Convex Optimization Methods for Computing Channel Capacity

Abhishek Sinha

Laboratory for Information and Decision Systems
MIT

May 15, 2014

The Communication Problem, Formally

The Problem Statement:

- The source possess M distinct messages, one of which it wishes to communicate with the destination.
- The noisy channel takes in one of the N input-symbol (say i) and produces one of the M output symbol with probability distribution $\mathbf{Q}_i$ *independently* of everything else.

The Communication Problem, Formally

The Problem Statement:

- The source possess M distinct messages, one of which it wishes to communicate with the destination.
- The noisy channel takes in one of the N input-symbol (say i) and produces one of the M output symbol with probability distribution $\mathbf{Q}_i$ *independently* of everything else.
- The source encodes each of the M messages using n input symbols and the destination decodes each of the received sequence to some message $\hat{M}$.

The Communication Problem, Formally

The Problem Statement:

- The source possess M distinct messages, one of which it wishes to communicate with the destination.
- The noisy channel takes in one of the N input-symbol (say i) and produces one of the M output symbol with probability distribution $\mathbf{Q}_i$ *independently* of everything else.
- The source encodes each of the M messages using n input symbols and the destination decodes each of the received sequence to some message $\hat{M}$.
- Rate of communication is defined as $\frac{\log M}{n}$

The Communication Problem, Formally

The Problem Statement:

- The source possess M distinct messages, one of which it wishes to communicate with the destination.
- The noisy channel takes in one of the N input-symbol (say i) and produces one of the M output symbol with probability distribution $\mathbf{Q}_i$ *independently* of everything else.
- The source encodes each of the M messages using n input symbols and the destination decodes each of the received sequence to some message $\hat{M}$.
- Rate of communication is defined as $\frac{\log M}{n}$

Maximum Achievable Rate

Over all encoding and decoding schemes, what is the maximum achievable rate, for arbitrarily small probability of error ?

$$\max \liminf \frac{\log M}{n} \tag{1}$$

s.t.

$$\mathbb{P}_n(M \neq \hat{M}) \searrow 0 \tag{2}$$

Where it all started - Shannon (1948)

34 *The Mathematical Theory of Communication*

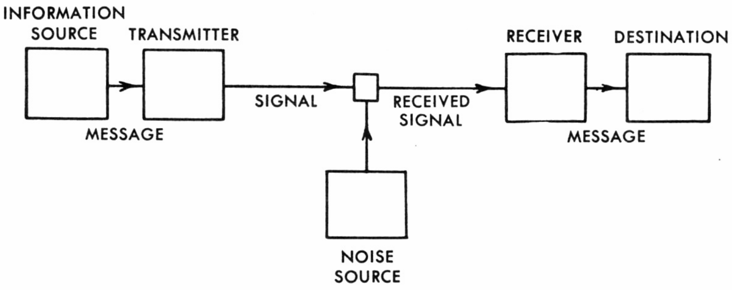
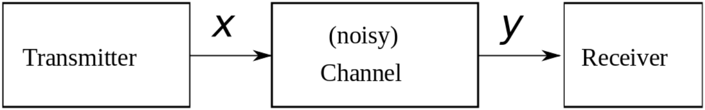
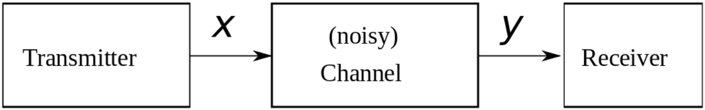


Fig. 1. — Schematic diagram of a general communication system.

The Fundamental Limit: Channel Capacity



The Fundamental Limit: Channel Capacity



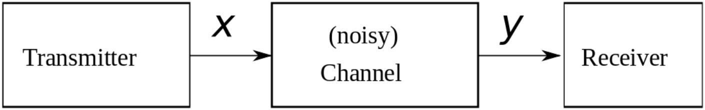
Theorem: Shannon 1948

For every channel matrix $\mathbf{Q}$, maximum achievable rate is given by

$$C = \max_{\mathbf{p}_X} I(X;Y) \tag{3}$$

Where $I(X;Y)$ denotes the *mutual information* between the random variables X and Y.

The Fundamental Limit: Channel Capacity



Theorem: Shannon 1948

For every channel matrix $\mathbf{Q}$, maximum achievable rate is given by

$$C = \max_{\mathbf{p}_X} I(X;Y) \tag{3}$$

Where $I(X;Y)$ denotes the *mutual information* between the random variables X and Y.

Objective of this talk

Solve the optimization problem 3.

Review of some useful functionals

- For two PMF **p** and **q** with the same support, the K-L divergence between **p** and **q** is given by,

$$D(\mathbf{p}||\mathbf{q}) = \sum_{x \in X} p_x \log \frac{p_x}{q_x}$$

Property:

$$D(\mathbf{p}||\mathbf{q}) \geq 0 \tag{4}$$

With equality iff $\mathbf{p} = \mathbf{q}$.

- **Mutual Information**

$$I(X;Y) = I(\mathbf{p}, \mathbf{Q}) = \sum_{i=1}^{N} p_i \Big(\sum_{j=1}^{M} Q_{ij} \log Q_{ij} \Big) - \sum_{j=1}^{M} q_j \log q_j \tag{5}$$

Where,

$$\mathbf{q} = \mathbf{p}\mathbf{Q} \tag{6}$$

The PMF **q** is known as the output distribution.

Some Properties of mutual information $I(X;Y) = I(\mathbf{p}, \mathbf{Q})$

Lemma

$I(X;Y) \equiv I(\mathbf{p}, \mathbf{Q})$ *is concave in the variable* $\mathbf{p}$.

Thus the problem 3 corresponds to maximizing a differentiable concave function over the probability simplex.

- All *off-the-shelf* constrained convex optimization methods are applicable.

Some Properties of mutual information $I(X;Y) = I(\mathbf{p}, \mathbf{Q})$

Lemma

$I(X;Y) \equiv I(\mathbf{p}, \mathbf{Q})$ *is concave in the variable* $\mathbf{p}$.

Thus the problem 3 corresponds to maximizing a differentiable concave function over the probability simplex.

- All *off-the-shelf* constrained convex optimization methods are applicable.
- Slow in practice as they do not take into account the structure of the problem.

We describe the celebrated Blahut-Arimoto Algorithm for solving the problem.

- we need to obtain a variational characterization of the mutual information $I(X;Y)$.

A Variational Characterization of $I(X;Y) = I(\mathbf{p}, \mathbf{Q})$

For a set of conditional input distributions $\Phi = \{\phi(\cdot|j), j \in \mathcal{Y}\}$ indexed by the output symbol j, define the functional

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

Proposition: For a fixed $\mathbf{Q}$ $\tilde{I}(\mathbf{p}, \mathbf{Q}; \phi)$ is concave individually in $\mathbf{p}$ and ϕ.

A Variational Characterization of $I(X;Y) = I(\mathbf{p}, \mathbf{Q})$

For a set of conditional input distributions $\Phi = \{\phi(\cdot|j), j \in \mathcal{Y}\}$ indexed by the output symbol j, define the functional

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

Proposition: For a fixed $\mathbf{Q}$ $\tilde{I}(\mathbf{p}, \mathbf{Q}; \phi)$ is concave individually in $\mathbf{p}$ and ϕ.

Theorem

For any matrix of conditional probabilities ϕ, we have

$$\max_{\phi} \tilde{I}(\mathbf{p}, \mathbf{Q}; \phi) = I(\mathbf{p}, \mathbf{Q}) \tag{7}$$

where maxima is achieved for $\phi(i|j) = \phi^*(i|j) = p_i \frac{Q_{ij}}{\sum_{i=1}^{N} p_i Q_{ij}}$.

Reformulation of the Optimization Problem

With the help from the previous theorem we can reformulate the original optimization problem OPT as follows

Capacity Reformulation

$$C = \max_{\mathbf{p}} \max_{\phi} \tilde{I}(\mathbf{p}, \mathbf{Q}; \phi) \tag{8}$$

Reformulation of the Optimization Problem

With the help from the previous theorem we can reformulate the original optimization problem OPT as follows

Capacity Reformulation

$$C = \max_{\mathbf{p}} \max_{\phi} \tilde{I}(\mathbf{p}, \mathbf{Q}; \phi) \tag{8}$$

- An intuitively obvious algorithm for solving the above problem would be to repeatedly fix one set of variables ($\mathbf{p}$ or ϕ) and optimize over the other.

Reformulation of the Optimization Problem

With the help from the previous theorem we can reformulate the original optimization problem OPT as follows

Capacity Reformulation

$$C = \max_{\mathbf{p}} \max_{\phi} \tilde{I}(\mathbf{p}, \mathbf{Q}; \phi) \tag{8}$$

- An intuitively obvious algorithm for solving the above problem would be to repeatedly fix one set of variables ($\mathbf{p}$ or ϕ) and optimize over the other.
- This is attractive in this case as there are closed form solutions for both the optimization problems.

Reformulation of the Optimization Problem

With the help from the previous theorem we can reformulate the original optimization problem OPT as follows

Capacity Reformulation

$$C = \max_{\mathbf{p}} \max_{\phi} \tilde{I}(\mathbf{p}, \mathbf{Q}; \phi) \tag{8}$$

- An intuitively obvious algorithm for solving the above problem would be to repeatedly fix one set of variables ($\mathbf{p}$ or ϕ) and optimize over the other.
- This is attractive in this case as there are closed form solutions for both the optimization problems.
- Concave character of $\tilde{I}(\mathbf{p}, \mathbf{Q}; \phi)$ guarantees that the method converges to optima.

Iterative Algorithm for solving OPT

Blahut-Arimoto Algorithm for Channel Capacity

Step 1: Initialize $\mathbf{p}^{(1)}$ 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 $\phi^{(t+1)}$ as follows:

$$\phi^{(t+1)}(i|j) = \frac{p_i^{(t)} Q_{ij}}{\sum_k p_k^{(t)} Q_{kj}}, \quad \forall i,j \tag{9}$$

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

$$p_i^{(t+1)} = \frac{r_i^{(t+1)}}{\sum_{k \in \mathcal{X}} r_k^{(t+1)}} \tag{10}$$

Where,

$$r_i^{(t+1)} = \exp\left(\sum_j Q_{ij} \log \phi^{(t+1)}(i|j)\right) \tag{11}$$

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

Convergence Rates and Improvements

Theorem

The BA algorithm has a convergence rate $\Theta(\frac{1}{t})$.

Can we do better ?

Convergence Rates and Improvements

Theorem

The BA algorithm has a convergence rate $\Theta(\frac{1}{t})$.

Can we do better ?

- By plugging-in the solution ϕ^* can re-write the BA iteration as follows

$$\mathbf{p}^{t+1} = \arg\max_{\mathbf{p}} \left(\sum_{i=1}^{N} p_i D(\mathbf{Q}_i || \mathbf{q}^t) - D(\mathbf{p} || \mathbf{p}^t) \right)$$

Interpreting the last term as a proximal term, the BA iteration nicely fits into the framework of proximal algorithms.

Convergence Rates and Improvements

Theorem

The BA algorithm has a convergence rate $\Theta(\frac{1}{t})$.

Can we do better ?

- By plugging-in the solution ϕ^* can re-write the BA iteration as follows

$$\mathbf{p}^{t+1} = \arg\max_{\mathbf{p}} \left(\sum_{i=1}^{N} p_i D(\mathbf{Q}_i || \mathbf{q}^t) - D(\mathbf{p} || \mathbf{p}^t) \right)$$

  Interpreting the last term as a proximal term, the BA iteration nicely fits into the framework of proximal algorithms.

- Using the idea of appropriately emphasizing/attenuating the penalty term via a weighting factor γ_t, we try the following iteration instead

$$\mathbf{p}^{t+1} = \arg\max_{\mathbf{p}} \left(\sum_{i=1}^{N} p_i D(\mathbf{Q}_i || \mathbf{q}^t) - \gamma_t D(\mathbf{p} || \mathbf{p}^t) \right)$$

Proximal Reformulations Contd.

The sequence $\{\gamma_t\}$ is chosen so that we have strict improvement of Capacity estimate at every iteration. Define the *maximum KLD-induced eigenvalue* of $\mathbf{Q}$ as

$$\lambda_{KL}^2(\mathbf{Q}) = \sup_{\mathbf{p}\neq\mathbf{p}'} \frac{D(\mathbf{pQ}||\mathbf{p}'\mathbf{Q})}{D(\mathbf{p}||\mathbf{p}')}$$

It can be shown that $0 \leq \lambda_{KL}^2(\mathbf{Q}) \leq 1$.

Proximal Reformulations Contd.

The sequence $\{\gamma_t\}$ is chosen so that we have strict improvement of Capacity estimate at every iteration. Define the *maximum KLD-induced eigenvalue* of $\mathbf{Q}$ as

$$\lambda^2_{KL}(\mathbf{Q}) = \sup_{\mathbf{p}\neq\mathbf{p}'} \frac{D(\mathbf{pQ}||\mathbf{p}'\mathbf{Q})}{D(\mathbf{p}||\mathbf{p}')}$$

It can be shown that $0 \leq \lambda^2_{KL}(\mathbf{Q}) \leq 1$.

Lemma

The capacity estimates improves at every iteration if we take $\gamma_t \geq \lambda^2_{KL}(\mathbf{Q})$.

Proximal Reformulations Contd.

The sequence $\{\gamma_t\}$ is chosen so that we have strict improvement of Capacity estimate at every iteration. Define the *maximum KLD-induced eigenvalue* of $\mathbf{Q}$ as

$$\lambda^2_{KL}(\mathbf{Q}) = \sup_{\mathbf{p}\neq\mathbf{p}'} \frac{D(\mathbf{pQ}||\mathbf{p}'\mathbf{Q})}{D(\mathbf{p}||\mathbf{p}')}$$

It can be shown that $0 \leq \lambda^2_{KL}(\mathbf{Q}) \leq 1$.

Lemma

The capacity estimates improves at every iteration if we take $\gamma_t \geq \lambda^2_{KL}(\mathbf{Q})$.

- However $\lambda^2_{KL}(\mathbf{Q})$ might be difficult to estimate.
- A step-size $\gamma_t = \frac{D(\mathbf{p}^{(t)}\mathbf{Q}||\mathbf{p}^{(t-1)}\mathbf{Q})}{D(\mathbf{p}^{(t)}||\mathbf{p}^{(t-1)})}$ is found to work well in practice.
- Convergence rate boosted by at least a factor of γ_∞^{-1}.

Accelerated BA Algorithm

Accelerated BA Algorithm

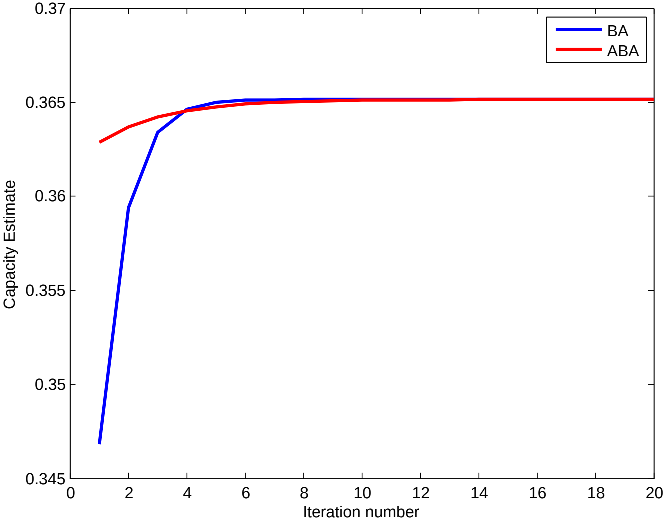
Step 1: Initialize $\mathbf{p}^{(1)}$ to the uniform distribution over $\mathcal{X}$, i.e. $p_i^{(1)} = \frac{1}{|\mathcal{X}|}$ for all $i \in \mathcal{X}$. Set t to 1.

Step 2: Repeat until convergence:

$$\mathbf{q}^{(t)} = \mathbf{p}^{(t)}\mathbf{Q} \tag{12}$$

$$p_i^{(t+1)} = p_i^{(t)} \frac{\exp\left(\gamma_t^{-1} D(\mathbf{Q}_i || \mathbf{q}^{(t)})\right)}{\sum_k p_k^{(t)} \exp\left(\gamma_t^{-1} D(\mathbf{Q}_k || \mathbf{q}^{(t)})\right)} \quad , \forall i \in \mathcal{X} \tag{13}$$

Numerical Simulation



Dual Approach

Finally we take the Lagrange dual of the problem OPT. By straight-forward calculations, it turns out to be the following Geometric Program

$$\min_{\mathbf{z}} \sum_{j=1}^{M} z_j$$

Subject to,

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

- The above GP is useful for deriving outer bounds on capacity.

Conclusion and References

- We have discussed both classical and accelarated Blahut-Arimoto Algorithm for computing Channel capacity of a discrete memoryless channel.
- We have discussed their convergence properties and connection with proximal algorithms

References:

- S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels,
- G. Matz and P. Duhamel, Information geometric formulation and interpretation of accelerated blahut-arimoto-type algorithms,
- M. Chiang and S. Boyd, Geometric programming duals of channel capacity and rate distortion,