alphabet $\mathcal{X}$ and output alphabet $\{0, 1, \ldots, M-1\}$. The finiteness of the output alphabet allows us to derive an approximation scheme, inspired by [9], to numerically approximate $C(W_M)$. Hence as a first step, we opt to quantify the capacity of a channel by its M-truncated version.

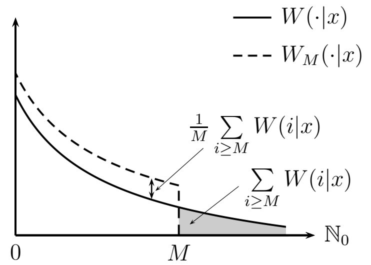

Fig. 1. Pictorial representation of the M-truncated channel counterpart

Theorem 1. *Suppose channel W satisfies Assumptions 1 with the order $k \in (0,1]$. Then, for any $M \in \mathbb{N}$ we have*

$$\big|C(W) - C(W_M)\big| \leq \frac{2\log(\mathrm{e})}{\mathrm{e}(1-k)} \Big[M^{1-k} \big(R_1(M)\big)^k + R_k(M) \Big],$$

where $R_k(M)$ is as defined in (2)*.*

Proof: The proof can be found in Appendix A-A. ■

Let $\mathcal{C}_b(\mathcal{X})$ denote the space of continuous bounded functions and $\mathbb{M}(\mathcal{X})$ the Banach space of finite signed measures on the Borel σ-algebra on $\mathcal{X}$. We define the bilinear form on $\mathbb{M}(\mathcal{X}) \times \mathcal{C}_b(\mathcal{X})$ by

$$\langle \mu, u \rangle := \int_{\mathcal{X}} u(x) \, \mathrm{d}\mu(x).$$

Consider a linear Operator $\mathcal{W} : \mathbb{R}^M \to \mathcal{C}_b(\mathcal{X})$ and its adjoint operator $\mathcal{W}^* : \mathbb{M}(\mathcal{X}) \to \mathbb{R}^M$, given by

$$\mathcal{W}\lambda(x) := \sum_{i=1}^{M} W_M(i-1|x)\lambda_i,$$
$$(\mathcal{W}^*\mu)_i := \int_X W_M(i-1|x) \, \mathrm{d}\mu(x).$$

We consider two types of input cost constraints: A peak-power constraint $\mathsf{Pr}[X \in \mathbb{A}] = 1$ for some compact set $\mathbb{A} \subset \mathcal{X}$ and an average-power constraint $\mathsf{E}[s(X)] \leq S$ for some $S \in \mathbb{R}_{\geq 0}$ and $s \in \mathcal{C}(\mathcal{X})$.

Proposition 2. *The optimization problem* (1) *is equivalent to*

$$C_{\mathbb{A},S} = \sup_{p \in \mathcal{D}(\mathbb{A})} \{ I(p, W) \ : \ \mathsf{E}[s(X)] \leq S \}. \tag{4}$$

where $\mathcal{D}(\mathbb{A})$ is the space of all probability densities on $\mathbb{A}$.

Proof: The proof can be found in Appendix A-B. ■

Define $r(\cdot) := -\sum_{j=0}^{M-1} W_M(j|\cdot) \log(W_M(j|\cdot))$, which is an element in $\mathcal{C}_b(\mathcal{X})$ by Assumption 1.

Lemma 3. *Let $S_{\max} := \max\limits_{p \in \mathcal{D}(\mathbb{A})} \mathsf{E}_p[s(X)]$. If $S \geq S_{\max}$ the optimization problem* (4) *with channel W_M is equivalent to*

$$\mathsf{P}: \quad \begin{cases} \sup\limits_{p,q} & -\langle p, r \rangle + H(q) \\ \mathrm{s.t.} & \mathcal{W}^* p = q \\ & p \in \mathcal{D}(\mathbb{A}), \ q \in \Delta_M. \end{cases}$$

If $S < S_{\max}$ the optimization problem (4) *with channel W_M is equivalent to*

$$\mathsf{P}: \quad \begin{cases} \sup\limits_{p,q} & -\langle p, r \rangle + H(q) \\ \mathrm{s.t.} & \mathcal{W}^* p = q \\ & \langle p, s \rangle = S \\ & p \in \mathcal{D}(\mathbb{A}), \ q \in \Delta_M. \end{cases} \tag{5}$$

Proof: The proof can be found in Appendix A-C. ■

Thus, problem (1) by following Proposition 2 and Lemma 3, is equivalent to (with or without additional input constraint $\langle p, s \rangle = S$, depending on S)

$$\mathsf{P}: \quad \begin{cases} \sup\limits_{p,q} & -\langle p, r \rangle + H(q) \\ \mathrm{s.t.} & \mathcal{W}^* p = q \\ & \langle p, s \rangle = S \\ & p \in \mathcal{D}(\mathbb{A}), \ q \in \Delta_M. \end{cases} \tag{6}$$

We call (6) the primal program. Its Lagrange dual program is given by

$$\mathsf{D}: \quad \inf_{\lambda \in \mathbb{R}^M} \{ G(\lambda) + F(\lambda) \}, \tag{7}$$

where

$$G(\lambda) = \begin{cases} \sup\limits_{p \in \mathcal{D}(\mathbb{A})} & \langle p, \mathcal{W}\lambda \rangle - \langle p, r \rangle \\ \mathrm{s.t.} & \langle p, s \rangle = S, \end{cases} \tag{8}$$

$$F(\lambda) = \max_{q \in \Delta_M} \{ H(q) - \lambda^\top q \}. \tag{9}$$

Lemma 4. *Strong duality holds between* (6) *and* (7)*.*

Proof: The primal program (6) clearly satisfies Slater's condition. Since it is a convex optimization problem, this implies strong duality. ■

$G(\lambda)$ is a linear program and as such non-smooth in λ in general, which prevents the dual program (7) from being solved efficiently. Therefore, we consider the smooth approximation

$$G_\nu(\lambda)=\begin{cases}\sup\limits_{p\in\mathcal{D}(\mathbb{A})} & \langle p,\mathcal{W}\lambda\rangle-\langle p,r\rangle+\nu h(p)-\nu\log\rho\\ \text{s.t.} & \langle p,s\rangle=S,\end{cases}\tag{10}$$

with smoothing parameter $\nu\in\mathbb{R}_{>0}$ and $\rho=\int_{\mathcal{X}}\mathbf{1}_{\mathbb{A}}(x)\,\mathrm{d}x$. We denote by p_ν^λ the optimizer to (10), that is unique since the objective function is strictly concave. Note that $h(p)\le\log\rho$ for all $p\in\mathcal{D}(\mathbb{A})$ and that there exists a function $\iota:\mathbb{R}_{>0}\to\mathbb{R}_{\ge 0}$ such that

$$G_\nu(\lambda)\le G(\lambda)\le G_\nu(\lambda)+\iota(\nu)\text{ for all }\lambda,\tag{11}$$

i.e., $G_\nu(\lambda)$ is a uniform approximation of the non-smooth function $G(\lambda)$. In Lemma 9, for the case of no additional input cost constraint, an explicit expression for ι is given, which implies that $\iota(\nu)\to 0$ as $\nu\to 0$. In oder to analyze (10) we consider the optimization problem

$$\begin{cases}\sup\limits_{p\in\mathcal{D}(\mathbb{A})} & h(p)+\langle p,c\rangle\\ \text{s.t.} & \langle p,s\rangle=S,\end{cases}\tag{12}$$

with $c,s\in\mathcal{C}_b(\mathcal{X})$, that has a closed form solution.

Lemma 5. *Let $p^*(x)=2^{\mu_1+c(x)+\mu_2 s(x)}$, where μ_1 and μ_2 are chosen such that p^* satisfies the constraints in* (12)*. Then p^* uniquely solves* (12)*.*

Proof: The proof can be found in Appendix A-D. ■

By considering a finite dimensional version of Lemma 5 one can find the smooth, closed form expression for $F(\lambda)$ given by

$$F(\lambda)=\log\left(\sum_{i=1}^{M}2^{-\lambda_i}\right).\tag{13}$$

Furthermore, Lemma 5 implies that $G_\nu(\lambda)$ has a (unique) analytical optimizer

$$p_\nu^\lambda(x,\mu)=2^{\mu_1+\frac{1}{\nu}(\mathcal{W}\lambda(x)-r(x))+\mu_2 s(x)},\quad x\in\mathbb{A},\tag{14}$$

where $\mu_1,\mu_2\in\mathbb{R}$ have to be chosen such that $\langle p_\nu^\lambda(\cdot,\mu),s\rangle=S$ and $p_\nu^\lambda(\cdot,\mu)\in\mathcal{D}(\mathbb{A})$. Having chosen $\mu_1,\mu_2\in\mathbb{R}$ as described we call the solution p_ν^λ.

Remark 1. In case of no input constraint $\langle p,s\rangle=S$, the unique optimizer to (10) is given by

$$p_\nu^\lambda(x)=\frac{2^{\frac{1}{\nu}(\mathcal{W}\lambda(x)-r(x))}}{\int_{\mathbb{A}}2^{\frac{1}{\nu}(\mathcal{W}\lambda(x)-r(x))}\,\mathrm{d}x},$$

whose straightforward evaluation is numerically difficult for small ν. A numerically stable technique to evaluate the above integral for small ν can be obtained by following the same lines as in [9, p. 148].

Remark 2. In case of additional input constraints, we seek for an efficient method to find the coefficients μ_i in (14). The problem of finding μ_i can be reduced to the finite dimensional convex optimization problem [12, p. 257 ff.]

$$\sup_{\mu\in\mathbb{R}^2}\left\{\langle y,\mu\rangle-\int_{\mathbb{A}}p_\nu^\lambda(x,\mu)\,\mathrm{d}x\right\},\tag{15}$$

where $y:=(1,S)$. Note that (15) is an unconstrained maximization of a concave function. However, unlike the finite input alphabet case, the evalutation of its gradient and Hessian involves computing moments of the measure $p_\nu^\lambda(x,\mu)\,\mathrm{d}x$, which we want to avoid in view of computational efficiency. In the case of having an odd number of moment constraints, there are efficient numerical schemes known, based on semidefinite programming, to compute the gradient and Hessian (see [12, p. 259 ff.] for details).

In the remainder of this paper we impose the following assumption on the channel W.

Assumption 2. $\gamma_M:=\min\limits_{x\in\mathbb{A},y<M}W_M(x|y)>0$

In case $\sum_{j\ge M}W(j|x)>0$ for all x, Assumption 2 holds according to (3) and a lower bound can be given by $\gamma_M\ge\frac{1}{M}\min_x\sum_{j\ge M}W(j|x)$.

Lemma 6. *Under Assumption 2, the dual program* (7) *is equivalent to $\min\limits_{\lambda\in Q}\{G(\lambda)+F(\lambda)\}$, where*
$Q:=\left\{\lambda\in\mathbb{R}^M\ :\ \|\lambda\|_1\le\frac{M}{2}\left(\log(\gamma_M^{-1})\vee 1\right)\right\}$.

Proof: The proof can be found in Appendix A-E. ■

We can show that the uniform approximation $G_\nu(\lambda)$ is smooth and has a Lipschitz continuous gradient, with known constant.

Theorem 7. *$G_\nu(\lambda)$ is well defined and continuously differentiable at any $\lambda\in Q$. Moreover, this function is convex and its gradient $\nabla G_\nu(\lambda)=\mathcal{W}^*p_\nu^\lambda$ is Lipschitz continuous with constant $L_\nu=\frac{1}{\nu}$.*

Proof: The proof can be found in Appendix A-F. ■

Finally, we consider the smooth, finite dimensional, convex optimization problem

$$\mathsf{D}_\nu:\quad\min_{\lambda\in Q}\{F(\lambda)+G_\nu(\lambda)\},\tag{16}$$

whose solution can be approximated with Nesterov's optimal scheme for smooth optimization [9]. Note that the objective function has a Lipschitz continuous gradient with constant $L_\nu\le 1+\frac{1}{\nu}$, according to (13) and Theorem 7. For $D_1:=$

Algorithm 1: Optimal Scheme for Smooth Optimization

For $k\ge 0$ **do**

Step 1:	Compute $\nabla F(\lambda_k)+\nabla G_\nu(\lambda_k)$
Step 2:	$y_k=-\frac{1}{L_\nu}\left(\nabla F(\lambda_k)+\nabla G_\nu(\lambda_k)\right)+\lambda_k$
Step 3:	$z_k=-\frac{1}{L_\nu}\sum_{i=0}^{k}\frac{i+1}{2}\left(\nabla F(\lambda_i)+\nabla G_\nu(\lambda_i)\right)$
Step 4:	$\lambda_{k+1}=\frac{2}{k+3}z_k+\frac{k+1}{k+3}y_k$

$\frac{1}{2}(\frac{M}{2}\log(\gamma_M^{-1}) \vee 1)^2$ we have the following result, when running Algorithm 1 on the problem (16).

Theorem 8. *Consider a smoothing parameter $\nu = \nu(n) = \frac{K}{n+1}$ for some positive constant K. Then after n iterations we can generate approximate solutions to the problems* (7) *and* (4)*, namely,*

$$\hat{\lambda} = y_n \in Q, \qquad \hat{p} = \sum_{i=0}^{n} \frac{2(i+1)}{(n+1)(n+2)} p_\nu^{\lambda_i} \in \mathcal{D}(\mathbb{A}),$$

which satisfy the following inequality:

$$\begin{aligned} 0 &\leq F(\hat{\lambda}) + G(\hat{\lambda}) - I(\hat{p}, W) \\ &\leq \iota\left(\frac{K}{n+1}\right) + \frac{4D_1}{K(n+1)} + \frac{4D_1}{(n+1)^2}. \end{aligned} \tag{17}$$

Proof: The theorem can be proven by following the proof in [9] while using Theorem 7. ■

Remark 3. By combining Theorem 1 and Theorem 8, we can quantify the approximation error of the presented method to find the capacity of any channel W, satisfying Assumptions 1 and 2, by

$$\begin{aligned} &|C(W) - C_{\text{approx}}(n)| \\ &\quad \leq |C(W_M) - C_{\text{approx}}(n)| + |C(W) - C(W_M)|, \end{aligned}$$

which can be expressed as an *a posteriori* approximation error

$$\begin{aligned} &|C(W) - C_{\text{approx}}(n)| \leq \\ &\quad \underbrace{F(\hat{\lambda}) + G(\hat{\lambda}) - I(\hat{p}, W)}_{(\star)} + \underbrace{|C(W) - C(W_M)|}_{(\star\star)}, \end{aligned}$$

where $|C(W) - C(W_M)|$ is given by Theorem 1.

Remark 4. Given a fixed number of iterations, the term $(\star)$ above is effected by the truncation level M for two reasons: the higher M the larger the size of the output as well as the lower the parameter γ_M. Therefore, term $(\star)$ increases as M increases. On the other hand, term $(\star\star)$ obviously has the opposite behavior. Namely, the higher M leads to the better approximation of the channel W by the truncated version W_M as defined in (3). Hence, given a channel W with the polynomial tail order k, there is an optimal value for the truncation parameter M, which thanks to the monotonicity explained above can be effectively computed in practice by techniques such as bisection.

A. Without Average-Power Constraint

In this subsection we discuss the setting considering only a peak-power constraint. Here our proposed methodology allows us to access a closed form expression for $G_\nu(\lambda)$ in (10) and its gradient

$$G_\nu(\lambda) = \nu \log\left(\int_{\mathbb{A}} 2^{\frac{1}{\nu}(\mathcal{W}\lambda(x) - r(x))}\,\mathrm{d}x\right) - \nu\log\rho \tag{18}$$

and

$$\nabla G_\nu(\lambda) = \frac{\int_{\mathbb{A}} 2^{\frac{1}{\nu}(\mathcal{W}\lambda(x) - r(x))} W_M(\cdot|x)\,\mathrm{d}x}{\int_{\mathbb{A}} 2^{\frac{1}{\nu}(\mathcal{W}\lambda(x) - r(x))}\,\mathrm{d}x}.$$

The following lemma gives an explicit expression for the function ι in (11). Note that $f_\lambda(\cdot)$ is Lipschitz continuous, uniformly in λ, by Assumption 2 and (3). Let L denote the Lipschitz constant.

Lemma 9. *In case of only a peak-power constraint, a possible choice of the function ι in* (11) *is given by*

$$\iota(\nu) = \begin{cases} \nu\left(\log\left(\frac{L\rho}{\nu}\right) + 1\right), & \nu < L\rho \\ \nu, & \textit{otherwise}, \end{cases}$$

where $\rho := \int_{\mathcal{X}} \mathbf{1}_{\mathbb{A}}\,\mathrm{d}x$.

Proof: The proof can be found in Appendix A-G. ■

Remark 5. Using Lemma 9, (17) provides an explicit error bound and implies that the duality gap vanishes in the limit $n \to \infty$, which shows that we approach the capacity of W_M.

Remark 6. By (11) and Theorem 8

$$\begin{aligned} 0 &\leq F(\hat{\lambda}) + G_\nu(\hat{\lambda}) + \iota(\nu) - I(\hat{p}, W) \\ &\leq \iota(\nu) + \frac{4}{n+1}\sqrt{D_1 D_2} + \frac{4D_1}{(n+1)^2}, \end{aligned} \tag{19}$$

which means that $F(\hat{\lambda}) + G_\nu(\hat{\lambda}) + \iota(\nu)$ is an upper bound for the channel capacity with *a priori* error (19). This bound can be particularly helpful in cases where an evaluation of $G(\lambda)$ for a given λ is hard.

III. Discrete-Time Poisson Channel

The discrete-time Poisson channel is a mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{N}$, such that conditioned on the input $x \geq 0$ the output is Poisson distributed with mean $x + \eta$, i.e.,

$$W(y|x) = e^{-(x+\eta)}\frac{(x+\eta)^y}{y!}, \quad y \in \mathbb{N},\ x \in \mathbb{R}_{\geq 0}, \tag{20}$$

where $\eta \geq 0$ denotes a constant sometimes referred to as *dark current*. A peak-power constraint on the transmitter is given by the peak-input constraint $X \leq A$ with probability one, i.e., $\mathbb{A} = [0, A]$ and an average-power constraint on the transmitter is considered by $\mathsf{E}[X] \leq S$.

Up to now, no analytic expression for the capacity of a discrete-time Poisson channel is known. However, for different scenarios lower and upper bounds exist. For the presence of only an average-power constraint, lower and upper bounds have been introduced in [13], [14]. In [15], for a peak-power and/or an average-power constraint, a lower bound and an asymptotic upper bound have been determined. The asymptotic upper bound contains an unknown error term that vanishes in the limit $A \to \infty$. In this section, we show how to use the methods developed in the preceding section, to approximate the capacity of a Poisson channel for the presence of only a peak-power constraint. In other words, we derive iterative lower and upper bounds that coincide when performing an infinite number of iterations and come very close for a finite number of iterations.

The following proposition provides an upper bound for the k-polynomial tail of the Poisson channel W as defined in (20).

Proposition 10 (Poisson Tail)**.** *The Poisson channel defined in* (20) *having a bounded input alphabet* $\mathcal{X} = [0, A]$ *and* dark current *parameter* η *has a* k*-polynomial tail for any* $k \in (0, 1]$ *in the sense of Definition 1, which is upper bounded for all* $M \geq A + \eta$ *by*

$$R_k(M) \leq \left(\alpha \mathrm{e}^{(\alpha-1)(A+\eta)} \frac{(A+\eta)^M}{M!}\right)^k, \quad \alpha := 2^{(k^{-1}-1)}.$$

Proof: The proof can be found in Appendix A-H. ■

We present an example to illustrate the theoretical results developed in the preceding section and their performance. First, note that for the discrete-time Poisson channel Assumption 2 clearly holds.

Example 1. Consider a discrete-time Poisson channel W as defined in (20) with a peak-power constraint A and a dark current $\eta = 3$. Up to now, the best known lower bound for the capacity is given by [15, Theorem 4]

$$C_{\mathbb{A}}(W) \geq \frac{1}{\ln 2}\left(\frac{1}{2}\ln A + \left(\frac{A}{3}+1\right)\ln\left(1+\frac{3}{A}\right) - 1 - \sqrt{\frac{\mu + \frac{1}{12}}{A}}\left(\frac{\pi}{4} + \frac{1}{2}\ln 2\right) - \frac{1}{2}\ln\frac{\pi e}{2}\right). \quad (21)$$

Due to the best of our knowledge no explicit upper bound for the capacity is known. According to Theorems 1 and 8, the algorithm introduced in this paper leads to an approximation error after n iterations that is given by

$$|C_{\text{approx}}(n) - C_{\mathbb{A}}(W)| \leq F(\hat{\lambda}) + G(\hat{\lambda}) - I(\hat{p}, W) + \mathcal{E},$$

where $\mathcal{E} = \frac{2\log(\mathrm{e})}{\mathrm{e}(1-k)}[M^{1-k}\big(R_1(M)\big)^k + R_k(M)]$, $R_\ell(M) = (\alpha \mathrm{e}^{(\alpha-1)(A+\eta)}\frac{(A+\eta)^M}{M!})^\ell$ and $\alpha := 2^{(\ell^{-1}-1)}$ for any $\ell \in (0, 1]$. This finally leads to the following upper and lower bounds for $C(W)$

$$2I(\hat{p}, W) - \left(F(\hat{\lambda}) + G(\hat{\lambda})\right) - \mathcal{E} \leq C_{\mathbb{A}}(W) \quad (22)$$
$$\leq 2\left(F(\hat{\lambda}) + G(\hat{\lambda})\right) - I(\hat{p}, W) + \mathcal{E}. \quad (23)$$

IV. Conclusion

In this paper we introduced a method to approximate the capacity for a large class of memoryless channels with a continuous input and a countable output alphabet. The key idea is to deploy the duality theory of convex programming together with a smoothing technique, which leads to entropy maximization problems whose analytical solutions are available. By invoking this favourable structure, the dual program can be seen as a finite dimensional convex optimization problem, which we solve with an efficient fast gradient method leading to explicit bounds on the approximation error.

Acknowledgment

The authors thank Renato Renner and Stefan Richter for helpful discussions and pointers to references. TS, PME and JL were supported by the HYCON2 Network of Excellence (FP7-ICT-2009-5) and the ETH grant (ETH-15 12-2). DS acknowledges support by the Swiss National Science Foundation (through the National Centre of Competence in Research 'Quantum Science and Technology' and grant No. 200020-135048) and by the European Research Council (grant No. 258932).

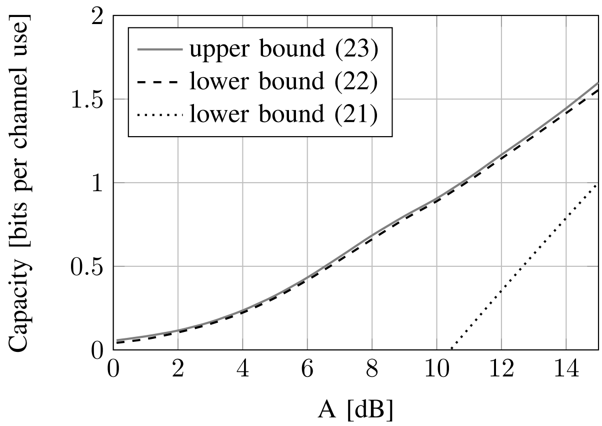

Fig. 2. This plot depicts the capacity of a discrete-time Poisson channel with dark current $\eta = 3$ as a function of the peak-power constraint parameter A.

Appendix A
Proofs

A. *Proof of Theorem 1*

To prove Theorem 1 we need a preliminary lemma.

Lemma 11. *Given* $k \in (0, 1)$ *and* $p \in [0, 1]$, *we have for all* $x \in [0, 1 - p]$

$$\big|(p + x)\log(p + x) - p\log p\big| \leq \frac{\log(\mathrm{e})}{\mathrm{e}(1-k)}x^k.$$

Proof: Note that for a fixed $x \in [0, 1]$, the mapping $p \mapsto (p+x)\log(p+x) - p\log p$ is non-decreasing; observe that the derivative of the mapping is non-negative for each $x \in [0, 1]$. Therefore, it suffices to verify the claim for $p \in \{0, 1\}$. For $p = 1$ and accordingly $x = 0$, Lemma 11 holds trivially. Let $p = 0$ and $h(x) := \frac{\log(\mathrm{e})}{\mathrm{e}(1-k)}x^{k-1} + \log x$. Note that $h(1) = \frac{\log(\mathrm{e})}{\mathrm{e}(1-k)} > 0$ and $h(x) \to \infty$ as $x \to 0$. Hence, by setting $\frac{\mathrm{d}}{\mathrm{d}x}h(x^\star) = 0$, it can be easily seen that

$$\min_{x\in(0,1]} h(x) = h(x^\star) = 0, \qquad x^\star := \mathrm{e}^{\frac{1}{k-1}}.$$

Thus $h(x) \geq 0$, and consequently $xh(x) \geq 0$ for all $x \in (0, 1]$, which concludes the proof. ■

Proof of Theorem 1: Note that

$$|C(W) - C(W_M)| = \big|\max_{p\in\mathcal{P}(\mathcal{X})} I(p, W) - \max_{p\in\mathcal{P}(\mathcal{X})} I(p, W_M)\big| \leq \max_{p\in\mathcal{P}(\mathcal{X})}\big|I(p, W) - I(p, W_M)\big|.$$

It thus suffices to bound the mutual information difference uniformly in the input probability distribution $p \in \mathcal{P}(\mathcal{X})$. Observe that

$$
\begin{aligned}
&\big|I(p,W) - I(p,W_M)\big| \\
&= \Big| \int_{\mathcal{X}} \Big[-h\big(W(\cdot,x)\big) + h\big(W_M(\cdot,x)\big)\Big] p(\,\mathrm{d}x) \\
&\quad + h\Big(\int_{\mathcal{X}} W(\cdot,x)p(\,\mathrm{d}x)\Big) - h\Big(\int_{\mathcal{X}} W_M(\cdot,x)p(\,\mathrm{d}x)\Big)\Big| \\
&= \Big| \int_{\mathcal{X}} \Big[\sum_{i\in\mathbb{N}} W(i|x)\log(W(i|x)) \\
&\quad - W_M(i|x)\log(W_M(i|x))\Big] p(\,\mathrm{d}x) \\
&\quad + \sum_{i\in\mathbb{N}} -\Big(\int_{\mathcal{X}} W(i|x)p(\,\mathrm{d}x)\Big)\log\Big(\int_{\mathcal{X}} W(i|x)p(\,\mathrm{d}x)\Big) \\
&\quad + \Big(\int_{\mathcal{X}} W_M(i|x)p(\,\mathrm{d}x)\Big)\log\Big(\int_{\mathcal{X}} W_M(i|x)p(\,\mathrm{d}x)\Big)\Big|.
\end{aligned}
$$

By the definition of the truncated channel in (3) and applying Lemma 11 to the above relation, we have

$$
\begin{aligned}
&\big|I(p,W) - I(p,W_M)\big| \\
&\le \frac{\log(\mathrm{e})}{\mathrm{e}(1-k)}\Big(\int_{\mathcal{X}} \Big[\sum_{i<M} \Big(\frac{1}{M}\sum_{j\ge M} W(j|x)\Big)^k \\
&\quad + \sum_{i\ge M} \big(W(i|x)\big)^k \Big] p(\,\mathrm{d}x) \\
&\quad + \sum_{i<M} \Big(\frac{1}{M}\sum_{j\ge M}\int_{\mathcal{X}} W(j|x)p(\,\mathrm{d}x)\Big)^k \\
&\quad + \sum_{i\ge M}\Big(\int_{\mathcal{X}} W(i|x)p(\,\mathrm{d}x)\Big)^k\Big) \\
&\le \frac{2\log(\mathrm{e})}{\mathrm{e}(1-k)}\Big(M\Big(\frac{R_1(M)}{M}\Big)^k + R_k(M)\Big),
\end{aligned}
$$

which concludes the proof. ∎

B. Proof of Proposition 2

Lemma 2 can be shown by proving that the optimization problem (1) is equivalent to

$$
C_{\mathbb{A},S} = \sup_{p\in\mathfrak{D}(\mathbb{A})} \left\{ I(p,W) \;:\; \mathsf{E}[s(X)] \le S \right\},
$$

where $\mathfrak{D}(\mathbb{A})$ is the space of probability measures on $\mathbb{A}$ that are absolutely continuous with respect to the Lebesgue measure.

It is known that the mapping $p \mapsto I(p,W)$ is weakly lower semicontinuous [16]. It then suffices to show that $\mathfrak{D}(\mathbb{A})$ is weakly dense in $\mathcal{P}(\mathbb{A})$. Let $\mathbb{B}$ be a countable dense subset of $\mathbb{A}$, and $\Delta(\mathbb{B})$ be the family of probability measures whose supports are finite subsets of $\mathbb{B}$. It is well known that $\Delta(\mathbb{B})$ is weakly dense in $\mathcal{P}(\mathbb{A})$, i.e., $\mathcal{P}(\mathbb{A}) = \overline{\Delta(\mathbb{B})}$ [17, Theorem 4, p. 237], where $\overline{\Delta}$ is the weak closure of Δ. Moreover, thanks to the Lebesgue differentiation theorem [18, Theorem 3.21, p. 98], we know that for any $b \in \mathbb{B}$ the point measure $\delta_{\{b\}} \in \Delta(\mathbb{B})$ can be arbitrarily weakly approximated by measures in $\mathfrak{D}(\mathbb{A})$, i.e., $\delta_{\{b\}} \in \overline{\mathfrak{D}(\mathbb{A})}$. Hence, we have $\Delta(\mathbb{B}) = \overline{\mathfrak{D}(\mathbb{A})}$, which in light of the preceding assertion implies $\mathcal{P}(\mathbb{A}) = \overline{\mathfrak{D}(\mathbb{A})}$. □

C. Proof of Lemma 3

In a first step, note that the mutual information $I(p,W)$ can be expressed as

$$
\begin{aligned}
I(p,W) &= \int_{\mathbb{A}} \sum_{j=0}^{M-1} W_M(j|x)p(x)\log\left(\frac{W_M(j|x)}{\int_{\mathbb{A}} W_M(j|z)p(z)\,\mathrm{d}z}\right)\mathrm{d}x \\
&= \int_{\mathbb{A}} \sum_{j=0}^{M-1} [W_M(j|x)p(x)\log W_M(j|x) \\
&\quad - W_M(j|x)p(x)\log\left(\int_{\mathbb{A}} W_M(j|z)p(z)\,\mathrm{d}z\right)\Big]\,\mathrm{d}x
\end{aligned}
$$

By adding the constraint $\int_{\mathbb{A}} p(x)W_M(j|x)\,\mathrm{d}x = q_j$ for all $j = 0,\dots,M-1$,

$$
\begin{aligned}
I(p,W) &= \int_{\mathbb{A}} \sum_{j=0}^{M-1} [W_M(j|x)p(x)\log W_M(j|x) \\
&\quad - \sum_{j=0}^{M-1} q_j \log q_j \\
&= -\langle p, r\rangle + H(q),
\end{aligned}
$$

where $p \in \mathcal{D}(\mathbb{A})$. In a second step we consider now the input constraints. Let $S_{\max} := \max_{p\in\mathcal{D}(\mathbb{A})} \mathsf{E}[s(X)]$. We can simplify the input cost constraint in optimization problem (4) as follows.

Lemma 12. *If $S \ge S_{\max}$ we can remove the input cost constraint. If $S < S_{\max}$ we can assume equality in the input cost constraint.*

Proof: By definition of $S_{\max}$ it is obvious that the input cost constraint is inactive for $S \ge S_{\max}$. In the following we assume $S < S_{\max}$ and denote C_S by $C(S)$. We first prove that $C(S)$ is a concave function. Let $S^{(1)}, S^{(2)} \ge 0$, $0 \le \lambda \le 1$ and $p^{(i)}$ being a probability mass function that achieves $C(S^{(i)})$ for $i \in \{1,2\}$. Consider the probability mass function $p^{(\lambda)} = \lambda p^{(1)} + (1-\lambda)p^{(2)}$. We can write

$$
\begin{aligned}
\mathsf{E}_{p^{(\lambda)}}[s(X)] &= \lambda\mathsf{E}_{p^{(1)}}[s(X)] + (1-\lambda)\mathsf{E}_{p^{(2)}}[s(X)] \\
&\le \lambda S^{(1)} + (1-\lambda)S^{(2)} \\
&=: S^{(\lambda)}. \qquad (24)
\end{aligned}
$$

Using the concavity of the mutual information in the input distribution, we obtain

$$
\begin{aligned}
\lambda C(S^{(1)}) + (1-\lambda)C(S^{(2)}) &= \lambda I\Big(X^{(1)},Y\Big) + (1-\lambda) I\Big(X^{(2)},Y\Big) \\
&\le I\Big(X^{(\lambda)},Y\Big) \\
&\le C(S^{(\lambda)}),
\end{aligned}
$$

where the final inequality follows by definition. Note that $S^{(\lambda)}$ is feasible has been shown in (24). Note that $C(S)$ non-decreasing in S, since enlarging S relaxes the input cost constraint. Furthermore, $C(S)$ is not constant in S in the presence of an active input-cost constraint. Thus, together with the concavity of $C(S)$ this implies that $C(S)$ is strictly increasing in S. Assume that $C(S)$ is achieved for some p^* such that $\mathsf{E}_{p^*}[s(X)] = \tilde{S} < S$. Then,

$$C(\tilde{S}) := \max_{p:\mathsf{E}[s(X)]\leq \tilde{S}} I(X,Y) \geq I(X^*,Y) = C(S),$$

which is a contradiction since $C(S)$ is strictly increasing in S. ■

D. Proof of Lemma 5

This proof is similar to the proof given in [19, Theorem 12.1.1]. Let q satisfy the constraints in (12). Then

$$\begin{aligned} J(q) &= h(q) + \langle q, c\rangle = -\int q(x)\log q(x)\,\mathrm{d}x + \langle q, c\rangle \\ &= -\int q(x)\log\left(\frac{q(x)}{p^*(x)}p^*(x)\right)\mathrm{d}x + \langle q, c\rangle \\ &= -D(q\|p^*) - \int q(x)\log p^*(x)\,\mathrm{d}x + \langle q, c\rangle \\ &\leq -\int q(x)\log p^*(x)\,\mathrm{d}x + \langle q, c\rangle \qquad (25) \\ &= -\int q(x)\left(\mu_1 + \mu_2 s(x)\right)\mathrm{d}x \qquad (26) \\ &= -\int p^*(x)\left(\mu_1 + \mu_2 s(x)\right)\mathrm{d}x\,\langle p^*, c\rangle - \langle p^*, c\rangle \qquad (27) \\ &= -\int p^*(x)\log p^*(x)\,\mathrm{d}x + \langle p^*, c\rangle = J(p^*). \end{aligned}$$

The inequality follows form the non-negativity of the relative entropy. Equality (26) follows by the definition of p^* and (27) uses the fact that both p^* and q satisfy the constraints in (12). Note that equality holds in (25) if and only if $q = p^*$. This proves the uniqueness. □

E. Proof of Lemma 6

Consider the following two convex optimization problems

$$\mathsf{P}_\beta: \quad \begin{cases} \max\limits_{p,q,\varepsilon} & -\langle p, r\rangle + H(q) - \beta\varepsilon \\ \text{s.t.} & |\mathcal{W}^* p - q| \leq \varepsilon\mathbf{1} \\ & \langle p, s\rangle = S \\ & p\in\mathcal{D}(\mathbb{A}),\ q\in\Delta_M,\ \varepsilon\in\mathbb{R}_{\geq 0} \end{cases} \quad \text{and}$$

$$\mathsf{D}_\beta: \quad \begin{cases} \min\limits_{\lambda} & F(\lambda) + G(\lambda) \\ \text{s.t.} & \|\lambda\|_1 \leq \frac{\beta}{2} \\ & \lambda\in\mathbb{R}^M, \end{cases}$$

which are duals of each other and strong duality holds as the existence of a Slater point is obviously guaranteed. Denote by $\varepsilon^*(\beta)$ the optimizer of P_β with the respective optimal value J^*_β. The main idea of the proof is to show that for a sufficiently large β, which we will quantify in the following, the optimizer $\varepsilon^*(\beta)$ of P_β is equal to zero. That is, in light of the duality relation, the constraint $\|\lambda\|_1 \leq \frac{\beta}{2}$ in D_β is inactive and as such D_β is equivalent to D. Note that for

$$J(\varepsilon) := \begin{cases} \max\limits_{p,q} & -\langle p, r\rangle + H(q) \\ \text{s.t.} & |\mathcal{W}^* p - q| \leq \varepsilon\mathbf{1} \\ & \langle p, s\rangle = S \\ & p\in\mathcal{D}(\mathbb{A}),\ q\in\Delta_M, \end{cases} \qquad (28)$$

the mapping $\varepsilon\mapsto J(\varepsilon)$, the so-called perturbation function, is concave [20, p. 268]. In the next step we write the optimization problem (28) in another equivalent form

$$J(\varepsilon) = \begin{cases} \max\limits_{p,v} & -\langle p, r\rangle + H(\mathcal{W}^* p + \varepsilon v) \\ \text{s.t.} & \|v\|_\infty \leq 1 \\ & \langle p, s\rangle = S \\ & p\in\mathcal{D}(\mathbb{A}),\ v\in\mathbb{R}^M. \end{cases} \qquad (29)$$

By using Taylor's theorem, there exists $y\in[0,\varepsilon]$ such that the entropy term in the objective function of (29) can be bounded as

$$\begin{aligned} &H(\mathcal{W}^* p + \varepsilon v) \\ &\quad = H(\mathcal{W}^* p) - \left(\log(\mathcal{W}^* p) + \mathbf{1}\right)^\top v\varepsilon \\ &\qquad - \sum_{j=1}^{M} \frac{v_j^2}{(\mathcal{W}^* p)_j + y v_j}\varepsilon^2 \\ &\quad \leq H(\mathcal{W}^* p) - \left(\log(\mathcal{W}^* p) + \mathbf{1}\right)^\top v\varepsilon + \frac{M}{\gamma}\varepsilon^2. \qquad (30) \end{aligned}$$

Thus, the optimal value of problem P_β can be expressed as

$$\begin{aligned} J^*_\beta &\leq \max_{\varepsilon}\left\{J(\varepsilon) - \beta\varepsilon\right\} \\ &\leq \max_{\varepsilon}\Big\{\max_{p,v}\big[-\langle p, r\rangle + H(\mathcal{W}^* p) \qquad (31a) \\ &\qquad - \left(\log(\mathcal{W}^* p) + \mathbf{1}\right)^\top v\varepsilon \ :\ \langle p, s\rangle = S\big] + \frac{M}{\gamma}\varepsilon^2 - \beta\varepsilon\Big\} \\ &\leq \max_{\varepsilon}\Big\{\max_{p,v}\big[-r^\top p + H(\mathcal{W}^* p) \ :\ \langle p, s\rangle = S\big] \\ &\qquad + (\rho - \beta)\varepsilon + \frac{M}{\gamma}\varepsilon^2\Big\} \qquad (31b) \\ &= J(0) + \max_{\varepsilon}\left\{(\rho-\beta)\varepsilon + \frac{M}{\gamma}\varepsilon^2\right\}, \qquad (31c) \end{aligned}$$

where $\rho = M\left(\log(\gamma^{-1})\vee 1\right)$. Note that (31a) follows from (29) and (30). The equation (31b) uses the fact that $-\left(\log(\mathcal{W}^* p) + \mathbf{1}\right)^\top v \leq M\left(\log(\gamma^{-1})\vee 1\right)$. Thus, for $\beta > \rho$ and $\varepsilon_1 = \frac{\gamma}{M}(\rho - \beta)$, we have $\max_{\varepsilon\leq\varepsilon_1}\left\{(\rho-\beta)\varepsilon + \frac{M}{\gamma}\varepsilon^2\right\} = 0$. Therefore, (31c) together with the concavity of the mapping $\varepsilon\mapsto J(\varepsilon)$ implies that $J(0)$ is the global optimum of $J(\varepsilon)$ and as such $\varepsilon^*(\beta) = 0$ for $\beta > \rho$, indicating that P_β is equivalent to P.

F. Proof of Theorem 7

To prove Theorem 7 we need a preliminary lemma.

Lemma 13. *The function $d : \mathfrak{D}(\mathbb{A}) \to \mathbb{R}_{\geq 0}$, $p\mapsto -h(p) + \log(\rho)$ as introduced in* (10) *is strongly convex with convexity parameter $\sigma = 1$.*

Proof: The proof follows the ideas of [9]. It can easily be shown that

$$\langle d''(p) \cdot g, g \rangle = \int_{\mathbb{A}} \frac{g(x)^2}{p(x)} \, \mathrm{d}x.$$

Cauchy-Schwarz then implies

$$\langle d''(p) \cdot g, g \rangle \geq \frac{\left(\int_{\mathbb{A}} g(x) \, \mathrm{d}x\right)^2}{\int_{\mathbb{A}} p(x) \, \mathrm{d}x} = \|g\|^2 .$$

■

Proof of Theorem 7: It is known according to Theorem 5.1 in [21], that $G_\nu(\lambda)$ is well defined and continuously differentiable at any $\lambda \in Q$ and that this function is convex and its gradient $\nabla G_\nu(\lambda) = \mathcal{W}^* p_\nu^\lambda$ is Lipschitz continuous with constant $L_\nu = \frac{1}{\nu} \|\mathcal{W}\|^2$, where we have also used Lemma 13. The operator norm can be simplified to

$$\begin{aligned}
\|\mathcal{W}\| &= \sup_{\lambda \in \mathbb{R}^M, p \in \mathcal{P}(\mathcal{X})} \left\{ \langle \mathcal{W}\lambda, p \rangle \ : \ \|\lambda\|_2 = 1, \|p\|_1 = 1 \right\} \\
&\leq \sup_{\lambda \in \mathbb{R}^M, p \in \mathcal{P}(\mathcal{X})} \left\{ \|\mathcal{W}^* p\|_2 \|\lambda\|_2 \ : \ \|\lambda\|_2 = 1, \|p\|_1 = 1 \right\} \\
&\leq \sup_{p \in \mathcal{P}(\mathcal{X})} \left\{ \|\mathcal{W}^* p\|_1 \ : \ \|p\|_1 = 1 \right\} \\
&= \sup_{p \in \mathcal{P}(\mathcal{X})} \left\{ \sum_{i=0}^{M-1} \int_{\mathcal{X}} W_M(i|x) p(x) \, \mathrm{d}x \ : \ \|p\|_1 = 1 \right\} \\
&= \sup_{p \in \mathcal{P}(\mathcal{X})} \left\{ \int_{\mathcal{X}} \|W_M(\cdot|x)\|_1 \, p(x) \, \mathrm{d}x \ : \ \|p\|_1 = 1 \right\} \\
&\leq \sup_{x \in \mathcal{X}} \|W_M(\cdot|x)\|_1 \\
&\leq 1.
\end{aligned}$$

■

G. Proof of Lemma 9

We start by the following definitions that simplify the proof below

$$\begin{aligned}
f_\lambda(x) &:= \mathcal{W}\lambda(x) - r(x), \\
\bar{f}_\lambda &:= \sup_{x \in \mathbb{A}} f_\lambda(x) \\
B_\varepsilon(\lambda) &:= \left\{ x \in \mathbb{A} \mid \bar{f}_\lambda - f_\lambda(x) < \varepsilon \right\}, \\
\eta_\varepsilon(\lambda) &:= \int_{B_\varepsilon(\lambda)} \mathrm{d}x.
\end{aligned}$$

Then we get the uniform lower bound

$$\eta_\varepsilon(\lambda) \leq \tfrac{\varepsilon}{L} \wedge \rho. \tag{32}$$

$$\begin{aligned}
&G(\lambda) - G_\nu(\lambda) \\
&\leq \bar{f}_\lambda - G_\nu(\lambda) \qquad \text{(33a)} \\
&= \nu \Bigg(-\log \Bigg(\int_{B_\varepsilon(\lambda)} 2^{\frac{1}{\nu}\left(f_\lambda(x) - \bar{f}_\lambda\right)} \, \mathrm{d}x \\
&\qquad + \int_{B_\varepsilon^c(\lambda)} 2^{\frac{1}{\nu}\left(f_\lambda(x) - \bar{f}_\lambda\right)} \, \mathrm{d}x \Bigg) + \log \rho \Bigg) \qquad \text{(33b)} \\
&\leq \nu \left(-\log \left(\int_{B_\varepsilon(\lambda)} 2^{\frac{1}{\nu}\left(f_\lambda(x) - \bar{f}_\lambda\right)} \, \mathrm{d}x \right) + \log \rho \right) \\
&\leq \nu \left(-\log \left(\eta_\varepsilon(\lambda) 2^{-\frac{\varepsilon}{\nu}} \right) + \log \rho \right) \qquad \text{(33c)} \\
&\leq \nu \left(-\log \left(\tfrac{\varepsilon}{L} \wedge \rho \right) + \frac{\varepsilon}{\nu} + \log \rho \right) \qquad \text{(33d)} \\
&= \nu \log \left(\frac{L\rho}{\varepsilon} \vee 1 \right) + \varepsilon,
\end{aligned}$$

where (33a) follows from (8) and (33b) is due to (18). The inequality (33c) results from the definitions of $B_\varepsilon(\lambda)$ and $\eta_\varepsilon(\lambda)$ above and (33d) is implied by (32). Finally, it can be seen that for small enough ν, the optimal choice for ε is ν, which concludes the proof.

H. Proof of Proposition 10

To prove Proposition 10, we need two lemmas.

Lemma 14. *For any $k \in (0, 1]$ and $a, b \geq 0$*

$$a^k + b^k \leq 2^{1-k}(a + b)^k.$$

Proof: Let $h(x) := 2^{1-k}(1 + x)^k - x^k$. By setting $\frac{\mathrm{d}}{\mathrm{d}x} h(x^\star) = 0$, one can easily see that $x^\star = 1$ is the minimizer of function h over the interval $[0, 1]$, i.e., $h(x) \geq h(1) = 1$ for all $x \in [0, 1]$. Suppose, without loss of generality, that $a \geq b$. By virtue of the preceding result of function h, we know that

$$1 \leq h\Big(\frac{b}{a}\Big) = 2^{1-k}\Big(1 + \frac{b}{a}\Big)^k - \Big(\frac{b}{a}\Big)^k,$$

where by multiplying a^k it readily leads to the desired assertion. ■

Lemma 15. *Let $(a_i)_{i \in \mathbb{N}}$ be a non-negative sequence of real numbers. For any $k \in (0, 1]$*

$$\sum_{i \in \mathbb{N}} a_i^k \leq \Big(\sum_{i \in \mathbb{N}} \alpha^i a_i \Big)^k, \qquad \alpha := 2^{(k^{-1} - 1)}.$$

Proof: For the proof we make use of an induction argument. Note that for any $a_1 \geq 0$ it trivially holds that $a_1^k \leq 2^{1-k} a_1^k$. We now assume that for any sequence $(a_i)_{i=1}^N \subset \mathbb{R}_{\geq 0}$ we have

$$\sum_{i=1}^N a_i^k \leq \Big(\sum_{i=1}^N 2^{(k^{-1}-1)i} a_i \Big)^k. \tag{34}$$

Let $(a_i)_{i=1}^{N+1} \subset \mathbb{R}_{\geq 0}$. Then,

$$\sum_{i=1}^{N+1} a_i^k = a_1^k + \sum_{i=2}^{N+1} a_i^k \leq a_1^k + \Big(\sum_{i=2}^{N+1} 2^{(k^{-1}-1)(i-1)} a_i \Big)^k \tag{35}$$

$$\leq 2^{1-k} \Big(a_1 + \sum_{i=2}^{N+1} 2^{(k^{-1}-1)(i-1)} a_i \Big)^k \tag{36}$$

$$= \Big(2^{(k^{-1}-1)} a_1 + \sum_{i=2}^{N+1} 2^{(k^{-1}-1)i} a_i \Big)^k$$

$$= \Big(\sum_{i=1}^{N+1} 2^{(k^{-1}-1)i} a_i \Big)^k,$$

where (35) (resp. (36)) follows from (34) (resp. Lemma 14). ■

Proof of Proposition 10: It is straightforward to see that

$$\max_{x \in [0,A]} \mathrm{e}^{-x} x^i = \mathrm{e}^{-\min\{A,i\}} \big(\min\{A,i\} \big)^i. \tag{37}$$

Moreover, based on a Taylor series expansion, it is well known that for all $M \in \mathbb{N}$ and $x \in \mathbb{R}_{\geq 0}$

$$\sum_{i \geq M} \frac{x^i}{i!} \leq \frac{\mathrm{e}^x}{M!} x^M. \tag{38}$$

Therefore, it follows that

$$R_k(M) := \sum_{i \geq M} \Big(\sup_{x \in [0,A]} \mathrm{e}^{-(x+\eta)} \frac{(x+\eta)^i}{i!} \Big)^k$$

$$\leq \sum_{i \geq M} \Big(\mathrm{e}^{-(A+\eta)} \frac{(A+\eta)^i}{i!} \Big)^k \tag{39a}$$

$$\leq \mathrm{e}^{-k(A+\eta)} \Big(\sum_{i \geq M} \alpha^{(i-M+1)} \frac{(A+\eta)^i}{i!} \Big)^k \tag{39b}$$

$$= \frac{\mathrm{e}^{-k(A+\eta)}}{\alpha^{k(M-1)}} \Big(\sum_{i \geq M} \frac{\big(\alpha(A+\eta)\big)^i}{i!} \Big)^k$$

$$\leq \frac{\mathrm{e}^{-k(A+\eta)}}{\alpha^{k(M-1)}} \Big(\frac{\mathrm{e}^{\alpha(A+\eta)}}{M!} \alpha^M (A+\eta)^M \Big)^k \tag{39c}$$

$$= \Big(\alpha \mathrm{e}^{(\alpha-1)(A+\eta)} \frac{(A+\eta)^M}{M!} \Big)^k,$$

where (39a) results from (37) and the assumption $M \geq A+\eta$, and (39b) (resp. (39c)) follows from Lemma 15 (resp. (38)). ■

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[2] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[3] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[4] C. Mung and S. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 245–258, 2004.

[5] D. Sutter, P. Mohajerin Esfahani, T. Sutter, and J. Lygeros, "Efficient approximation of discrete memoryless channel capacities." submitted to ISIT 2014.

[6] J. Huang and S. P. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2336–2351, 2005.

[7] A. Ben-Tal, L. Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton Series in Applied Mathematics, Princeton University Press, 2009.

[8] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[9] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[10] S. M. Moser, "Duality-based bounds on channel capacity," *PhD thesis, ETH Zurich*, 2005.

[11] S. Shamai, "Capacity of a pulse amplitude modulated direct detection photon channel," *IEE Proceedings on Communications, Speech and Vision*, vol. 137, no. 6, pp. 424–430, 1990.

[12] J. B. Lasserre, *Moments, Positive Polynomials and Their Applications*. Imperial College Press optimization series, Imperial College Press, 2009.

[13] D. Brady and S. Verdú, "The asymptotic capacity of the direct detection photon channel with a bandwidth constraint," in *28th Annual Allerton Conference on Communication, Control, and Computing*, pp. 691–700, October 1990.

[14] A. Martinez, "Spectral efficiency of optical direct detection," *J. Opt. Soc. Am. B*, vol. 24, pp. 739–749, Apr 2007.

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

[16] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1289–1301, 2012.

[17] P. Billingsley, *Convergence of probability measures*. Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics, Wiley, 1968.

[18] G. B. Folland, *Real analysis: modern techniques and their applications*. Pure and applied mathematics, Wiley, 1999.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, 2006.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004. Sixth printing with corrections, 2008.

[21] O. Devolder, F. Glineur, and Y. Nesterov, "Double smoothing technique for large-scale linearly constrained convex optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 702–727, 2012.