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 $Q_i$ *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 $Q_i$ 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 $Q_i$ \textit{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 $Q_i$, 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?

$$\max \lim \inf \frac{\log M}{n}$$

s.t.

$$\mathbb{P}_n(M \neq \hat{M}) \downarrow 0$$
Convex Optimization Methods for Computing Channel Capacity

Where it all started - Shannon (1948)

34

*The Mathematical Theory of Communication*

![Diagram of a communication system](image)

Fig. 1. — Schematic diagram of a general communication system.
The Fundamental Limit: Channel Capacity

Theorem: Shannon 1948

For every channel matrix $Q$, maximum achievable rate is given by

$$C = \max_{p} I(X;Y)$$

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

Objective of this talk

Solve the optimization problem 3.
The Fundamental Limit: Channel Capacity

For every channel matrix $Q$, maximum achievable rate is given by

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

(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.
The Fundamental Limit: Channel Capacity

Theorem: Shannon 1948

For every channel matrix $Q$, maximum achievable rate is given by

$$C = \max_{p_X} I(X; Y)$$  \hspace{1cm} (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,

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

Property:

$$D(p||q) \geq 0 \quad (4)$$

With equality iff $p = q$.

- Mutual Information

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

Where,

$$q = pQ \quad (6)$$

The PMF $q$ is known as the output distribution.
Some Properties of mutual information $I(X; Y) = I(p, Q)$

Lemma

$I(X; Y) = I(p, Q)$ is concave in the variable $p$.

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 $I(X; Y) = I(p, Q)$

Lemma

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

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 $I(X; Y) = I(p, Q)$

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

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

Proposition: For a fixed $Q$ $\tilde{I}(p, Q; \phi)$ is concave individually in $p$ and $\phi$. 
A Variational Characterization of $I(X; Y) = I(p, Q)$

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

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

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

**Theorem**

*For any matrix of conditional probabilities $\phi$, we have*

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

*where maxima is achieved for $\phi(i|j) = \phi^*(i|j) = p_i \frac{Q_{ij}}{\sum_{i=1}^{N} p_i Q_{ij}}$.***
With the help from the previous theorem we can reformulate the original optimization problem OPT as follows

### Capacity Reformulation

\[
C = \max_p \max_{\phi} \tilde{I}(p, Q; \phi) \quad (8)
\]
With the help from the previous theorem we can reformulate the original optimization problem \( \text{OPT} \) as follows

**Capacity Reformulation**

\[
C = \max_p \max_{\phi} \tilde{I}(p, Q; \phi)
\]

(8)

- An intuitively obvious algorithm for solving the above problem would be to repeatedly fix one set of variables (\( p \) or \( \phi \)) 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:

\[
C = \max_p \max_\phi \tilde{I}(p, Q; \phi)
\] (8)

- An intuitively obvious algorithm for solving the above problem would be to repeatedly fix one set of variables (\(p\) or \(\phi\)) 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 \( \text{OPT} \) as follows

\[
C = \max_p \max_{\phi} \tilde{I}(p, Q; \phi)
\]  

(8)

- An intuitively obvious algorithm for solving the above problem would be to repeatedly fix one set of variables (\( p \) or \( \phi \)) 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 \( \tilde{I}(p, Q; \phi) \) guarantees that the method converges to optima.
Iterative Algorithm for solving OPT

Blahut-Arimoto Algorithm for Channel Capacity

Step 1: Initialize $p^{(1)}$ 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:

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

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

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

Where,

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

Step 4: Set $t \leftarrow t + 1$ and goto Step 2.
The BA algorithm has a convergence rate $\Theta\left(\frac{1}{t}\right)$.

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 $\phi^*$ can re-write the BA iteration as follows

$$p_{t+1} = \arg \max \limits_{p} \left( \sum_{i=1}^{N} p_i D(Q_i||q^t) - D(p||p^t) \right)$$

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

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

Can we do better?

- By plugging-in the solution $\phi^*$ can re-write the BA iteration as follows

$$p^{t+1} = \arg \max_p \left( \sum_{i=1}^{N} p_i D(Q_i||q^t) - D(p||p^t) \right)$$

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 $\gamma_t$, we try the following iteration instead

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

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

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

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

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

**Lemma**

The capacity estimates improves at every iteration if we take \( \gamma_t \geq \lambda_{KL}^2(Q) \).
The sequence \( \{\gamma_t\} \) is chosen so that we have strict improvement of Capacity estimate at every iteration. Define the \textit{maximum KLD-induced eigenvalue} of \( Q \) as

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

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

\textbf{Lemma}

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

- However \( \lambda_{KL}^2(Q) \) might be difficult to estimate.
- A step-size \( \gamma_t = \frac{D(p^{(t)}Q \| p^{(t-1)}Q)}{D(p^{(t)} \| p^{(t-1)})} \) is found to work well in practice.
- Convergence rate boosted by at least a factor of \( \gamma^{-1} \).
Accelerated BA Algorithm

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

**Step 2:** Repeat until convergence:

$$ q^{(t)} = p^{(t)} Q $$

$$ p^{(t+1)}_i = p^{(t)}_i \frac{\exp \left( \gamma_t^{-1} D(Q_i \| q^{(t)}) \right)}{\sum_k p^{(t)}_k \exp \left( \gamma_t^{-1} D(Q_k \| q^{(t)}) \right)}, \forall i \in \mathcal{X} $$
Numerical Simulation

Convex Optimization Methods for Computing Channel Capacity
Finally we take the Lagrange dual of the problem OPT. By straight-forward calculations, it turns out to be the following Geometric Program

\[
\min_z \sum_{j=1}^{M} z_j
\]

Subject to,

\[
\prod_{j=1}^{M} z_j^{P_{ij}} \geq \exp(-H(Q_i)), \quad i = 1, 2, \ldots, N
\]

\[
z \geq 0
\]

- The above GP is useful for deriving outer bounds on capacity.
We have discussed both classical and accelerated 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,