Alternating maximization procedure for finding the global maximum of directed information

Iddo Naiss and Haim Permuter

Ben Gurion university

November 18th, 2010
Outline

- Motivation
- Causal conditioning
- Capacity of channels with feedback
- The algorithm, its extensions, and the connection to the regular BA
- Rate distortion with feed forward
- Numerical examples
Motivation

Channel capacity was shown by Shannon to be

\[ C = \lim_{n \to \infty} \frac{1}{n} \max_{p(x^n)} I(X^n; Y^n). \]
Motivation

- Channel capacity was shown by Shannon48 to be

$$ C = \lim_{n \to \infty} \frac{1}{n} \max_{p(x^n)} I(X^n; Y^n). $$

- The problem: obtaining a numerical solution to

$$ \max_{p(x^n)} I(X^n; Y^n). $$
Motivation

- Channel capacity was shown by Shannon48 to be

\[ C = \lim_{n \to \infty} \frac{1}{n} \max_{p(x^n)} I(X^n; Y^n). \]

- The problem: obtaining a numerical solution to

\[ \max_{p(x^n)} I(X^n; Y^n). \]

- Blahut/Arimoto72 solution: solving

\[ \max_{p(x^n), p(x^n|y^n)} I(X^n; Y^n), \]

using the *Alternating maximization procedure*. 

Iddo Naiss and Haim Permuter

Extension of the BAA for maximizing directed information
In our work we calculate bounds and estimators for feedback channel capacity, and feedforward rate distortion, using the alternating optimization procedure.
The Alternating maximization procedure

Let $f(u_1, u_2)$, be a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$.

The goal: to find

$$\max_{A_1, A_2} f(u_1, u_2).$$

The procedure:

$$u_{1}^{(k+1)} = \arg \max_{A_1} f(u_{1}^{(k)}, u_{2}^{(k)}).$$

$$u_{2}^{(k+1)} = \arg \max_{A_2} f(u_{1}^{(k+1)}, u_{2}^{(k)}).$$

$$f^{(k)} = f(u_{1}^{(k)}, u_{2}^{(k)}).$$
Convergence of the procedure

Lemma (Yeung08 (also by Csiszar/Tusnady84))

Let $f(u_1, u_2)$, be a function $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, that is

- concave and bounded from above,
- continuous and has continuous partial derivatives,
- and the sets $A_1, A_2$, which we maximize $f$ over are convex.

If $\sup_{u_i \in A_i} f(u_1, u_2)$ is attained at $A_i$ for $i \in 1, 2$ then

$$\lim_{k \rightarrow \infty} f^{(k)} = \max_{A_1, A_2} f(u_1, u_2).$$
We denote as \( p(x^n || y^{n-d}) \) the probability mass function (PMF) of \( X^n \) causally conditioned on \( Y^{n-d} \), given by

\[
p(x^n || y^{n-d}) \triangleq \prod_{i=1}^{n} p(x_i | x^{i-1} y^{i-d}).
\]

This leads to the causally conditioned entropy given by

\[
H(X^n || Y^n) \triangleq -\mathbb{E} \left[ \log p(X^n || Y^n) \right].
\]

The directed information from \( X^n \) to \( Y^n \) is denoted by

\[
I(X^n \rightarrow Y^n) \triangleq H(Y^n) - H(Y^n || X^n)
\]

\[
= \sum_{y^n, x^n} p(y^n || x^n) r(x^n || y^{n-1}) \log \frac{q(x^n | y^n)}{r(x^n || y^{n-1})}. 
\]

Iddo Naiss and Haim Permuter

Extension of the BAA for maximizing directed information
In the work of P./Weissman/Goldsmith09, Tatikonda/Mitter09 and Kim08 it was shown that

\[ C = \lim_{n \to \infty} \frac{1}{n} \max_{p(x^n|y^{n-1})} I(X^n \to Y^n). \]
In the work of P./Weissman/Goldsmith09, Tatikonda/Mitter09 and Kim08 it was shown that

\[ C = \lim_{n \to \infty} \frac{1}{n} \max_{p(x^n||y^{n-1})} I(X^n \to Y^n). \]

Problem: obtaining a numerical solution to

\[ C_n = \max_{p(x^n||y^{n-1})} I(X^n \to Y^n). \]
In the work of P./Weissman/Goldsmith09, Tatikonda/Mitter09 and Kim08 it was shown that

\[ C = \lim_{n \to \infty} \frac{1}{n} \max_{p(x^n||y^{n-1})} I(X^n \to Y^n). \]

Problem: obtaining a numerical solution to

\[ C_n = \max_{p(x^n||y^{n-1})} I(X^n \to Y^n). \]

Solution: we show that we can solve the following-

\[ C_n = \max_{p(x^n||y^{n-1}), p(x^n|y^n)} I(X^n \to Y^n). \]

using the alternating maximization procedure.

Iddo Naiss and Haim Permuter  Extension of the BAA for maximizing directed information
Iterative algorithm for calculating
\[
\max_{p(x^n||y^{n-1})} I(X^n \rightarrow Y^n), \text{ where } p(y^n||x^n) \text{ is fixed}
\]

Set \(q(x^n|y^n) = 2^{-n}\) to be the start point.
Iterative algorithm for calculating \( \max p(x^n||y^{n-1}) I(X^n \rightarrow Y^n) \), where \( p(y^n||x^n) \) is fixed

1. Set \( q(x^n|y^n) = 2^{-n} \) to be the start point.
2. Starting from \( i = n \), compute \( r(x_i|x_{i-1}, y_{i-1}) \)

\[
r(x_i|x_{i-1}, y_{i-1}) = \frac{r'(x_i, y_{i-1})}{\sum_{x_i} r'(x_i, y_{i-1})},
\]

where

\[
r'(x_i, y_{i-1}) = \prod_{x_{i+1}^n, y_i^n} \left[ \frac{q(x^n|y^n)}{\prod_{j=i+1}^n r(x_j|x_{j-1}, y_{j-1})} \right] \prod_{j=i}^n p(y_j|x^j, y_{j-1}) \prod_{j=i+1}^n r(x_j|x_{j-1}, y_{j-1}),
\]

and do so \textbf{backwards} until \( i = 1 \).
The algorithm for channel capacity

Once you have $r(x_i | x_{i-1}, y_{i-1})$ for all $i \in \{1, \ldots, n\}$, compute $r(x^n | y^{n-1}) = \prod_{i=1}^{n} r(x_i | x_{i-1}, y_{i-1})$. 
The algorithm for channel capacity

3. Once you have $r(x_i|x_i^{i-1}, y_{i-1}^{i-1})$ for all $i \in \{1, \ldots, n\}$, compute $r(x^n|y^{n-1}) = \prod_{i=1}^n r(x_i|x_i^{i-1}, y_{i-1}^{i-1})$.

4. Compute

$$q(x^n|y^n) = \frac{r(x^n|y^{n-1})p(y^n|x^n)}{\sum_{x^n} r(x^n|y^{n-1})p(y^n|x^n)}.$$
The algorithm for channel capacity

3. Once you have $r(x_i|x^{i-1}, y^{i-1})$ for all $i \in \{1, \ldots, n\}$, compute $r(x^n||y^{n-1}) = \prod_{i=1}^{n} r(x_i|x^{i-1}, y^{i-1})$.

4. Compute

$$q(x^n|y^n) = \frac{r(x^n||y^{n-1})p(y^n|x^n)}{\sum_{x^n} r(x^n||y^{n-1})p(y^n|x^n)}.$$ 

5. Calculate

$$I(X^n \rightarrow Y^n) = \sum_{y^n, x^n} p(y^n|x^n) r(x^n||y^{n-1}) \log \frac{q(x^n|y^n)}{r(x^n||y^{n-1})}$$
The algorithm for channel capacity

3. Once you have $r(x_i|x_{i-1}, y_{i-1})$ for all $i \in \{1, \ldots, n\}$, compute $r(x^n|y^{n-1}) = \prod_{i=1}^{n} r(x_i|x_{i-1}, y_{i-1})$.

4. Compute

$$q(x^n|y^n) = \frac{r(x^n|y^{n-1})p(y^n|x^n)}{\sum_{x^n} r(x^n|y^{n-1})p(y^n|x^n)}.$$  

5. Calculate

$$I(X^n \rightarrow Y^n) = \sum_{y^n, x^n} p(y^n|x^n) r(x^n|y^{n-1}) \log \frac{q(x^n|y^n)}{r(x^n|y^{n-1})}$$

6. Repeat steps (2)-(5) until convergence.
Backwards maximization

Backwards maximization is needed due to the causal conditioning PMF properties, which we take under consideration when we use Lagrange’s multipliers method.

- Regular conditioned constraint:

\[
\sum_{x^n} p(x^n | y^n) = 1.
\]
Backwards maximization

Backwards maximization is needed due to the causal conditioning PMF properties, which we take under consideration when we use Lagrange’s multipliers method.

- Regular conditioned constraint:
  \[
  \sum_{x^n} p(x^n|y^n) = 1.
  \]

- Causal conditioned constraint:
  \[
  \forall (i \in \{1, 2, \ldots, n\}) : \sum_{x_i} p(x_i|x^{i-1}, y^i) = 1.
  \]
Extensions to our algorithm

- General delay. The problem we solve:

$$\max_{r(x^n||y^{n-d})} I(X^n \rightarrow Y^n).$$
General delay. The problem we solve:

$$\max_{r(x^n||y^{n-d})} I(X^n \rightarrow Y^n).$$

Delayed function of the feedback. The problem we solve:

$$\max_{r(x^n||z^{n-d})} I(X^n \rightarrow Y^n),$$

where $z^{n-d} = g(y^{n-d}).$
General delay. The problem we solve:

\[
\max_{r(x^n||y^{n-d})} I(X^n \rightarrow Y^n).
\]

Delayed function of the feedback. The problem we solve:

\[
\max_{r(x^n||z^{n-d})} I(X^n \rightarrow Y^n),
\]

where \( z^{n-d} = g(y^{n-d}) \).

Both cases are solved in the paper.
Connection to BAA: Block length $n = 1$

For $n = 1$ we agree with the Blahut-Arimoto algorithm. Instead of steps (2), (3) we have

$$r(x) = \frac{\prod_y q(x|y)p(y|x)}{\sum_x \prod_y q(x|y)p(y|x)},$$

and step (4) is replaced by

$$q(x|y) = \frac{r(x)p(y|x)}{\sum_x r(x)p(y|x)}.$$

For $d = n$ we reduce to the same equations for greater alphabets.
Venkataramanan/Pradhan07 had introduced and solved the following general problem:

\[ X^N \xrightarrow{\text{Encoder}} W(X^N) \xrightarrow{\text{Decoder}} \hat{X}(W, X^{n-d}) \]

**Figure:** Source coding with feed forward.
For Ergodic and Stationary source, it can be shown that:

\[ R(D) = \lim_{n \to \infty} \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n,\hat{x}^n)] \leq D} \frac{1}{n} I(\hat{X}^n \to X^n), \]

and

\[ R_n(D) \searrow R(D) \quad (1) \]
For Ergodic and Stationary source, it can be shown that:

\[ R(D) = \lim_{n \to \infty} \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n, \hat{x}^n)] \leq D} \frac{1}{n} I(\hat{X}^n \to X^n), \]

and

\[ R_n(D) \preceq R(D) \quad (1) \]

Problem: solving

\[ R_n(D) = \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n, \hat{x}^n)] \leq D} I(\hat{X}^n \to X^n). \]
For Ergodic and Stationary source, it can be shown that:

\[
R(D) = \lim_{n \to \infty} \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n, \hat{x}^n)] \leq D} \frac{1}{n} I(\hat{X}^n \to X^n),
\]
and

\[
R_n(D) \downarrow R(D) \tag{1}
\]

Problem: solving

\[
R_n(D) = \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n, \hat{x}^n)] \leq D} I(\hat{X}^n \to X^n).
\]

Solution: solving

\[
R_n(D) = \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n, \hat{x}^n)] \leq D, p(\hat{x}^n||x^{n-1})} I(\hat{X}^n \to X^n),
\]
using the alternating minimization procedure over

\[
I(\hat{X}^n \to X^n) = \sum_{\hat{x}^n, x^n} p(x^n) r(\hat{x}^n|x^n) \log \frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n||x^{n-1})}.
\]
Numerical examples: Memoryless Binary symmetric channel; capacity

Consider a memoryless BSC with probability of \( p = 0.3 \)

![Figure: Binary symmetric Channel](image)

The capacity of this BSC is known to be

\[
C = 1 - H(0.3) = 0.1187.
\]
Numerical examples: Memoryless Binary symmetric channel; capacity

Shannon feedback does not increase capacity of memoryless channels with feedback

Our algorithm performance (for block length $n = 5$).

Figure: Performance of our algorithm over BSC(0.3)
Estimators for the feedback channel capacity

- Bounds: P./Weissman/Goldsmith09 showed-

\[
\overline{C}_n = \max_{s_0} \max_{r(x^n||y^{n-1})} \frac{1}{n} I(X^n \rightarrow Y^n|s_0) + \frac{1}{n}.
\]

\[
\underline{C}_n = \max_{r(x^n||y^{n-1})} \min_{s_0} \frac{1}{n} I(X^n \rightarrow Y^n|s_0) - \frac{1}{n}.
\]
Estimators for the feedback channel capacity

- **Bounds:** P./Weissman/Goldsmith09 showed-

\[
\overline{C}_n = \max_{s_0} \max_{r(x^n||y^{n-1})} \frac{1}{n} I(X^n \rightarrow Y^n | s_0) + \frac{1}{n}.
\]

\[
\underline{C}_n = \max_{r(x^n||y^{n-1})} \min_{s_0} \frac{1}{n} I(X^n \rightarrow Y^n | s_0) - \frac{1}{n}.
\]

- **Directed information rate**

**Lemma**

*If* \( C_n \rightarrow C \) *then*

\[
\lim_{n \rightarrow \infty} nC_n - (n - 1)C_{n-1} = C.
\]
Estimators for the feedback channel capacity

- Bounds: P./Weissman/Goldsmith09 showed-

\[
\bar{C}_n = \max_{s_0} \max_{r(x^n||y^{n-1})} \frac{1}{n} I(X^n \to Y^n|s_0) + \frac{1}{n}.
\]
\[
C_n = \max_{r(x^n||y^{n-1})} \min_{s_0} \frac{1}{n} I(X^n \to Y^n|s_0) - \frac{1}{n}.
\]

- Directed information rate

**Lemma**

*If* \( C_n \to C *then*

\[
\lim_{n \to \infty} nC_n - (n - 1)C_{n-1} = C.
\]

- We use our algorithm to compute both bounds and estimator.
The trapdoor channel is a model introduced by David Blackwell in 1961.

![Trapdoor Channel Diagram](image)

**Figure:** Trapdoor Channel

Each ball has same probability.
2 states trapdoor channel with feedback; capacity

P./Cuff/Van Roy/Weissman08 showed that, $C = 0.69424191$.

Here $C_{12} = 0.6706533$ and $12C_{12} - 11C_{11} = 0.6942285$.

The lower bound: use $r(x^n || y^{n-1})$ that achieves $C_n$, and

$$C^* = \min_{s_0} I(X^n \rightarrow Y^n).$$

Figure: (a) Bounds for capacity. (b) Directed information rate.
We can consider the following extension to the trap door channel.

**Figure:** Trapdoor channel with $M$ states.
$M = 3$ states trapdoor channel; capacity

Our algorithm performance:

Figure: Directed information rate estimator.

Capacity estimation: $C \sim 0.542$
Effect of the number of cells on the capacity

Cell number from 1 to 15; directed information rate estimator for \( n = 12 \).

\[ 12C_{12} - 11C_{11} \]

Figure: Change of \( 12C_{12} - 11C_{11} \) over the number of cells in the trapdoor channel with feedback with delay 1.
The Ising channel modulates intersymbol interference.

\[ x_n = 0 \quad \begin{array}{c}
\begin{array}{c}
0 \\
1
\end{array}
\end{array} \quad 0 \quad \begin{array}{c}
\begin{array}{c}
\frac{1}{2} \\
\frac{1}{2}
\end{array}
\end{array} \quad 1 \quad x_n+1 \]

\[ x_n = 1 \quad \begin{array}{c}
\begin{array}{c}
0 \\
1
\end{array}
\end{array} \quad 0 \quad \begin{array}{c}
\begin{array}{c}
\frac{1}{2} \\
\frac{1}{2}
\end{array}
\end{array} \quad 0 \quad y_n \quad x_n+1 \]

\[ x_n = 1 \quad \begin{array}{c}
\begin{array}{c}
0 \\
1
\end{array}
\end{array} \quad 0 \quad \begin{array}{c}
\begin{array}{c}
\frac{1}{2} \\
\frac{1}{2}
\end{array}
\end{array} \quad 0 \quad y_n \]

\[ x_n = 0 \quad \begin{array}{c}
\begin{array}{c}
0 \\
1
\end{array}
\end{array} \quad 0 \quad \begin{array}{c}
\begin{array}{c}
\frac{1}{2} \\
\frac{1}{2}
\end{array}
\end{array} \quad 1 \quad y_n \]

**Figure:** The Ising Channel.

The channel (without feedback) was introduced by Berger/Bonomi90.
Ising channel; effects of the delay on the capacity

Delay from 2 to 12; directed information rate estimator for \( n = 12 \).

![Figure: Change of \( 12C_{12} - 11C_{11} \)](image)
Consider an i.i.d. source distributed $X \sim B(0.5)$. Weissman/Merhav03 showed that feed forward does not decrease $R(D)$. Our algorithm performance (with block length $n = 5$) is presented here.

**Figure:** Rate distortion function; red line is the theoretical calculation
Consider the following Markov source

\[ p = 0.3, \; q = 0.2, \text{ and a single letter distortion function} \]

\[ d(x^n, \hat{x}^n) = \sum_{i=1}^{n} d(x_i, \hat{x}_i), \]

\[ X_0 \text{ has stationary distribution } [0.4, 0.6]. \]
The rate distortion function for this example can be analytically calculated, and shown here by the red line.

Figure: The arrow indicates the way $R_n(D)$ responds to $n$ increasing.

This example’s solution is very similar to the solution of Weissman and Merhav, for the case where $p = q$. 
Consider the Markov source from previous example, and the distortion function given by

<table>
<thead>
<tr>
<th>Table: Distortion $e(\hat{x}<em>i, x</em>{i-1}, x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{x}_i$</td>
</tr>
<tr>
<td>$\hat{x}_i = 0$</td>
</tr>
<tr>
<td>$\hat{x}_i = 1$</td>
</tr>
</tbody>
</table>

and

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} e(\hat{x}_i, x_{i-1}, x_i).$$

This problem was introduced and solved by Venkataramanan/Pradhan07.
Markov source with general distortion function; $R(D)$

For the parameters $p = 0.3$, $q = 0.2$, our algorithm performance is presented here.

Figure: $R(D)$ for the stock market example

(a) Graph of $R_n(D)$. (b) Graph of $12D_{12}(R) - 11D_{11}(R)$.

Red line is the theoretical calculation
We showed that we can estimate and bound the channel capacity with feedback using-

\[ C_n = \frac{1}{n} \max_{p(x^n|z^{n-d})} I(X^n \to Y^n), \]

and rate distortion with feed forward using-

\[ R_n(D) = \frac{1}{n} \min_{p(\hat{x}^n|x^n): \mathbb{E}[d(x^n,\hat{x}^n)] \leq D} I(\hat{X}^n \to X^n). \]

We solve these optimization problems by using an alternating optimization procedure, thus obtaining a numerical solution.

Thank You!