# Efficient Approximation of Discrete Memoryless Channel Capacities

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

***Abstract*—We propose an iterative method for efficiently approximating the capacity of discrete memoryless channels, possibly having additional constraints on the input distribution. Based on duality of convex programming, we derive explicit upper and lower bounds for the capacity. To find an $\varepsilon$-approximation of the capacity, in case of no additional input constraints, the presented method has a computational complexity $O(\frac{1}{\varepsilon}M^2N\sqrt{\log N})$, where $N$ and $M$ denote the input and output alphabet size, and a single iteration has a complexity $O(MN)$.**

## I. Introduction

Shannon proved in his seminal 1948 paper [1] that the channel capacity for any discrete memoryless channel (DMC) consisting of a finite input alphabet $\mathcal{X} = \{1, 2, \dots, N\}$, a finite output alphabet $\mathcal{Y} = \{1, 2, \dots, M\}$, and a conditional probability mass function $P_{Y|X}(y|x)$ denoted by $W(y|x)$ expressing the probability of observing the output symbol $y$ given the input symbol $x$, is

$$C(W) = \max_{p \in \Delta_N} I(p, W)\,,$$

where $\Delta_N := \{x \in \mathbb{R}^N : \ x \geq 0, \ \sum_{i=1}^N x_i = 1\}$ denotes the $N$-simplex and the mutual information is denoted by $I(p, W) := \sum_{x\in\mathcal{X}} p(x) D(W(\cdot|x)\|(pW)(\cdot))$. $W(y|x) = \mathbb{P}[Y = y|X = x]$ describes the channel law, $(pW)(\cdot)$ is the probability distribution of the channel output induced by $p$ and $W$, i.e., $(pW)(y) := \sum_{x\in\mathcal{X}} p(x)W(y|x)$. $D(\cdot\|\cdot)$ denotes the relative entropy that is defined as $D(W(\cdot|x)\|(pW)(\cdot)) := \sum_{y\in\mathcal{Y}} W(y|x) \log \frac{W(y|x)}{(pW)(y)}$. Shannon also showed that if there is an additional cost constraint on the input distribution of the form $\mathbb{E}[s(X)] \leq S$, the capacity is given by

$$C_S(W) = \max_{p \in \Delta_N} \{I(p, W) \ : \ \mathbb{E}[s(X)] \leq S\}. \tag{1}$$

For a few DMCs it is known that the capacity can be computed analytically, however in general there is no closed-form solution. It is therefore of interest to have an algorithm that solves (1) efficiently. Since for a fixed channel the mutual information is known to be a concave function in $p$, the optimization problem (1) is a finite dimensional convex optimization problem. Solving (1) with convex programming solvers, however, turned out to be computationally inefficient even for small alphabet sizes [2].

*Previous Work and Contributions.*— Historically one of the first attempts to numerically solve (1) is the so-called *Blahut-Arimoto algorithm* [2], [3]. It exploits the special structure of the mutual information and approximates iteratively the capacity of any DMC. Each iteration step has a computational complexity $O(MN^2)$. It was shown that this algorithm, in case of no additional input constraints has an explicit error bound (also called *a priori* error bound) of the form $|C(W) - C^{(n)}_{\text{approx}}(W)| \leq \frac{\log N}{n}$, where $n$ denotes the number of iterations [3, Corollary 1]. Hence, the overall computational complexity to find an $\varepsilon$-solution is given by $O(\frac{1}{\varepsilon}MN^2 \log N)$. As such the computational cost required for an acceptable accuracy for channels with large input alphabets can be computationally hard. This undesirable property prevents the algorithm from being useful for a large class of channels, e.g., a Rayleigh channel with a discrete input alphabet [4]. There have been several improvements of the Blahut-Arimoto algorithm [5], [6], [7], which achieve a better convergence for certain channels. However, since they all rely on the original Blahut-Arimoto algorithm they inherit its complexity per iteration step. Therefore, even with improved Blahut-Arimoto algorithms, computing the capacity for channels without favorable structure having large input alphabets is computationally expensive.

Mung and Boyd [8] presented an efficient method to derive upper bounds on the channel capacity problem, based on geometric programming. A totally different approach to solve (1) was taken by Huang and Meyn [9]. Their approach is based on cutting plane methods, where the mutual information is iteratively approximated by linear functionals. In each iteration step, a finite dimensional linear program has to be solved. It has been shown that their method converges to the optimal value, however no explicit error bound is provided.

In this article, we present a new approach to solve (1) that is based on its dual formulation. It turns out that the dual problem of (1) has a particular structure that allows us to use Nesterov's smoothing method [10]. For no input cost constraint, this leads to an explicit error bound of the order $|C(W) - C^{(n)}_{\text{approx}}(W)| \leq O(\frac{M\sqrt{\log N}}{n})$, where $n$ denotes the number of iterations and each iteration step has a computational complexity of $O(NM)$. Thus, the overall computational complexity of finding an $\varepsilon$-solution is given by $O(\frac{1}{\varepsilon}M^2N\sqrt{\log N})$. In particular for large input alphabets our method gives a considerable computational efficiency improvement compared to the Blahut-Arimoto algorithm. In addition, the novel method provides us with an *a posteriori*

error which, after having run a number of iterations, states how far the approximated solution is away from the optimal value. This is desirable as often a priori error bounds are conservative in practice.

*Structure*.— The remainder of this article is structured as follows. In Section II we reformulate the capacity problem (1) which then helps to derive its dual program in Section III. We then show how to efficiently approximate the capacity by applying smoothing techniques to the dual problem. Section IV comments on the scenario where we do not have an additional input cost constraint. We demonstrate the performance of the new method in Section V for two DMCs having large input and output alphabets. We conclude in Section VI with a summary of our work.

*Notation*.— All the logarithms in this article are with respect to the basis 2. We consider DMCs having a finite input alphabet $\mathcal{X} = \{1, 2, \ldots, N\}$ and a finite output alphabet $\mathcal{Y} = \{1, 2, \ldots, M\}$. The channel law is summarized in a matrix $\mathsf{W} \in \mathbb{R}^{N\times M}$, where $\mathsf{W}_{ij} := \mathbb{P}[Y = j|X = i] = W(j|i)$. The input and output probability mass functions are denoted by the vectors $p \in \mathbb{R}^N$ and $q \in \mathbb{R}^M$. The input cost constraint can be written as $\mathbb{E}[s(X)] = p^\top s \leq S$, where $s \in \mathbb{R}^N$ denotes the cost vector and $S \in \mathbb{R}_{\geq 0}$ is the given total cost. We define the standard $n-$simplex as $\Delta_N := \{x \in \mathbb{R}^N : \ x \geq 0, \ \sum_{i=1}^N x_i = 1\}$. For a probability mass function $p \in \Delta_N$ we denote its entropy by $H(p) := \sum_{i=1}^N -p_i \log p_i$. It is convenient to introduce an additional variable for the conditional entropy of $Y$ given $\{X = i\}$ as $r \in \mathbb{R}^N$, where $r_i = -\sum_{j=1}^M \mathsf{W}_{ij} \log \mathsf{W}_{ij}$. For two vectors $x, y \in \mathbb{R}^n$ the canonical inner product is $\langle x, y\rangle := x^\top y$. We denote the maximum between $a$ and $b$ by $a \vee b$.

## II. Preliminaries

We start by reformulating problem (1). To keep notation simple we consider a single average-input cost constraint. The extension to multiple average-input cost constraints is straightforward. In a first step, we introduce the output distribution $q \in \Delta_M$ as an additional decision variable, as done in [11], [8], [12].

**Lemma 1.** *Let* $\mathcal{F} := \arg\max_{p\in\Delta_N} I(p, W)$ *and* $S_{\max} := \min_{p\in\mathcal{F}} s^\top p$. *If* $S \geq S_{\max}$ *the optimization problem* (1) *is equivalent to*

$$\mathsf{P}: \quad \begin{cases} \max\limits_{p,q} & -r^\top p + H(q) \\ \text{s.t.} & \mathsf{W}^\top p = q \\ & p \in \Delta_N, \ q \in \Delta_M. \end{cases}$$

*If* $S < S_{\max}$ *the optimization problem* (1) *is equivalent to*

$$\mathsf{P}: \quad \begin{cases} \max\limits_{p,q} & -r^\top p + H(q) \\ \text{s.t.} & \mathsf{W}^\top p = q \\ & s^\top p = S \\ & p \in \Delta_N, \ q \in \Delta_M. \end{cases} \tag{2}$$

*Proof:* The proof given in Appendix A. ∎

We tackle this optimization problem with an approach that is based on its Lagrangian dual problem. The dual function turns out to be a non-smooth function. As such, it is known that the efficiency estimate of a black-box first-order method is of the order $O\left(\frac{1}{\varepsilon^2}\right)$ if no specific problem structure is used, where $\varepsilon$ is the desired abolute accuracy of the approximate solution in function value [13]. Our problem, has a certain structure that allows us to use Nesterov's approach of approximating non-smooth problems with smooth ones [10]. This leads to a significant efficiency improvement in the estimate of the original (non-smooth) problem, i.e., an efficiency estimate of the order $O\left(\frac{1}{\varepsilon}\right)$. This, together with the low complexity of each iteration step in the approximation scheme that uses a fast gradient method, leads to a numerical method for the channel capacity problem that has a very attractive computational complexity.

Some preliminaries are needed in order to present our approximation scheme. We begin with the following optimization problem, that has an analytical solution

$$\begin{cases} \max\limits_{p} & J(p) := H(p) - c^\top p \\ \text{s.t.} & s^\top p = S \\ & p \in \Delta_N. \end{cases} \tag{3}$$

**Lemma 2.** *Let* $p^* = [p_1^*, \ldots, p_N^*]$ *with* $p_i^* = 2^{\mu_1 - c_i + \mu_2 s_i}$, *where* $\mu_1$ *and* $\mu_2$ *are chosen such that* $p^*$ *satisfies the constraints in* (3). *Then* $p^*$ *uniquely solves* (3).

*Proof:* See Appendix B. ∎

For the channel law matrix $\mathsf{W} \in \mathbb{R}^{N\times M}$ we consider the norm

$$\|\mathsf{W}\| := \max_{\lambda\in\mathbb{R}^M, \ p\in\mathbb{R}^N} \left\{\langle \mathsf{W}\lambda, p\rangle \ : \ \|\lambda\|_2 = 1, \ \|p\|_1 = 1\right\},$$

and note that an upper bound is given by

$$\begin{aligned} \|\mathsf{W}\| &= \max_{\|p\|_1=1} \max_{\|\lambda\|_2=1} \lambda^\top \mathsf{W}^\top p \leq \max_{\|p\|_1=1} \|\mathsf{W}^\top p\|_2 \\ &\leq \max_{\|p\|_1=1} \|\mathsf{W}^\top p\|_1 = \max_{\|p\|_1=1} \|p\|_1 = 1. \end{aligned} \tag{4}$$

## III. Dual Smooth Reformualtion

Consider the convex optimizaton problem (2), whose optimal value, according to Lemma 1 is the capacity $C_S$. Our approach, having special emphasis on keeping the computational complexity low, strongly exploits the specific structure of the mentioned optimization problem and tries to proceed with analytical steps as far as possible. The Lagrange dual function for (2) is given by $G(\lambda) + F(\lambda)$, where $F, G : \mathbb{R}^M \to \mathbb{R}$ are

$$G(\lambda) = \begin{cases} \max\limits_{p} & -r^\top p + \lambda^\top \mathsf{W}^\top p \\ \text{s.t.} & s^\top p = S \\ & p \in \Delta_N \end{cases} \quad \text{and}$$

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

Note that $G(\lambda)$ is a convex and piecewise linear function and non-smooth in general. $F(\lambda)$ can be shown to be a smooth

function and has a closed form expression (6). The Lagrange dual program to (2) is

$$\mathsf{D}: \quad \min_{\lambda} \left\{ G(\lambda) + F(\lambda) \ : \ \lambda \in \mathbb{R}^M \right\}. \tag{5}$$

Note that since the coupling constraint $\mathsf{W}^\top p = q$ in the primal program (2) is affine, the set of optimal solutions to the dual program (5) is nonempty [14, Proposition 5.3.1] and as such the optimum is attained. In order to assure that the set of dual optimizers is compact and to precisely characterize its size (with respect to the one-norm), we need to impose the following assumption on the channel matrix $\mathsf{W}$, that we will maintain for the remainder of this article.

**Assumption 1.** $\gamma := \min_{i,j} \mathsf{W}_{ij} > 0$

Since for a fixed input distribution the mutual information is a convex function in the channel law and since we are in a finite dimensional setup, this implies that it is continuous in the channel law in its relative interior [15]. Hence, in case of a channel matrix having zero entries we can slightly perturb these entries without considerably changing the capacity.

**Lemma 3.** *Under Assumption 1, the dual program* (5) *is equivalent to*

$$\min_{\lambda} \left\{ G(\lambda) + F(\lambda) \ : \ \lambda \in Q \right\},$$

*where* $Q := \left\{ \lambda \in \mathbb{R}^M \ : \ \|\lambda\|_2 \leq \frac{M}{2} \left( \log(\gamma^{-1}) \vee 1 \right) \right\}$.

*Proof:* See Appendix C. ∎

For later use, we define the function $Q \ni \lambda \mapsto d_1(\lambda) := \frac{1}{2} \|\lambda\|_2^2 \in \mathbb{R}$ and the number $D_1 := \max_{\lambda} \{ d_1(\lambda) \ : \ \lambda \in Q \}$, which gives $D_1 = \frac{1}{2} \left( \frac{M}{2} \left( \log(\gamma^{-1}) \vee 1 \right) \right)^2$.

**Lemma 4.** *Strong duality holds between* (2) *and* (5).

*Proof:* The primal program (2) clearly satisfies Slater's condition. Since it is a convex optimization problem, this implies strong duality. ∎

The goal is to efficiently approximate the dual program (5), while quantifying the approximation error explicitly. Note that the optimization problem defining $F(\lambda)$ is of the form given in (3), i.e., according to Lemma 2, $F(\lambda)$ admits a unique optimizer $q^*$ with components $q_j^* = 2^{\mu - \lambda_j}$, where $\mu \in \mathbb{R}$ needs to be chosen such that $q^* \in \Delta_M$, which gives

$$\mu = -\log \left( \sum_{i=1}^M 2^{-\lambda_i} \right).$$

Therefore,

$$\begin{aligned} F(\lambda) &= \sum_{i=1}^M \left( -q_i^* \log q_i^* - \lambda_i q_i^* \right) = -\sum_{i=1}^M \mu \, 2^{\mu - \lambda_i} \\ &= -\mu \, 2^{\mu} \sum_{i=1}^M 2^{-\lambda_i} = \log \left( \sum_{i=1}^M 2^{-\lambda_i} \right), \end{aligned} \tag{6}$$

which clearly is a smooth function. We will later use its gradient that is given by the closed form expression

$$(\nabla F(\lambda))_i = \frac{-2^{-\lambda_i}}{\sum_{j=1}^M 2^{-\lambda_j}}. \tag{7}$$

The main difficulty in solving (5) efficiently is that $G(\cdot)$ is non-smooth. It however is in a particular favourable structure, which allows to use Nesterov's smoothing technique [10]. This method is based on approximating $G(\cdot)$ by a function with a Lipschitz continuous gradient and an explicitly given Lipschitz constant. The smoothing step is computationally cheap due to the particular structure of (5). The approximating function can then be minimized with a rate of convergence $O\left(\frac{1}{n^2}\right)$, where $n$ denotes the number of iterations. This finally leads to a rate of convergence of the original (non-smooth) problem of $O\left(\frac{1}{n}\right)$. In the light of [10] consider

$$G_\nu(\lambda) = \begin{cases} \max\limits_{p} \ \left\langle \mathsf{W}\lambda, p \right\rangle - r^\top p + \nu H(p) - \nu \log N \\ \text{s.t.} \quad s^\top p = S \\ \qquad p \in \Delta_N, \end{cases} \tag{8}$$

with smoothing parameter $\nu \in \mathbb{R}_{>0}$ and denote by $p_\nu(\lambda)$ the optimal solution. Note that $p_\nu(\lambda)$ is unique since the objective function is strictly concave. Clearly for any $p \in \Delta_N$, $G_\nu(\lambda)$ is a uniform approximation of the non-smooth function $G(\lambda)$, since $G_\nu(\lambda) \leq G(\lambda) \leq G_\nu(\lambda) + \nu D_2$ with $D_2 := \log(N)$. Similarly as above, by using Lemma 2 an analytical optimizer $p_\nu(\lambda)$ to (8) is given by

$$p_\nu(\lambda, \mu)_i = 2^{\mu_1 + \frac{1}{\nu}(\mathsf{W}\lambda - r)_i + \mu_2 s_i}, \tag{9}$$

where $\mu_1, \mu_2 \in \mathbb{R}$ have to be chosen such that $s^\top p_\nu(\lambda, \mu) = S$ and $p_\nu(\lambda, \mu) \in \Delta_N$. Having chosen $\mu_1, \mu_2 \in \mathbb{R}$ as described, we call the solution $p_\nu(\lambda)$.

**Remark 1.** In case of no input constraints, the unique optimizer to (8) is given by

$$p_\nu(\lambda)_i = \frac{2^{\frac{1}{\nu}(\mathsf{W}\lambda - r)_i}}{\sum_{i=1}^N 2^{\frac{1}{\nu}(\mathsf{W}\lambda - r)_i}},$$

whose straightforward evaluation is numerically difficult for small $\nu$. We present a numerically stable method for this evaluation. Define $K \in \mathbb{R}^N$, by its components

$$\begin{aligned} K_i &= \nu \log \left( p_\nu(\lambda)_i \right) \\ &= (\mathsf{W}\lambda - r)_i - \nu \log \left( \sum_{i=1}^N 2^{\frac{1}{\nu}(\mathsf{W}\lambda - r)_i} \right), \end{aligned}$$

which can be computed numerically stable using the technique in [10, p. 148]. Finally, $p_\nu(\lambda)_i = 2^{\frac{K_i}{\nu}}$.

**Remark 2.** In case of an additional input constraint, we need an efficient method to find the coefficients $\mu_1$ and $\mu_2$ in (9). In particular if there are multiple input constraints (which will lead to multiple $\mu_i$) the efficiency of the method computing them becomes important. Instead of solving a system of non-linear equations, it turns out that the $\mu_i$ can be found by

solving the following convex optimization problem [16, p. 257 ff.]

$$\sup_{\mu\in\mathbb{R}^2}\left\{\langle y,\mu\rangle-\sum_{i=1}^{N}p_\nu(\lambda,\mu)_i\right\}, \tag{10}$$

where $y:=(1,S)$. Note that (10) is an unconstrained maximization of a concave function, whose gradient and Hessian can be easily computed, which would allow us to use second-order methods, e.g., Newton's method.

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

**Proposition 5.** *$G_\nu(\lambda)$ is well defined and continuously differentiable at any $\lambda\in Q$. Moreover, this function is convex and its gradient $\nabla G_\nu(\lambda)=\mathsf{W}^\top p_\nu(\lambda)$ is Lipschitz continuous with constant $\tilde{L}_\nu\le\frac{1}{\nu}$.*

*Proof:* The proof follows directly from the proof of Theorem 1 and Lemma 3 in [10] together with (4). ■

We consider the smooth, convex optimization problem

$$\mathsf{D}_\nu:\min_{\lambda\in Q}\{F(\lambda)+G_\nu(\lambda)\}, \tag{11}$$

whose solution can be approximated with Nesterov's optimal scheme for smooth optimization [10]. Consider the following algorithm where $\pi_Q(x)$ denotes the projection operator of the set $Q$, defined in Lemma 3, with $R:=\frac{M}{2}\left(\log(\gamma^{-1})\vee 1\right)$

$$\pi_Q(x):=\begin{cases}R\frac{x}{\|x\|_2}, & \|x\|_2>R\\ x, & \text{otherwise.}\end{cases}$$

**Algorithm 1:** Optimal scheme for smooth optimization

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

**Step 1:** Compute $\nabla F(x_k)+\nabla G_\nu(x_k)$

**Step 2:** $y_k=\pi_Q\left(-\frac{1}{L_\nu}\left(\nabla F(x_k)+\nabla G_\nu(x_k)\right)+x_k\right)$

**Step 3:** $z_k=\pi_Q\left(-\frac{1}{L_\nu}\sum_{i=0}^{k}\frac{i+1}{2}\left(\nabla F(x_i)+\nabla G_\nu(x_i)\right)\right)$

**Step 4:** $x_{k+1}=\frac{2}{k+3}z_k+\frac{k+1}{k+3}y_k$

Assume that the smooth function $F$ has a Lipschitz continuous gradient with constant $K\ge 0$. It can be directly seen by (7) that $K\le 1$ and as such by invoking Proposition 5 $L_\nu\le 1+\frac{1}{\nu}$. The following theorem provides explicit error bounds for the solution of the above algorithm after $n$ iterations. Recall that $D_1=\frac{1}{2}(\frac{M}{2}\log(\gamma^{-1})\vee 1)^2$ and $D_2=\log N$.

**Theorem 6** ([10])**.** *For $n\in\mathbb{N}$ consider a smoothing parameter $\nu=\nu(n)=\frac{2}{n+1}\sqrt{\frac{D_1}{D_2}}$. Then after $n$ iterations we can generate the approximate solutions to the problems* (5) *and* (1)*, namely,*

$$\hat{\lambda}=y_n\in Q,\quad \hat{p}=\sum_{i=0}^{n}\frac{2(i+1)}{(n+1)(n+2)}p_\nu(\lambda_i)\in\Delta_N, \tag{12}$$

*which satisfy the following inequality:*

$$\begin{aligned}0&\le F(\hat{\lambda})+G(\hat{\lambda})-I(\hat{p},W)\\&\le\frac{4}{n+1}\sqrt{D_1D_2}+\frac{4D_1}{(n+1)^2}.\end{aligned} \tag{13}$$

*Thus, the complexity of finding an $\varepsilon$-solution to the problems* (5) *and* (1) *by the smoothing technique does not exceed*

$$4\sqrt{D_1D_2}\;\frac{1}{\varepsilon}+2\sqrt{\frac{D_1}{\varepsilon}}.$$

Note that Theorem 6 provides an explicit error bound given in (13), also called *a priori error*. In addition this theorem predicts an approximation to the optimal input distribution (12), i.e., the optimizer of the primal problem. Thus, by comparing the values of the primal and the dual optimization problem Theorem 6 enables us to compute an *a posteriori error* which is the difference of the dual and the primal problem, namely $F(\hat{\lambda})+G(\hat{\lambda})-I(\hat{p},W)$.

## IV. No Input Cost Constraints

In the special case of no input cost constraints, one can derive an analytical expression for $G_\nu(\lambda)$ and its gradient as

$$\begin{aligned}G_\nu(\lambda)&=\nu\log\left(\sum_{i=1}^{N}2^{\frac{1}{\nu}(\mathsf{W}\lambda-r)_i}\right)-\nu\log N\\ \nabla G_\nu(\lambda)&=\frac{1}{S(\lambda)}\sum_{i=1}^{N}2^{\frac{1}{\nu}(\mathsf{W}\lambda-r)_i}\mathsf{W}_{i,\cdot},\end{aligned} \tag{14}$$

where $S(\lambda):=\sum_{i=1}^{N}2^{\frac{1}{\nu}(\mathsf{W}\lambda-r)_i}$. In order to achieve an $\varepsilon$-precise solution the smoothing factor $\nu$ has to be chosen in the order of $\varepsilon$, according to Theorem 6. A straightforward computation of $\nabla G_\nu(\lambda)$ via (14) for a small enough $\nu$ is numerically difficult. In the light of [10, p. 148], we present a numerically stable technique for computing $\nabla G_\nu(\lambda)$. By considering the functions $\mathbb{R}^M\ni\lambda\mapsto f(\lambda)=\mathsf{W}\lambda-r\in\mathbb{R}^N$ and $\mathbb{R}^N\ni x\mapsto R_\nu(x)=\nu\log\left(\sum_{i=1}^{N}2^{\frac{x_i}{\nu}}\right)\in\mathbb{R}$ it is clear that $\nabla_\lambda R_\nu(f(\lambda))=\nabla G_\nu(\lambda)$. The basic idea is to define $\bar{f}(\lambda):=\max_{1\le i\le N}f_i(\lambda)$ and then consider a function $g:\mathbb{R}^M\to\mathbb{R}^N$ given by $g_i(\lambda)=f_i(\lambda)-\bar{f}(\lambda)$, such that all components of $g(\lambda)$ are non-positive. One can show that

$$\nabla_\lambda R_\nu(f(\lambda))=\nabla_\lambda R_\nu(g(\lambda))+\nabla\bar{f}(\lambda),$$

where the term on the right-hand side can be computed with a small numerical error.

## V. Simulation Results

This section presents two examples to illustrate the theoretical results developed in the preceding sections and their performance. All the simulations in this section are performed on a 2.3 GHz Intel Core i7 processor with 8 GB RAM.

**Example 1.** Consider a DMC with a channel matrix $\mathsf{W}\in\mathbb{R}^{N\times M}$, where $N=10000$ and $M=100$, such that $\mathsf{W}_{ij}=\frac{V_{ij}}{\sum_{j=1}^{M}V_{ij}}$, and $V_{ij}$ is chosen i.i.d. according to a uniform distribution having support $[0,1]$ for all $1\le i\le N$,

$1 \leq j \leq M$. Table I compares the performance of the Blahut-Arimoto algorithm with the algorithm introduced in this paper, which has the following a priori error bound as predicted by Theorem 6

$$C_{\text{UB}} - C \leq \frac{4}{n+1}\sqrt{D_1 D_2} + \frac{4D_1}{(n+1)^2},$$

where $n$ denotes the number of iterations, $D_1 = \frac{1}{2}(\frac{M}{2}(\log(\gamma^{-1}) \vee 1))^2$, where $\gamma$ is equal to the smallest entry in the channel matrix $\mathsf{W}$ and $D_2 = \log N$. Recall that the Blahut-Arimoto algorithm has an a priori error bound of the form $C - C_{\text{LB}} \leq \frac{\log N}{n}$ [3, Corollary 1]. As explained after Theorem 6, the new method provides additionally an a posteriori error.

TABLE I
CAPACITY OF A DMC GIVEN IN EXAMPLE 1 WITH PARAMETERS $D_1 = 8.7597 \cdot 10^5$ AND $D_2 = \log 10000$. THE A PRIORI AND A POSTERIORI ERRORS ARE DENOTED BY $\mathrm{e}_{\text{apriori}}$ AND $\mathrm{e}_{\text{apost.}}$.

| | Blahut-Arimoto Algorithm | | | | Fast-Gradient Method | | | |
|---|---|---|---|---|---|---|---|---|
| Iterations | 10 | $10^2$ | $10^3$ | $10^4$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| Time [s] | 5.3 | 52 | 528 | 5359 | 4.0 | 40 | 406 | 4263 |
| $C_{\text{UB}}$ | — | — | — | — | 0.443 | 0.443 | 0.415 | 0.409 |
| $C_{\text{LB}}$ | 0.288 | 0.391 | 0.409 | 0.409 | 0.279 | 0.300 | 0.405 | 0.409 |
| $\mathrm{e}_{\text{apriori}}$ | 1.329 | 0.133 | 0.013 | 0.001 | 17.13 | 1.400 | 0.137 | 0.001 |
| $\mathrm{e}_{\text{apost.}}$ | — | — | — | — | 0.164 | 0.143 | 0.010 | $5.5{\cdot}10^{-4}$ |

**Example 2.** Consider a DMC with a channel matrix $\mathsf{W} \in \mathbb{R}^{N \times M}$, where $N = 100000$ and $M = 10$, such that $\mathsf{W}_{ij} = \frac{V_{ij}}{\sum_{j=1}^{M} V_{ij}}$, and $V_{ij}$ is chosen i.i.d. according to a uniform distribution having support $[0, 1]$ for all $1 \leq i \leq N$ and $1 \leq j \leq M$. Table II shows the performance of the algorithm introduced in this article, which has the following a priori error bound as predicted by Theorem 6

$$C_{\text{UB}} - C \leq \frac{4}{n+1}\sqrt{D_1 D_2} + \frac{4D_1}{(n+1)^2},$$

where $n$ denotes the number of iterations, $D_1 = \frac{1}{2}(\frac{M}{2}(\log(\gamma^{-1}) \vee 1))^2$, where $\gamma$ is equal to the smallest entry in the channel matrix $\mathsf{W}$ and $D_2 = \log N$. Note that for the Blahut-Arimoto algorithm we were not able to do a single iteration as we run out of memory.

TABLE II
PERFORMANCE OF THE NEW METHOD FOR A DMC GIVEN IN EXAMPLE 2 WITH PARAMETERS $D_1 = 5.7034 \cdot 10^3$ AND $D_2 = \log 100000$. THE A PRIORI AND A POSTERIORI ERRORS ARE DENOTED BY $\mathrm{e}_{\text{apriori}}$ AND $\mathrm{e}_{\text{apost.}}$.

| Iterations | 10 | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
|---|---|---|---|---|---|---|
| Time [s] | 0.3 | 1.9 | 18.3 | 184 | 1394 | 14025 |
| $C_{\text{UB}}$ | 1.271 | 1.270 | 1.112 | 1.055 | 1.052 | 1.051 |
| $C_{\text{LB}}$ | 0.273 | 0.302 | 0.959 | 1.051 | 1.051 | 1.051 |
| $\mathrm{e}_{\text{apriori}}$ | 300.5 | 14.43 | 1.252 | 0.123 | 0.012 | 0.001 |
| $\mathrm{e}_{\text{apost.}}$ | 0.994 | 0.968 | 0.154 | $3.91{\cdot}10^{-3}$ | $3.15{\cdot}10^{-4}$ | $2.98{\cdot}10^{-5}$ |

## VI. CONCLUSION

We introduced a new approach to approximate the capacity of DMCs possibly having constraints on the input distribution. The dual problem of Shannon's capacity formula turns out to have a particular structure such that the Lagrange dual function admits a closed form solution. Applying smoothing techniques to the non-smooth dual function allows us to finally solve the dual problem efficiently. This new approach, in the case of no constraints on the input distribution, has a computational complexity per iteration step of $O(MN)$. In comparison, the Blahut-Arimoto algorithm has a computational cost of $O(MN^2)$ per iteration step. More precisely for no input power constraint, the total computational cost to find an $\varepsilon$-close solution is $O(\frac{1}{\varepsilon}M^2N\sqrt{\log N})$ for the algorithm developed in this article, whereas the Blahut-Arimoto algorithm requires $O(\frac{1}{\varepsilon}MN^2 \log N)$. We would like to emphasize that the computational cost of the smallest unit, i.e., the cost of one iteration is strictly better for the algorithm introduced in this article. As highlighted by Example 2, this can make a substantial difference especially for large input alphabets. Another strength of the new approach is that it provides an a posteriori error, i.e., after having run a certain number of iterations we can precisely estimate the actual error.

The method introduced in this article can be extended to approximate the capacity of memoryless channels having a continuos input alphabet and a countable discrete output alphabet, fulfilling a mild assumption on the decay rate of the channels tail [17].


## ACKNOWLEDGMENT

The authors thank Renato Renner, Yurii Nesterov and Stefan Richter for helpful discussions and pointers to references. 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). TS, PME and JL were supported by the HYCON2 Network of Excellence (FP7-ICT-2009-5) and the ETH grant (ETH-15 12-2).


## APPENDIX A
## PROOF OF LEMMA 1

The mutual information $I(p, W)$ can be expressed as

$$\begin{aligned} &I(p, W) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} \mathsf{W}_{ij} p_i \log\left(\frac{\mathsf{W}_{ij}}{\sum_{k=1}^{N} \mathsf{W}_{kj} p_k}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} \left[p_i \mathsf{W}_{ij} \log(\mathsf{W}_{ij}) - p_i \mathsf{W}_{ij} \log\left(\sum_{k=1}^{N} \mathsf{W}_{kj} p_k\right)\right]. \end{aligned}$$

By adding the constraint $\sum_{i=1}^{N} p_i \mathsf{W}_{ij} = q_j$ for all $j = 1, \ldots, M$,

$$\begin{aligned} I(p, W) &= \sum_{i=1}^{N}\sum_{j=1}^{M} [p_i \mathsf{W}_{ij} \log(\mathsf{W}_{ij}) - p_i \mathsf{W}_{ij} \log(q_j)] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} p_i \mathsf{W}_{ij} \log(\mathsf{W}_{ij}) - \sum_{j=1}^{M} q_j \log(q_j) \\ &= -r^\top p + H(q), \end{aligned}$$

where $p \in \Delta_N$. Since $q = \mathsf{W}^\top p$ and $\mathsf{W}^\top$ is a stochastic matrix, this implies $q \in \Delta_M$. By definition of $S_{\max}$ it is obvious that the input cost constraint $s^\top p \leq S$ is inactive for $S \geq S_{\max}$, leading to the first optimization problem in Lemma 1. It remains to show that for $S < S_{\max}$, the input constraint can be written with equality, leading to the second optimization problem in Lemma 1. In oder to keep the notation simple we define $\mathsf{C}(S) := C_S(W)$ for a fixed channel $W$. We show that $\mathsf{C}(S)$ is concave in $S$ for $S \in [0, S_{\max}]$. Let $S^{(1)}, S^{(2)} \in [0, S_{\max}]$, $0 \leq \lambda \leq 1$ and $p^{(i)}$ probability mass functions that achieve $\mathsf{C}(S^{(i)})$ for $i \in \{1, 2\}$. Consider the probability mass function $p^{(\lambda)} = \lambda p^{(1)} + (1-\lambda)p^{(2)}$. We can write

$$\begin{aligned} s^\top p^{(\lambda)} &= \lambda s^\top p^{(1)} + (1-\lambda) s^\top p^{(2)} \\ &\leq \lambda S^{(1)} + (1-\lambda) S^{(2)} \\ &=: S^{(\lambda)} \in [0, S_{\max}]. \end{aligned} \tag{15}$$

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

$$\begin{aligned} &\lambda \mathsf{C}(S^{(1)}) + (1-\lambda)\mathsf{C}(S^{(2)}) \\ &\quad = \lambda I\Big(p^{(1)}, W\Big) + (1-\lambda) I\Big(p^{(2)}, W\Big) \\ &\quad \leq I\Big(p^{(\lambda)}, W\Big) \\ &\quad \leq \mathsf{C}(S^{(\lambda)}), \end{aligned}$$

where the final inequality follows by Shannon's formula for the capacity given in (1). $\mathsf{C}(S)$ clearly is non-decreasing in $S$ since enlarging $S$ relaxes the input cost constraint. Furthermore, we show that

$$\mathsf{C}(S_{\max} - \varepsilon) < \mathsf{C}(S_{\max}), \quad \text{for all } \varepsilon > 0. \tag{16}$$

Suppose $\mathsf{C}(S_{\max} - \varepsilon) = \mathsf{C}(S_{\max})$ and denote $\mathsf{C}^\star := \max_{p \in \Delta_N} I(p, W)$. This then implies that there exists $\bar{p} \in \Delta_N$ such that $I(\bar{p}, W) = \mathsf{C}^\star$ and $s^\top \bar{p} = S_{\max} - \varepsilon$, which contradicts the definition of $S_{\max}$. Hence, the concavity of $\mathsf{C}(S)$ together with the non-decreasing property and (16) imply that $\mathsf{C}(S)$ is strictly increasing in $S$. Assume that $\mathsf{C}(S)$ is achieved for some $p^\star$ such that $s^\top p^\star = \tilde{S} < S$. Then,

$$\mathsf{C}(\tilde{S}) := \max_{p : s^\top p \leq \tilde{S}} I(p, W) \geq I(p^\star, W) = \mathsf{C}(S),$$

which is a contradiction since $\mathsf{C}(S)$ is strictly increasing in $S$ for $S \in [0, S_{\max}]$. □

## Appendix B
## Proof of Lemma 2

This proof is similar to the proof given in [18, Theorem 12.1.1]. Let $q$ satisfy the constraints in (3). Then

$$\begin{aligned} J(q) &= H(q) - c^\top q \\ &= -\sum_{i=1}^N q_i \log q_i - c^\top q \\ &= -\sum_{i=1}^N q_i \log\left(\frac{q_i}{p_i^*} p_i^*\right) - c^\top q \\ &= -D(q \| p^*) - \sum_{i=1}^N q_i \log p_i^* - c^\top q \\ &\leq -\sum_{i=1}^N q_i \log p_i^* - c^\top q \qquad (17) \\ &= -\sum_{i=1}^N q_i \left(\mu_1 + \mu_2 s_i\right) \qquad (18) \\ &= -\sum_{i=1}^N p_i^* \left(\mu_1 + \mu_2 s_i\right) - c^\top p^* + c^\top p^* \qquad (19) \\ &= -\sum_{i=1}^N p_i^* \log p_i^* - c^\top p^* = J(p^*). \end{aligned}$$

The inequality follows form the non-negativity of the relative entropy. Equality (18) follows by the definition of $p^*$ and (19) uses the fact that both $p^*$ and $q$ satisfy the constraints in (3). Note that equality holds in (17) if and only if $q = p^*$. This proves the uniqueness. □

## Appendix C
## Proof of Lemma 3

Consider the following two convex optimization problems

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

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 $\mathsf{P}_\beta$ with the respective optimal value $J_\beta^*$. The main idea of the proof is to show that for a sufficiently large $\beta$, which we will quantify in the following, the optimizer $\varepsilon^*(\beta)$ of $\mathsf{P}_\beta$ is equal to zero. That is, in light of the duality relation, the constraint $\|\lambda\|_1 \leq \frac{\beta}{2}$ in $\mathsf{D}_\beta$ is inactive and as such

$\mathsf{D}_\beta$ is equivalent to $\mathsf{D}$. Note that for

$$J(\varepsilon) := \left\{ \begin{array}{ll} \max\limits_{p,q} & -r^\top p + H(q) \\ \text{s.t.} & |\mathsf{W}^\top p - q| \leq \varepsilon \mathbf{1} \\ & s^\top p = S \\ & p \in \Delta_N,\ q \in \Delta_M \end{array} \right. , \tag{20}$$

the mapping $\varepsilon \mapsto J(\varepsilon)$, the so-called perturbation function, is concave [19, p. 268]. In the next step we write the optimization problem (20) in another equivalent form

$$J(\varepsilon) = \left\{ \begin{array}{ll} \max\limits_{p,v} & -r^\top p + H(\mathsf{W}^\top p + \varepsilon v) \\ \text{s.t.} & \|v\|_\infty \leq 1 \\ & s^\top p = S \\ & p \in \Delta_N,\ v \in \mathbb{R}^M \end{array} \right. . \tag{21}$$

By using Taylor's theorem, there exists $y \in [0, \varepsilon]$ such that the entropy term in the objective function of (21) can be bounded as

$$\begin{aligned} & H(\mathsf{W}^\top p + \varepsilon v) && (22) \\ & = H(\mathsf{W}^\top p) - (\log(\mathsf{W}^\top p) + \mathbf{1})^\top v\varepsilon - \sum_{j=1}^{M} \frac{v_j^2 \varepsilon^2}{\sum_{i=1}^{N} \mathsf{W}_{ij} p_i + y v_j} \\ & \leq H(\mathsf{W}^\top p) - (\log(\mathsf{W}^\top p) + \mathbf{1})^\top v\varepsilon + \frac{M}{\gamma}\varepsilon^2. && (23) \end{aligned}$$

Thus, the optimal value of problem $\mathsf{P}_\beta$ can be expressed as

$$\begin{aligned} J_\beta^* & \leq \max_\varepsilon \{J(\varepsilon) - \beta\varepsilon\} \\ & \leq \max_\varepsilon \left\{ \max_{p,v} \left[ -r^\top p + H(\mathsf{W}^\top p) - (\log(\mathsf{W}^\top p) + \mathbf{1})^\top v\varepsilon : \right. \right. \\ & \qquad \left. \left. s^\top p = S \right] + \frac{M}{\gamma}\varepsilon^2 - \beta\varepsilon \right\} && (24a) \\ & \leq \max_\varepsilon \left\{ \max_{p,v} \left[ -r^\top p + H(\mathsf{W}^\top p) \ :\ s^\top p = S \right] \right. \\ & \qquad \left. + (\rho - \beta)\varepsilon + \frac{M}{\gamma}\varepsilon^2 \right\} && (24b) \\ & = J(0) + \max_\varepsilon \left\{ (\rho - \beta)\varepsilon + \frac{M}{\gamma}\varepsilon^2 \right\}, && (24c) \end{aligned}$$

where $\rho = M\left(\log(\gamma^{-1}) \vee 1\right)$. Note that (24a) follows from (21) and (23). The equation (24b) uses the fact that $-(\log(\mathsf{W}^\top p) + \mathbf{1})^\top v \leq M\left(\log(\gamma^{-1}) \vee 1\right)$. Thus, for $\beta > \rho$ and $\varepsilon_1 = \frac{\gamma}{M}(\rho - \beta)$, we have $\max\limits_{\varepsilon \leq \varepsilon_1} \left\{ (\rho - \beta)\varepsilon + \frac{M}{\gamma}\varepsilon^2 \right\} = 0$. Therefore, (24c) 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 $\mathsf{P}_\beta$ is equivalent to $\mathsf{P}$ in the sense that $J_\beta^\star = J_0^\star$. By strong duality this implies that the constraint $\|\lambda\|_1 \leq \frac{\beta}{2}$ in $\mathsf{D}_\beta$ is inactive. Finally, $\|\lambda\|_2 \leq \|\lambda\|_1$ concludes the proof. □

## 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] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1290–1301, 2001.

[5] J. Sayir, "Iterating the Arimoto-Blahut algorithm for faster convergence," *Proceedings IEEE International Symposium on Information Theory (ISIT)*, p. 235, 2000.

[6] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated Blahut-Arimoto-type algorithms," *Proceedings Information Theory Workshop (ITW)*, pp. 66–70, 2004.

[7] Y. Yu, "Squeezing the Arimoto-Blahut algorithm for faster convergence," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3149–3157, 2010.

[8] 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.

[9] 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.

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

[11] A. Ben-Tal and M. Teboulle, "Extension of some results for channel capacity using a generalized information measure," *Applied Mathematics and Optimization*, vol. 17, no. 1, pp. 121–132, 1988.

[12] C. Mung, "Geometric programming for communication systems," *Foundations and Trends in Communications and Information Theory*, vol. 2, pp. 1–154, July 2005.

[13] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization, Springer, 2004.

[14] D. P. Bertsekas, *Convex Optimization Theory*. Athena Scientific optimization and computation series, Athena Scientific, 2009.

[15] T. R. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

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

[17] T. Sutter, P. Mohajerin Esfahani, D. Sutter, and J. Lygeros, "Capacity approximation of memoryless channels with countable output alphabets," *Proceedings IEEE International Symposium on Information Theory (ISIT)*, June 2014.

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

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