Convex Optimization methods for Computing Channel Capacity

Abhishek Sinha
Laboratory for Information and Decision Systems (LIDS), MIT
sinhaa@mit.edu

May 15, 2014

We consider a classical computational problem from Information Theory, namely, numerically determining the Shannon Capacity of a given discrete memoryless channel. We formulate the problem as a convex optimization problem and review a classical algorithm, namely, the Blahut-Arimoto (BA) algorithm [1] that exploits the particular structure of the problem. This algorithm is an example of an *alternating minimizing* algorithm with a guaranteed rate of convergence $\Theta(\frac{1}{k})$. Moreover, if the optimal solution is unique, this algorithm achieves an exponential rate of convergence. Then we review some recent advances made on this problem using methods of convex optimization. First, we review [2] where the authors present two related algorithms, based on natural gradient and proximal point methods respectively, that are potentially faster than the original Blahut-Arimoto algorithm. Finally, we review [4] that considers the problem from a dual perspective and presents a dual algorithm that is shown to be a *geometric program*. We then critically evaluate the relative performance of these methods on specific problems. Finally, we present some directions for further research on this interesting problem.

1 Introduction

Claude Shannon's 1948 paper [5] marked the beginning the field of mathematical study of Information and reliable transmission of Information over a noisy communication channel, known as *Information Theory*. In that paper, through some ingenious mathematical arguments, he showed that information can be *reliably* transmitted over a *noisy* communication channel if the rate of transmission of information is less than the *channel capacity*, a fundamental quantity determined by the statistical description of the channel. In particular, the paper shows the startling fact that presence of noise in a communication channel limits only the rate of communication and not the probability of error in information transmission.
In the simplest case of discrete memoryless channel (DMC), the channel capacity is expressed as a convex program with the *input probability distribution* as the optimiza-

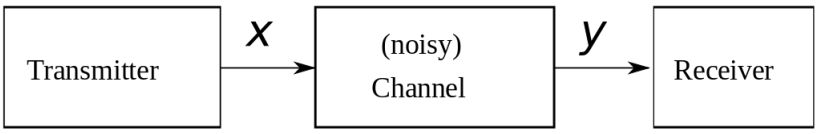

Figure 1: A Communication Channel

tion variables. Although this program can be solved explicitly in some special cases, no closed form formula is known for arbitrary DMCs. Hence one needs to resort to the techniques of convex optimization algorithms to evaluate the channel capacity of an arbitrary DMC. In this expository article, we will discuss an elegant iterative algorithm obtained by Suguru Arimoto, presented in *IEEE Transactions on Information Theory* in 1972. In this article, we will strive to provide complete proof of the key results starting from the first principles.

2 Preliminary Definitions and Results

In this section, we will define some standard Information Theoretic functionals that will be extensively used in the rest of the paper. All random variables discussed in this paper are assumed to take value from a finite set (i.e. discrete) with strictly positive probabilities and all logarithms are taken with respect to base 2, unless specified otherwise.

Definition The **Entropy** $H(X)$ of a random variable X taking value from a finite alphabet $\mathcal{X}$ with Probability Mass Function $p_X(x)$ is defined as:

$$H(X) = \mathbb{E}(-\log p_X(X)) = -\sum_{x \in \mathcal{X}} p_X(x) \log(p_X(x)) \equiv H(p_X) \tag{1}$$

Note that $H(X)$ depends only the probability measure of the random variable X and not on the particular values that X takes.

Definition The **Relative Entropy** $D(p_X||q_X)$ of two PMFs $p_X(\cdot)$ and $q_X(\cdot)$ (with $q_X(x) > 0, \forall x \in \mathcal{X}$) supported on the same alphabet space $\mathcal{X}$ is defined as:

$$D(p_X||q_X) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)} \tag{2}$$

Lemma 2.1. *For any two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ with the same support, we have*

$$D(\boldsymbol{p}||\boldsymbol{q}) \geq 0 \tag{3}$$

With equality holding iff $\boldsymbol{p} = \boldsymbol{q}$.

Proof. Although this result can be proved directly using *Jensen's inequality*, we opt to give an elementary proof here. The fundamental inequality that we use is

$$\exp(x) \geq 1 + x, \quad \forall x \in \mathbb{R} \tag{4}$$

with the equality holding iff $x = 0$.
The proof of this result follows from simple calculus. Taking natural logarithm of both sides of the above inequality, we conclude that for all $x \in \mathbb{R}$

$$\ln(x) \leq x - 1 \tag{5}$$

with the equality holding iff $x = 1$.

Now we write,

$$\begin{aligned} -D(\boldsymbol{p}||\boldsymbol{q}) &= \sum_{i=1}^{N} p_i \log \frac{q_i}{p_i} \\ &\leq \sum_{i=1}^{N} p_i (\frac{q_i}{p_i} - 1) \qquad (6) \\ &= \sum_{i=1}^{N} q_i - \sum_{i=1}^{N} p_i \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

Where inequality (6) follows from Eqn. (5). Hence we have

$$D(\boldsymbol{p}||\boldsymbol{q}) \geq 0 \tag{7}$$

Where the equality holds iff the equality holds in Eqn. 6, i.e. if $\boldsymbol{p} = \boldsymbol{q}$. □

Definition The mutual information $I(X;Y)$ between two random variables X and Y, taking values from the alphabet set $\mathcal{X} \times \mathcal{Y}$ with joint distribution $p_{XY}(\cdot)$ and marginal distributions $p_X(\cdot)$ and $p_Y(\cdot)$ respectively, is defined as follows:

$$I(X;Y) = D(p_{XY}(\cdot,\cdot)||p_X(\cdot)p_Y(\cdot)) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \tag{8}$$

Writing $p_{XY}(x,y)$ as $p_X(x)p_{Y|X}(y|x)$, the above quantity may be re-written as

$$\begin{aligned} I(X;Y) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_X(x)p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{p_Y(y)} \\ &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_X(x)p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{\sum_{z\in X} p_X(x)p_{Y|X}(y|z)} \qquad (9) \end{aligned}$$

Definition A Discrete Memoryless Channel [6], denoted by $(\mathcal{X}, p_{Y|X}(y|x), \mathcal{Y})$ consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of probability mass functions $\{p_{Y|X}(\cdot|x), x \in \mathcal{X}\}$, with the interpretation that X is the input and Y is the output of the channel.

The capacity C of the DMC is defined as the maximum possible rate of information transmission with arbitrary small probability of error. Shannon established the following fundamental result in his seminal paper [5].

Theorem 2.2 (The Noisy Channel Coding Theorem)**.**

$$C = \max_{p_X} I(X;Y) \tag{10}$$

In the rest of this article, we discuss algorithms that solve the optimization problem 10 for a given DMC.

3 Some Convexity Results

In this section we will establish the convexity of the optimization problem 10, starting from the first principles. To simplify notations, we re-label the input symbols as $[1 \dots N]$ and output symbols as $[1 \dots M]$ where $N = |\mathcal{X}|$ and $M = |\mathcal{Y}|$. We denote the $1 \times N$ input probability vector by $\boldsymbol{p}$, the $N \times M$ channel matrix by $\boldsymbol{Q}$ and the $1 \times M$ output probability vector by $\boldsymbol{q}$. Then by the laws of probability, we have

$$\boldsymbol{q} = \boldsymbol{p}\boldsymbol{Q} \tag{11}$$

Hence the objective function $I(X;Y)$ can be re-written as

$$\begin{aligned} I(X;Y) = I(\boldsymbol{p}, \boldsymbol{Q}) &= \sum_{i=1}^{N}\sum_{j=1}^{M} p_i Q_{ij} \log \frac{Q_{ij}}{q_j} \\ &= \sum_{i=1}^{N} p_i \big(\sum_{j=1}^{M} Q_{ij} \log Q_{ij} \big) - \sum_{j=1}^{M} \big(\sum_{i=1}^{N} p_i Q_{ij} \big) \log q_j \\ &= \sum_{i=1}^{N} p_i \big(\sum_{j=1}^{M} Q_{ij} \log Q_{ij} \big) - \sum_{j=1}^{M} q_j \log q_j \end{aligned} \tag{12}$$

Where we have utilized equation 11 in the last equation.

Lemma 3.1. *For a fixed channel-matrix $\boldsymbol{Q}$, $I(\boldsymbol{p}, \boldsymbol{Q})$ is concave in the input probability distribution $\boldsymbol{p}$ and hence the problem 10 has an optimal solution.*

Proof. We first establish that the function $f(x) = x \log x, x \geq 0$ is convex in x. To establish this, just note that $f''(x) > 0, \forall x > 0$. Hence the function $f(\boldsymbol{q}) = \sum_{i=1}^{M} q_i \log q_i$ is convex in $\boldsymbol{q}$. Now from Eqn. 11 we note that $\boldsymbol{q}$ is a linear transformation of the input probability vector $\boldsymbol{p}$. Hence, viewed as a function of $\boldsymbol{p}$, the second term on the right of Eqn. 12 is convex in $\boldsymbol{p}$. Since the first term is linear in $\boldsymbol{p}$, the result follows. □

Since the constraint set in the optimization problem 10 is the probability simplex $\boldsymbol{p}\mathbf{1} = 1, p_i \geq 0$, the above lemma establishes that the optimization problem 10 is that of maximizing a concave function over a convex constraint set. We record this fact in the following theorem

Theorem 3.2. *The optimization problem given in 10 is convex.*

4 A Variational Characterization of Mutual Information

In this section we will express the mutual information $I(X;Y)$ as variational problem. This will lead us directly to an *alternating minimization* algorithm for solving the optimization problem 10.
Let us denote the set of all conditional distributions on the alphabet $\mathcal{X}$, indexed by the output alphabet $\mathcal{Y}$ by $\Phi = \{\phi(\cdot|j), j \in \mathcal{Y}\}$. For any $\phi \in \Phi$ define the quantity $\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi})$ as follows:

$$\boxed{\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) = \sum_{i=1}^{N}\sum_{j=1}^{M} p_i Q_{ij} \log \frac{\phi(i|j)}{p_i}} \tag{13}$$

The concavity of $\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi})$ w.r.t. $\boldsymbol{p}$ and $\boldsymbol{\phi}$ is readily apparent.

Lemma 4.1. *For fixed $\mathbf{p}$ and Q, $\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi})$ is* concave *in $\boldsymbol{\phi}$. Similarly, for fixed $\boldsymbol{\phi}$ and Q, $\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi})$ is* concave *in $\boldsymbol{p}$.*

Proof. This follows from the concavity of the functions $\log(x)$ and $x \log \frac{1}{x}$ and the definition of $\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi})$. □

Clearly, from the defining Eqn. 12, it follows that for the particular choice $\phi^*(i|j) = p_i \frac{Q_{ij}}{\sum_{i=1}^{N} p_i Q_{ij}}$, we have

$$\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}^*) = I(\boldsymbol{p}, \boldsymbol{Q}) \tag{14}$$

The following lemma shows that $\boldsymbol{\phi}^*$ maximizes $\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \cdot)$.

Lemma 4.2. *For any matrix of conditional probabilities $\boldsymbol{\phi}$, we have*

$$\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) \leq I(\boldsymbol{p}, \boldsymbol{Q}) \tag{15}$$

Proof. We have,

$$I(\boldsymbol{p}, \boldsymbol{Q}) - \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) = \sum_{i=1}^{N}\sum_{j=1}^{M} p_i Q_{ij} \log \frac{Q_{ij}}{q_j} - \sum_{i=1}^{N}\sum_{j=1}^{M} p_i Q_{ij} \log \frac{\phi(i|j)}{p_i} \tag{16}$$

$$= \sum_{j=1}^{M} q_j \sum_{i=1}^{N} \frac{p_i Q_{ij}}{q_j} \log \frac{p_i Q_{ij}/q_j}{\phi(i|j)} \tag{17}$$

$$= \sum_{j=1}^{M} q_j \sum_{i=1}^{N} \phi^*(i|j) \log \frac{\phi^*(i|j)}{\phi(i|j)} \tag{18}$$

Define $r(i|j) = p_i Q_{ij}/q_j$ which can be interpreted as *a posteriori* input probability distribution given the output variable to take value j. Then we can write down the above equation as follows

$$I(\boldsymbol{p}, \boldsymbol{Q}) - \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) \quad = \quad \sum_{j=1}^{M} q_j D(\phi^*(\cdot|j)||\phi(\cdot|j)) \tag{19}$$

Which is non-negative and equality holds iff $\phi(i|j) = \phi^*(i|j), \forall i, j$, by virtue of lemma 2.1. □

Combining the above results, we have the following variational characterization of mutual information

Theorem 4.3. *For any input distribution $\boldsymbol{p}$ and any channel matrix $\boldsymbol{Q}$ we have*

$$I(\boldsymbol{p}, \boldsymbol{Q}) = \max_{\phi \in \Phi} \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) \tag{20}$$

And the conditional probability matrix that achieves the maximum is given by

$$\phi(i|j) = \phi^*(i|j) = p_i \frac{Q_{ij}}{\sum_{i=1}^{N} p_i Q_{ij}} \tag{21}$$

Based on the above theorem, we can recast the optimization problem 10 as follows

$$\boxed{C = \max_{\boldsymbol{p}_X} \max_{\boldsymbol{\phi} \in \Phi} \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi})} \tag{22}$$

Since the channel matrix $\boldsymbol{Q}$ is fixed, we can view the optimization problem 22 as optimizing over two different sets of variables, $\boldsymbol{p}$ and $\boldsymbol{\phi}$. One natural iterative approach to solve the problem is to fix one set of variables and optimize over the other and vice versa. This method is especially attractive when closed form solution for both the maximizations are available. As we will see in the following theorem, this is precisely the case here. This is in essence the Blahut-Arimoto (BA) algorithm for obtaining the capacity of a Discrete Memoryless Channel [1].

5 The geometric idea of alternating optimization

We consider the following problem, given two convex sets A and b in $\mathbb{R}^n$ as shown in Figure 2, we wish to determine the minimum distance between them. More precisely, we wish to determine

$$d_{\min} = \min_{a \in A, b \in B} d(a, b) \tag{23}$$

Where $d(a, b)$ is the euclidean distance between a and b. An obvious algorithm to do this would be to take any point $x \in A$, and find the $y \in B$ that is closest to it. Then fix this y and find the closest point in A. Repeating this process, it is clear that the distance is non-increasing at each stage. But it is not obvious whether the algorithm

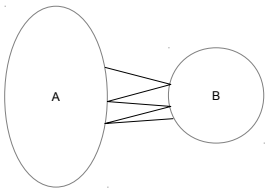

Figure 2: Alternating Minimization

converges to the optimal solution. However, we will show that if the sets are probability distributions and the distance measure is the relative entropy then the algorithm does converge to the minimum relative entropy between two sets. To use the above idea of alternating optimization in problem 22, if possible, it is advantageous to have a closed form expression for the solution of either optimization problem. The theorem below indicates that this is indeed possible and finds the solution of either optimization in closed form.

Theorem 5.1. *For a fixed $\boldsymbol{p}$ and $\boldsymbol{Q}$, we have*

$$\arg\max_{\phi\in\Phi} \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) = \boldsymbol{\phi}^* \tag{24}$$

Where,

$$\phi^*(i|j) = p_i \frac{Q_{ij}}{\sum_{k=1}^{N} p_k Q_{kj}} \tag{25}$$

And for a fixed $\boldsymbol{\phi}$ and $\boldsymbol{Q}$, we have

$$\arg\max_{\boldsymbol{p}} \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) = \boldsymbol{p}^* \tag{26}$$

Where the components of $\boldsymbol{p}^*$ are given by,

$$p^*(i) = \frac{r_i}{\sum_k r_k} \tag{27}$$

And the maximum value is given by

$$\max_{\boldsymbol{p}} \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) = \log(\sum_i r_i) \tag{28}$$

Where,

$$r_i = \exp\left(\sum_j Q_{ij} \log \phi(i|j)\right) \tag{29}$$

Proof. The first part of the theorem is already proved as part of the theorem 4.3. We prove here the second part only.
As with any constrained-optimization problem with equality constraint, a straightforward way to approach the problem is to use the method of lagrange-multiplier. However an elegant way to solve the problem is to use lemma 2.1 in a clever way to tightly upper-bound the objective function and then to find an optimal input-distribution $\boldsymbol{p}^*$ which achieves the bound. We take this approach here.

Consider an input distribution $\boldsymbol{p}^*$ such that $p^*(i) = Dr(i), \forall i \in \mathcal{X}$ where,

$$\log r(i) = \sum_j Q_{ij} \log \phi(i|j) \tag{30}$$

And D is the normalization constant, i.e. $D = (\sum_{i \in \mathcal{X}} r(i))^{-1}$. From Lemma 2.1, we have

$$D(\boldsymbol{p}||\boldsymbol{p}^*) \geq 0 \tag{31}$$

i.e.,

$$\sum_{i=1}^{N} p_i \log p_i \geq \sum_{i=1}^{N} p_i \log p^*(i) = \log D + \sum_{i=1}^{N} p_i \sum_{j=1}^{M} Q_{ij} \log \phi(i|j) \tag{32}$$

Rearranging the above equation, we have

$$\tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i Q_{ij} \log \frac{\phi(i|j)}{p_i} \leq -\log D \tag{33}$$

With the equality holding iff $\boldsymbol{p} = \boldsymbol{p}^*$. Clearly the optimal value is given by $-\log D = \log \sum_i r_i$. □

Equipped with Theorem 5.1, we are now ready to describe the Blahut-Arimoto (BA) algorithm formally.

Step 1: Initialize $\boldsymbol{p}^{(1)}$ to the uniform distribution over $\mathcal{X}$, i.e. $p_i^{(1)} = \frac{1}{|\mathcal{X}|}$ for all $i \in \mathcal{X}$. Set t to 1.
Step 2: Find $\boldsymbol{\phi}^{(t+1)}$ as follows:

$$\phi^{(t+1)}(i|j) = \frac{p_i^{(t)} Q_{ij}}{\sum_k p_k^{(t)} Q_{kj}}, \quad \forall i, j \tag{34}$$

Step 3: Update $\boldsymbol{p}^{(t+1)}$ as follows:

$$p_i^{(t+1)} = \frac{r_i^{(t+1)}}{\sum_{k \in \mathcal{X}} r_k^{(t+1)}} \tag{35}$$

Where,

$$r_i^{(t+1)} = \exp\left(\sum_j Q_{ij} \log \phi^{(t+1)}(i|j)\right) \tag{36}$$

Step 4: Set $t \leftarrow t+1$ and goto Step 2.

We can combine **Step 2** and **Step 3** as follows. Denote the output distribution induced by the input distribution $\boldsymbol{p}^t$ by $\boldsymbol{q}^t$, i.e. $\boldsymbol{q}^t = \boldsymbol{p}^t\boldsymbol{Q}$. Hence from Eqn. 34 we have

$$\phi^{(t+1)}(i|j) = \frac{p_i^{(t)} Q_{ij}}{q^{(t)}(j)} \tag{37}$$

We now evaluate the term inside the exponent of Eqn 36 as follows

$$\sum_j Q_{ij} \log \phi^{(t+1)}(i|j) = \sum_j Q_{ij} \log \frac{p_i^{(t)} Q_{ij}}{q^{(t)}(j)} = D(\boldsymbol{Q}_i||\boldsymbol{q}^{(t)}) + \log p_i^{(t)} \tag{38}$$

Where $\boldsymbol{Q}_i$ denotes the i^{th} row of the channel matrix $\boldsymbol{Q}$. Hence from Eqn. 36 we have

$$r_i^{(t+1)} = p_i^{(t)} \exp\left(D(\boldsymbol{Q}_i||\boldsymbol{q}^{(t)})\right) \tag{39}$$

Thus the above algorithm has the following simplified description

Simplified Blahut-Arimoto Algorithm

Step 1: Initialize $\boldsymbol{p}^{(1)}$ to the uniform distribution over $\mathcal{X}$, i.e. $p_i^{(1)} = \frac{1}{|\mathcal{X}|}$ for all $i \in \mathcal{X}$. Set t to 1.
Step 2: Repeat until convergence:

$$\boldsymbol{q}^{(t)} = \boldsymbol{p}^{(t)}\boldsymbol{Q} \tag{40}$$

$$p_i^{(t+1)} = p_i^{(t)} \frac{\exp\left(D(\boldsymbol{Q}_i||\boldsymbol{q}^{(t)})\right)}{\sum_k p_k^{(t)} \exp\left(D(\boldsymbol{Q}_k||\boldsymbol{q}^{(t)})\right)} \quad \forall i \in \mathcal{X} \tag{41}$$

6 Proximal Point Reformulation : Accelerated Blahut-Arimoto Algorithm

In this section we re-examine the alternating minimization procedure of the Blahut-Arimoto algorithm [2]. Plugging in the optimal solution $\boldsymbol{\phi}^t$ from the first optimization

to the second optimization, we have

$$\begin{aligned}
\boldsymbol{p}^{t+1} &= \arg\max_{\boldsymbol{p}} \tilde{I}(\boldsymbol{p}, \boldsymbol{Q}; \boldsymbol{\phi}^t) \\
&= \arg\max_{\boldsymbol{p}} \sum_{i=1}^{N}\sum_{j=1}^{M} p_i Q_{ij} \log \frac{\phi^t(i|j)}{p_i} \\
&= \arg\max_{\boldsymbol{p}} \sum_{i=1}^{N}\sum_{j=1}^{M} p_i Q_{ij} \log \frac{p_i^t Q_{ij}}{p_i q_j^t} \\
&= \arg\max_{\boldsymbol{p}} \left(\sum_{i=1}^{N} p_i D(\boldsymbol{Q}_i||\boldsymbol{q}^t) - D(\boldsymbol{p}||\boldsymbol{p}^t) \right) \qquad (42)
\end{aligned}$$

The Eqn. (42) can be interpreted a maximization of $\sum_{i=1}^{N} p_i D(\boldsymbol{Q}_i||\boldsymbol{q}^t)$ with a penalty term $D(\boldsymbol{p}||\boldsymbol{p}^t)$ which ensures that the update $\boldsymbol{p}^{t+1}$ remains in the vicinity of $\boldsymbol{p}^t$ [2]. Algorithms of this type are known as *proximal point methods*, since they force the update to stay in the proximity of the current guess. This is reasonable in our case because the first term in 42 is an approximation of the mutual information $\boldsymbol{I}(\boldsymbol{p}; \boldsymbol{Q})$, by replacing the KLDs $D(\boldsymbol{Q}_i||\boldsymbol{q}^*)$ with $D(\boldsymbol{Q}_i||\boldsymbol{q}^t)$. The penalty term $D(\boldsymbol{p}||\boldsymbol{p}^k)$ ensures that the maximization is restricted to a neighbourhood of $\boldsymbol{p}^k$ for which the approximation $D(\boldsymbol{Q}_i||\boldsymbol{q}^*) \approx D(\boldsymbol{Q}_i||\boldsymbol{q}^t)$ is accurate. In fact we have the following equality

$$\boldsymbol{p}^{k+1} = \arg\max_{\boldsymbol{p}} \left(\tilde{I}^t(\boldsymbol{p}) - D(\boldsymbol{p}||\boldsymbol{p}^t) \right) \qquad (43)$$

Where $\tilde{I}^t(\boldsymbol{p}) = I(\boldsymbol{p}^t, \boldsymbol{Q}) + \sum_{i=1}^{N} (p_i - p_i^t) D(\boldsymbol{Q}_i||\boldsymbol{q}^t)$, which can be shown to be a first-order Taylor series approximation of $I(\boldsymbol{p})$. Thus the original Blahut-Arimoto algorithm can be thought of as a proximal point method maximizing the first-order Taylor series approximation of $I(\boldsymbol{p}, \boldsymbol{Q})$ with a proximity penalty expressed by $D(\boldsymbol{p}||\boldsymbol{p}^t)$.
It is now natural to modify (43) by an emphasizing/attenuating the penalty term via a weighting factor, i.e.,consider the following iteration

$$\boldsymbol{p}^{t+1} = \arg\max_{\boldsymbol{p}} \left(\tilde{I}^t(\boldsymbol{p}) - \gamma_t D(\boldsymbol{p}||\boldsymbol{p}^t) \right) \qquad (44)$$

The idea is that close to the optimal solution the K-L distance of $\boldsymbol{p}$ to $\boldsymbol{p}^t$ would be small and hence the proximity constraint γ_t can be gradually relaxed by decreasing γ_t. One such possible choice of the $\{\gamma_t\}_{t\geq 1}$ sequence. In the following sub-section we derive a sequence of step-sizes that guarantees non-decreasing mutual information estimates $I(\boldsymbol{p}^{(t)}, \boldsymbol{Q})$.
We note below the accelerated BA algorithm as derived above

Step 1: Initialize $\boldsymbol{p}^{(1)}$ to the uniform distribution over $\mathcal{X}$, i.e. $p_i^{(1)} = \frac{1}{|\mathcal{X}|}$ for all $i \in \mathcal{X}$. Set t to 1.

Step 2: Repeat until convergence:

$$\boldsymbol{q}^{(t)} = \boldsymbol{p}^{(t)}\boldsymbol{Q} \tag{45}$$

$$p_i^{(t+1)} = p_i^{(t)} \frac{\exp\left(\gamma_t^{-1} D(\boldsymbol{Q}_i||\boldsymbol{q}^{(t)})\right)}{\sum_k p_k^{(t)} \exp\left(\gamma_t^{-1} D(\boldsymbol{Q}_k||\boldsymbol{q}^{(t)})\right)} \quad ,\forall i \in \mathcal{X} \tag{46}$$

6.1 Suitable choice of step-sizes for accelerated BA algorithm

A fundamental property of the BA algorithm is that the mutual information $I(\boldsymbol{p}^t, \boldsymbol{Q})$, which represents the current capacity estimate at the t th iteration is non-decreasing. For the accelerated BA algorithm, we need to choose a sequence $\{\gamma_t\}_{t\geq 1}$ that preserves this property. For this we need the following lemma.

Lemma 6.1. *For any iteration t, we have*

$$D(\boldsymbol{q}^{(t+1)}||\boldsymbol{q}^{(t)}) \geq \sum_{i=1}^{N} p_i^{(t+1)} D(\boldsymbol{Q}_i||\boldsymbol{q}^{(t)}) \tag{47}$$

Proof. Recall that,

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{p}^{(t+1)}\boldsymbol{Q} = \sum_{i=1}^{N} p_i^{(t+1)} \boldsymbol{Q}_i \tag{48}$$

The above equation expresses the output probability vector $\boldsymbol{q}^{k+1}$ as a convex combination of the rows of the matrix $\boldsymbol{Q}$. Since the relative entropy $D(\cdot||\cdot)$ is convex in both the arguments, we have

$$D(\boldsymbol{q}^{(t+1)}||\boldsymbol{q}^{(t)}) \geq \sum_{i=1}^{N} p_i^{(t+1)} D(\boldsymbol{Q}_i||\boldsymbol{q}^{(t)})$$

□

Equipped with the above lemma, we now establish a lower bound on increment of mutual information $I(\boldsymbol{p}, \boldsymbol{Q})$ at each stage.

Lemma 6.2. *For every stage t of the accelerated BA algorithm , we have*

$$I(\boldsymbol{p}^{(t+1)}, \boldsymbol{Q}) \geq I(\boldsymbol{p}^{(t)}, \boldsymbol{Q}) + \gamma_t D(\boldsymbol{p}^{(t+1)}||\boldsymbol{p}^{(t)}) - D(\boldsymbol{q}^{(t+1)}||\boldsymbol{q}^{(t)}) \tag{49}$$

Proof. We have from Eqn. 44 of the accelerated BA iteration

$$\tilde{I}^t(\boldsymbol{p}^{(t+1)}) - \gamma_t D(\boldsymbol{p}^{(t+1)}||\boldsymbol{p}^{(t)}) \geq \tilde{I}^t(\boldsymbol{p}^{(t)}) \tag{50}$$

Plugging in the expression for $\tilde{I}^t(\cdot)$ from above, we have

$$I(\boldsymbol{p}^{(t+1)}, \boldsymbol{Q}) + \sum_{i=1}^{N} (p_i^{(t+1)} - p_i^{(t)}) D(\boldsymbol{Q}_i || \boldsymbol{q}^{(t)}) \geq I(\boldsymbol{p}^{(t)}, \boldsymbol{Q}) + \gamma_t D(\boldsymbol{p}^{(t+1)} || \boldsymbol{p}^{(t)}) \quad (51)$$

Now using the lemma 6.1 and using the non-negativity of K-L divergence, the result follows. □

From the above lemma, it follows that a sufficient condition for $I(\boldsymbol{p}^{(t+1)}, \boldsymbol{Q})$ to be non-decreasing is

$$\frac{1}{\gamma_t} \leq \frac{D(\boldsymbol{p}^{(t+1)} || \boldsymbol{p}^{(t)})}{D(\boldsymbol{p}^{(t+1)} \boldsymbol{Q} || \boldsymbol{p}^{(t)} \boldsymbol{Q})} \quad (52)$$

Now define the *maximum KLD-induced eigenvalue* of $\boldsymbol{Q}$ as

$$\lambda_{KL}^2(\boldsymbol{Q}) = \sup_{\boldsymbol{p} \neq \boldsymbol{p}'} \frac{D(\boldsymbol{p}\boldsymbol{Q} || \boldsymbol{p}'\boldsymbol{Q})}{D(\boldsymbol{p} || \boldsymbol{p}')} \quad (53)$$

Using the above definition, we conclude that a sufficient condition for $I(\boldsymbol{p}^{(t+1)}, \boldsymbol{Q})$ to be non-decreasing is given by

$$\gamma_t \geq \lambda_{KL}^2(\boldsymbol{Q}) \quad (54)$$

7 Convergence Statements of the Accelerated BA Algorithm

In the previous section we proved that for any step-size sequence $\gamma_t \geq \lambda_{KL}^2(\boldsymbol{Q})$, the accelerated BA algorithm has the potential for increased convergence speed. For lack of space, we only give the statements of the theorem. Complete proofs of these theorems may be found in [2].

Theorem 7.1. *Consider the accelerated BA algorithm with* $I^t = \sum_i p_i^t D(\boldsymbol{Q}_i || \boldsymbol{q}^{(t)})$ *and* $L_t = \gamma_t \log(\sum_i p_i^{(t)} \exp(\gamma_t^{-1} D(\boldsymbol{Q}_i || \boldsymbol{q}^{(t)}))$. *Assume that* $\gamma_{\inf} = \inf_t \gamma_t^{-1} > 0$ *and* (54) *is satisfied for all* t. *Then*

$$\lim_{t \to \infty} L_t = \lim_{t \to \infty} I^t = C \quad (55)$$

And the convergence rate is at least proportional to $1/t$, *i.e.*

$$C - L^t < \frac{D(\boldsymbol{p}^* || \boldsymbol{p}^0)}{\mu_{\inf} t} \quad (56)$$

8 Dual Approach : Geometric Programming

In this section, we take a dual approach to solve the problem 10 and show that the dual problem reduces to a simple Geometric Program [4]. We also derive several useful upper bounds on the channel capacity from the dual program.
First we rewrite the mutual information functional as follows.

$$I(X;Y) = H(Y) - H(Y|X) = -\sum_{j=1}^{M} q_j \log q_j - \boldsymbol{pr} \tag{57}$$

Where,

$$r_i = -\sum_{j=1}^{M} Q_{ij} \log Q_{ij} \tag{58}$$

Subject to,

$$\boldsymbol{q} = \boldsymbol{pQ} \tag{59}$$

$$\boldsymbol{p1} = 1, \;\; \boldsymbol{p} >= \boldsymbol{0} \tag{60}$$

Hence the optimization problem 10 may be rewritten as follows

$$\max \;\; -\boldsymbol{pr} - \sum_{j=1}^{M} q_j \log q_j \tag{61}$$

Subject to,

$$\begin{aligned} \boldsymbol{pQ} &= \boldsymbol{q} \\ \boldsymbol{p1} &= 1 \\ \boldsymbol{p} &\geq \boldsymbol{0} \end{aligned}$$

It is to be noted that keeping two sets of optimization variables $\boldsymbol{p}$ and $\boldsymbol{q}$ and introducing the equality constraint $\boldsymbol{pQ} = \boldsymbol{q}$ in the primal problem is a key step to derive an explicit and simple Lagrange dual problem of 61.

Theorem 8.1. *The Lagrange dual of the channel capacity problem 61 is given by the following problem*

$$\min_{\boldsymbol{\alpha}} \;\; \log \sum_{j=1}^{M} \exp(\alpha_j) \tag{62}$$

Subject to,

$$\boldsymbol{Q\alpha} \geq -\boldsymbol{r} \tag{63}$$

An equivalent version of the above Lagrange dual problem is the following Geometric program (in the standard form):

$$\min_{\boldsymbol{z}} \sum_{j=1}^{M} z_j \tag{64}$$

Subject to,

$$\begin{aligned} \prod_{j=1}^{M} z_j^{P_{ij}} &\geq \exp\big(-H(\boldsymbol{Q}_i)\big), \quad i = 1, 2, \ldots, N \\ \boldsymbol{z} &\geq 0 \end{aligned}$$

From the Lagrange dual problem, we immediately have the following upper bound on the channel capacity.

- *Weak Duality:* $\log\left(\sum_{j=1}^{M} \exp(\alpha_j)\right) \geq C$, for all $\boldsymbol{\alpha}$ that satisfy $\boldsymbol{Q}\boldsymbol{\alpha} + \boldsymbol{r} \geq \mathbf{0}$.
- *Strong Duality:* $\log\left(\sum_{j=1}^{M} \exp(\alpha_j)\right) = C$, for the optimal dual variable $\boldsymbol{\alpha}^*$.

8.1 Bounding From the Dual

Because the inequality constraints in the dual problem 64 are affine, it is easy to obtain a dual feasible $\boldsymbol{\alpha}$ by finding any solution to a system of linear inequalities, and the resulting value of the dual objective function provides an easily derivable upper bound on channel capacity. The following is one such non-trivial bound.

Corollary 8.2. *Channel capacity is upper-bounded in terms of a maximum-likelihood receiver selecting* $\arg\max_i P_{ij}$ *for each output symbol* j

$$C \leq \log \sum_{j=1}^{M} \max_i P_{ij}$$

which is tight iff the optimal output distribution $\boldsymbol{q}^*$ *is*

$$q_j^* = \frac{\max_i P_{ij}}{\sum_{k=1}^{M} P_{ik}} \tag{65}$$

As is readily apparent, the geometric program Lagrange dual 64 generates a broader class of upper bounds on capacity. This upper bounds can be effectively used to terminate an iterative optimization procedure for channel capacity.

9 Conclusion

In this report, we have surveyed various convex optimization algorithms for solving the channel capacity problem. In particular, we have derived the classical Blahut-Arimoto algorithm from first principles. Then we established a connection with Proximal algorithms and original BA iteration. Using a proper step-size sequence, we have derived an acceerated version of the BA algorithm. Finally we have considered the dual of the channel capacity problem and have shown that its Lagrange dual is given by a Geometric Program (GP). The GP have been effectively utilized to derive non-trivial upper bound on the channel capacity.

References

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

[2] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated blahut-arimoto-type algorithms," in *Information theory workshop, 2004. IEEE*. IEEE, 2004, pp. 66–70.

[3] Y. Yu, "Squeezing the arimoto-blahut algorithm for faster convergence," *arXiv preprint arXiv:0906.3849*, 2009.

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

[5] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[6] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.