Optimization of the Directed Information

Haim Permuter

Ben-Gurion University, Israel

1st Munich Workshop on Bidirectional Communication and Directed Information
May 2012
Notation

**Causal Conditioning pmf**

\[
P(y^n || x^n) \triangleq \prod_{i=1}^{n} P(y_i | x^i, y^{i-1})
\]

\[
P(y^n | x^n) = \prod_{i=1}^{n} P(y_i | x^n, y^{i-1})
\]

**Causal Conditioning entropy**

\[
H(Y^n || X^n) \triangleq E[ - \log P(Y^n || X^n) ]
\]

\[
H(Y^n | X^n) \triangleq E[ - \log P(Y^n | X^n) ]
\]

**Directed Information**

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

\[
I(X^n; Y^n) \triangleq H(Y^n) - H(Y^n | X^n)
\]
**Notation**

**Causal Conditioning pmf**

\[
P(y^n \mid x^n) \triangleq \prod_{i=1}^{n} P(y_i \mid x^i, y^{i-1})
\]

\[
P(y^n \mid x^{n-1}) \triangleq \prod_{i=1}^{n} P(y_i \mid x^{i-1}, y^{i-1})
\]

**Causal Conditioning entropy**

\[
H(Y^n \mid X^n) \triangleq E[-\log P(Y^n \mid X^n)]
\]

\[
H(Y^n \mid X^{n-1}) \triangleq E[-\log P(Y^n \mid X^{n-1})]
\]

**Directed Information**

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

\[
I(X^{n-1} \rightarrow Y^n) \triangleq H(Y^n) - H(Y^n \mid X^{n-1})
\]
Directed information and causal conditioning characterizes

1. rate reduction in **losless compression** due to causal side information at the decoder,

2. the gain in growth rate in **horse-race gambling** due to causal side information

3. **channel capacity** with feedback,

4. **rate distortion** with feedforward,

5. **causal MMSE** for additive Gaussian noise,

6. **stock investment** with causal side information,

7. measure of **causal relevance** between processes,

8. **actions with causal constraint** such as “to feed or not to feed back”,

H. Permuter  Optimization of the Directed Information
Directed information optimization

How to find

$$\max_{p(x^n \mid y^{n-1})} I(X^n \rightarrow Y^n).$$

Recall

$$I(X^n \rightarrow Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1})$$

$$= H(Y^n) - H(Y^n \mid X^n)$$

$$= \sum_{y^n, x^n} p(x^n, y^n) \log \frac{p(y^n \mid x^n)}{p(y^n)}$$

$$P(x^n, y^n)$$ can be expressed by the chain-rule

$$p(x^n, y^n) = p(x^n \mid y^{n-1})p(y^n \mid x^n)$$
Property of the optimization problem

\[
\max_{p(x^n||y^{n-1})} I(X^n \rightarrow Y^n)
\]

Good news

- \( I(X^n \rightarrow Y^n) \) is convex in \( p(x^n||y^{n-1}) \) for a fixed \( p(y^n||x^n) \).
- \( p(x^n||y^{n-1}) \) is a convex set.
Property of the optimization problem

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

**Good news**
- \(I(X^n \to Y^n)\) is convex in \(p(x^n||y^{n-1})\) for a fixed \(p(y^n||x^n)\).
- \(p(x^n||y^{n-1})\) is a convex set.

**Bad news**
- Not easy to describe \(p(x^n||y^{n-1})\) using linear equations.
  Contrary to \(p(x^n)\) where
  \[
  \begin{align*}
  p(x^n) &\geq 0 \ \forall x^n. \\
  \sum_{x^n} p(x^n) &= 1.
  \end{align*}
  \]
Property of the optimization problem

\[
\max_{p(x^n || y^{n-1})} I(X^n \rightarrow Y^n)
\]

**Good news**
- \( I(X^n \rightarrow Y^n) \) is convex in \( p(x^n || y^{n-1}) \) for a fixed \( p(y^n || x^n) \).
- \( p(x^n || y^{n-1}) \) is a convex set.

**Bad news**
- Not easy to describe \( p(x^n || y^{n-1}) \) using linear equations.
  - Contrary to \( p(x^n) \) where
    \[
    p(x^n) \geq 0 \quad \forall x^n. \\
    \sum_{x^n} p(x^n) = 1.
    \]
  - \( I(X^n \rightarrow Y^n) \) non-convex in \( p(x_1), \ldots, p(x_n | x^{n-1}, y^{n-1}) \)
Property of the optimization problem

$$\max_{p(x^n||y^{n-1})} I(X^n \rightarrow Y^n)$$

**Good news**

- $I(X^n \rightarrow Y^n)$ is convex in $p(x^n||y^{n-1})$ for a fixed $p(y^n||x^n)$.
- $p(x^n||y^{n-1})$ is a convex set.

**Bad news**

- Not easy to describe $p(x^n||y^{n-1})$ using linear equations.
  Contrary to $p(x^n)$ where

$$p(x^n) \geq 0 \quad \forall x^n.$$

$$\sum_{x^n} p(x^n) = 1.$$

- $I(X^n \rightarrow Y^n)$ non-convex in $p(x_1), \ldots, p(x_n|x^{n-1}, y^{n-1})$
- Cannot optimize each term in $\sum_i I(X^i; Y_i|Y^{i-1})$ or in $\sum_i I(X_i; Y^n_i|X^{i-1}, Y^{i-1})$, separately.
The Alternating maximization procedure

Lemma (double maximization)

\[
\max_{p(x^n \| y^{n-1})} I(X^n \rightarrow Y^n) = \max_{p(x^n \| y^{n-1}), q(x^n | y^n)} I(X^n \rightarrow Y^n).
\]
The Alternating maximization procedure

Lemma (double maximization)

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

Let \( f(u_1, u_2) \), be a convex fun and we want to find

\[
\max_{u_1 \in A_1, u_2 \in A_2} f(u_1, u_2).
\]

The procedure is

\[
u_1^{(k+1)} = \arg \max_{u_1 \in A_1} f(u_1^{(k)}, u_2^{(k)}), \quad u_2^{(k+1)} = \arg \max_{u_2 \in A_2} f(u_1^{(k+1)}, u_2^{(k)}).
\]

\[
f^{(k)} = f(u_1^{(k)}, u_2^{(k)}).
\]

Theorem (The Alternating maximization procedure)

\[
\lim_{k \rightarrow \infty} f^{(k)} = \max_{u_1 \in A_1, u_2 \in A_2} f(u_1, u_2).
\]
Compute by the alternating maximization procedure

\[
\max_{p(x^n || y^{n-1})} \max_{q(x^n | y^n)} I(X^n \rightarrow Y^n).
\]
Compute by the alternating maximization procedure

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

1st Step

**Lemma \((\max_{q(x^n \| y^n)} I(X^n \rightarrow Y^n))\)**

*For fixed \(p(x^n \| y^{n-1})\), \(q^*(x^n \| y^n)\) that achieves \(\max_{q(x^n \| y^n)} I(X^n \rightarrow Y^n)\), is*

\[
q^*(x^n \| y^n) = \frac{p(x^n \| y^{n-1})p(y^n \| x^n)}{\sum_{x^n} p(x^n \| y^{n-1})p(y^n \| x^n)}. 
\]
Lemma \((\max_p (x^n \parallel y^{n-1}) I(X^n \rightarrow Y^n))\)

For fixed \(q(x^n|y^n)\), \(p^*(x^n \parallel y^{n-1})\) that achieves \(\max_p (x^n \parallel y^{n-1}) I(X^n \rightarrow Y^n)\), is:

Starting from \(i = n\), compute \(p(x_i|x_{i-1}, y_{i-1})\)

\[
p_i = p^*(x_i|x_{i-1}, y_{i-1}) = \frac{p'(x^i, y^{i-1})}{\sum x_i p'(x^i, y^{i-1})},
\]

where

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

and do so \textbf{backwards} until \(i = 1\).
Main ideas of 2nd Step

- Exchange $p(x^n \| y^{n-1})$ by the set $\{p_i\}_{i=1}^n$ where
  \[ p_i = p(x_i | x_{i-1}, y_{i-1}) \]

\[
\max_{p(x^n \| y^{n-1})} \ I(X^n \rightarrow Y^n) = \max_{p_1} \max_{p_2} \ldots \max_{p_n} \ I(X^n \rightarrow Y^n)
\]
Main ideas of 2nd Step

- Exchange $p(x^n \Vert y^{n-1})$ by the set $\{p_i\}_{i=1}^n$ where $p_i = p(x_i | x^{i-1}, y^{i-1})$

\[
\max_{p(x^n \Vert y^{n-1})} I(X^n \rightarrow Y^n) = \max_{p_1} \max_{p_2} \ldots \max_{p_n} I(X^n \rightarrow Y^n)
\]

- $I(X^n \rightarrow Y^n)$ is concave in each $p_i$. 

H. Permuter
Optimization of the Directed Information
Main ideas of 2nd Step

- Exchange \( p(x^n \parallel y^{n-1}) \) by the set \( \{p_i\}_{i=1}^n \) where
  \[ p_i = p(x_i|x^{i-1}, y^{i-1}) \]

- \( I(X^n \rightarrow Y^n) \) is concave in each \( p_i \).

- For fixed \( q(x^n|y^n) \), \( p_i^* \) that achieves \( \max_{p_i} I(X^n \rightarrow Y^n) \), depends only on
  \[ q(x^n|y^n), p_{i+1}, p_{i+2}, \ldots, p_n \]
**Main ideas of 2nd Step**

- **Exchange** $p(x^n \| y^{n-1})$ by the set $\{p_i\}_{i=1}^n$ where
  
  $$p_i = p(x_i | x_{i-1}, y_{i-1})$$

  
  $$\max_{p(x^n \| y^{n-1})} I(X^n \rightarrow Y^n) = \max_{p_1} \max_{p_2} ... \max_{p_n} I(X^n \rightarrow Y^n)$$

- $I(X^n \rightarrow Y^n)$ is concave in each $p_i$.

- For fixed $q(x^n \| y^n)$, $p_i^*$ that achieves $\max_{p_i} I(X^n \rightarrow Y^n)$, depends **only on**
  
  $$q(x^n \| y^n), p_{i+1}, p_{i+2}, ..., p_n$$

- Hence we can find
  
  $$\max_{p_1} ... \left( \max_{p_{n-1}} \left( \max_{p_n} I(X^n \rightarrow Y^n) \right) \right)$$

  despite being nonconvex.
Using Step 1 and 2 we can compute

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

which converges from below to \( \max_{p(x^n \| y^{n-1})} I(X^n \rightarrow Y^n) \)

We also have an upper bound

\[
I_U = \max_{x_1} \sum_{y_1} \max_{x_2} \cdots \sum_{y_{n-1}} \max_{x_n} \sum_{y_n} p(y^n \| x^n) \log \frac{p(y^n \| x^n)}{\sum_{x'_{n}} p(y^n \| x'_{n}) p(x'_{n} \| y^{n-1})}.
\]

The algorithm terminate when

\[
|I_U - I_L| \leq \epsilon
\]
maximizing the directed information for BSC(0.3)

Value

Iteration

Upper bound, $I_U$

Lower bound, $I_L$

True capacity

H. Permuter

Optimization of the Directed Information
Bounds on capacity of any FSC

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

\[
C_n = \max_{p(x^n \| y^{n-1})} \min_{s_0} \frac{1}{n} I(X^n \rightarrow Y^n | s_0) - \frac{1}{n}.
\]
Directed information rate

\[ I(X^{n+1} \rightarrow Y^{n+1}) - I(X^n \rightarrow Y^n) \]

True cap.
Infinite-letter case

For two cases we have analytical solution using dynamic programming for unifilar channels. 

First case: Trapdoor channel.

\[ C_{fb} = \log \phi \] 

Golden Ratio: \( \phi = \frac{\sqrt{5} + 1}{2} \)

(a) Ash book

(b) D. Blackwell

Fig. 7.1 A simple two-state channel.
Ising Channel

- Introduced by Berger and Bonomi [1990].
Ising Channel

- Introduced by Berger and Bonomi [1990].
- If $x_t = x_{t-1}$, then $y_t = x_t$.
- If $x_t \neq x_{t-1}$, then $Y_t \sim Bernoulli\left(\frac{1}{2}\right)$.
Ising Channel

- Introduced by Berger and Bonomi [1990].
- If $x_t = x_{t-1}$, then $y_t = x_t$.
- If $x_t \neq x_{t-1}$, then $Y_t \sim \text{Bernoulli}(\frac{1}{2})$.

The Ising channel graphical model:
Ising Channel

- Introduced by Berger and Bonomi [1990].
- If \( x_t = x_{t-1} \), then \( y_t = x_t \).
- If \( x_t \neq x_{t-1} \), then \( Y_t \sim Bernoulli(\frac{1}{2}) \).

The Ising channel graphical model:

Q: How can one achieve \( R = \frac{1}{2} \)?
Ising channel

- Simple model for inference inter-symbol.
Ising channel

- Simple model for inference inter-symbol.
- The zero-error capacity of the Ising channel is 0.5 bit per channel use.
Ising channel

- Simple model for inference inter-symbol.
- The zero-error capacity of the Ising channel is 0.5 bit per channel use.
- The capacity \textit{without feedback} found to be bounded approximately by $0.5031 \leq C \leq 0.6723$. 
Ising channel

- Simple model for inference inter-symbol.
- The zero-error capacity of the Ising channel is 0.5 bit per channel use.
- The capacity *without feedback* found to be bounded approximately by $0.5031 \leq C \leq 0.6723$.
- The feedback capacity is $C = \max_{0 \leq a \leq 1} \frac{2H(a)}{3+a} \approx 0.575522$, where $z \approx 0.4503$. 
Ising channel

- Simple model for inference inter-symbol.
- The zero-error capacity of the Ising channel is 0.5 bit per channel use.
- The capacity without feedback found to be bounded approximately by $0.5031 \leq C \leq 0.6723$.
- The feedback capacity is $C = \max_{0 \leq a \leq 1} \frac{2H(a)}{3+a} \approx 0.575522$, where $z \approx 0.4503$.
- We formulate an equivalent problem using dynamic programming (DP).
Ising channel

- Simple model for inference inter-symbol.
- The zero-error capacity of the Ising channel is 0.5 bit per channel use.
- The capacity *without feedback* found to be bounded approximately by $0.5031 \leq C \leq 0.6723$.
- The feedback capacity is $C = \max_{0 \leq a \leq 1} \frac{2H(a)}{3+a} \approx 0.575522$, where $z \approx 0.4503$.
- We formulate an equivalent problem using dynamic programming (DP).
- The DP leads to a simple capacity achieving coding scheme.
Channel notation and DP formulation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>Time $\in \mathbb{N}$</td>
</tr>
<tr>
<td>$x_t$</td>
<td>Channel Input at time $t \in \mathcal{X}$</td>
</tr>
<tr>
<td>$s_t \equiv x_{t-1}$</td>
<td>Channel State at time $t \in \mathcal{S}$</td>
</tr>
<tr>
<td>$y_t$</td>
<td>Channel Output at time $t \in \mathcal{Y}$</td>
</tr>
</tbody>
</table>
## Channel notation and DP formulation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>Time $(\in \mathbb{N})$</td>
</tr>
<tr>
<td>$x_t$</td>
<td>Channel Input at time $t$ $(\in \mathcal{X})$</td>
</tr>
<tr>
<td>$s_t(=x_{t-1})$</td>
<td>Channel State at time $t$ $(\in \mathcal{S})$</td>
</tr>
<tr>
<td>$y_t$</td>
<td>Channel Output at time $t$ $(\in \mathcal{Y})$</td>
</tr>
</tbody>
</table>

## Ising channel

- $p(s_t = 0 | y^t)$, prob. of the channel state to be 0 given the output
- $y_t$, the channel output
- $p(x_t | s_{t-1})$, channel input prob. given the channel state at time $t-1$
- $p(s_t = 0 | y^t)$ as a function of $p(s_{t-1} = 0 | y^{t-1})$ and input dist.
- $I(X_t, S_{t-1}; Y_t | y^{t-1})$

## DP

- $z_t$, the DP state
- $w_t$, the DP disturbance
- $u_t$, the DP action
- $z_t = F(z_{t-1}, u_{t-1}, w_{t-1})$, states evolving
- $g(z_{t-1}, u_t)$, reward function
DP numerical evaluation

value fun. on the $20^{th}$ iteration, $J_{20}$

Histogram of $Z$

action parameter, $\delta$

action parameter, $\gamma$
DP and its relation to the coding scheme

<table>
<thead>
<tr>
<th></th>
<th>$z_t = p_0$</th>
<th>$z_t = p_1$</th>
<th>$z_t = p_2$</th>
<th>$z_t = p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_t = 0$</td>
<td>$z_{t+1} = p_3$</td>
<td>$z_{t+1} = p_3$</td>
<td>$z_{t+1} = p_3$</td>
<td>$z_{t+1} = p_2$</td>
</tr>
<tr>
<td>$y_t = 1$</td>
<td>$z_{t+1} = p_1$</td>
<td>$z_{t+1} = p_0$</td>
<td>$z_{t+1} = p_0$</td>
<td>$z_{t+1} = p_0$</td>
</tr>
<tr>
<td>$p(x_t = 1</td>
<td>x_{t-1} = 1)$</td>
<td>$a$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$p(x_t = 0</td>
<td>x_{t-1} = 0)$</td>
<td>irrelevant</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
### DP and its relation to the coding scheme

<table>
<thead>
<tr>
<th></th>
<th>( z_t = p_0 )</th>
<th>( z_t = p_1 )</th>
<th>( z_t = p_2 )</th>
<th>( z_t = p_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_t = 0 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_t = 1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p(x_t = 1</td>
<td>x_{t-1} = 1) )</td>
<td>( \alpha )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( p(x_t = 0</td>
<td>x_{t-1} = 0) )</td>
<td>irrelevant</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**DP states**

- **DP state** \( p_0 \)
  - D: \( p(x_t = 0 | y_t) = 0 \)
  - E: \( p(x_{t+1} = x_t) = \alpha \)

- **DP state** \( p_3 \)
  - D: \( p(x_t = 0 | y_t) = 1 \)
  - E: \( p(x_{t+1} = x_t) = \alpha \)

- **DP states** \( p_1, p_2 \)
  - E: \( x_{t+1} = x_t \)
  - D: Waits

- **DP state** \( p_0 \)
  - D: \( p(x_t = 0 | y_t) = 0 \)
  - E: \( p(x_{t+1} = x_t) = \alpha \)

- **DP state** \( p_3 \)
  - D: \( p(x_t = 0 | y_t) = 1 \)
  - E: \( p(x_{t+1} = x_t) = \alpha \)
DP and its relation to the coding scheme

- Alternate between 0 and 1 with prob. \( 1 - \alpha \).

\[
\begin{align*}
\text{DP state } p_0: & \\
\text{D: } & p(x_t = 0|y_t) = 0 \\
\text{E: } & p(x_{t+1} = x_t) = \alpha
\end{align*}
\]

\[
\begin{align*}
\text{DP state } p_3: & \\
\text{D: } & p(x_t = 0|y_t) = 1 \\
\text{E: } & p(x_{t+1} = x_t) = \alpha
\end{align*}
\]

\[
\begin{align*}
\text{DP states } p_1, p_2: & \\
\text{E: } & x_{t+1} = x_t \\
\text{D: } & \text{Waits}
\end{align*}
\]
DP and its relation to the coding scheme

- Alternate between 0 and 1 with prob. $1 - a$.
- If the output $y_{t+1} \neq s_t$, then decode $x_{t+1} = y_{t+1}$
DP and its relation to the coding scheme

- Alternate between 0 and 1 with prob. $1 - a$.
- If the output $y_{t+1} \neq s_t$, then decode $x_{t+1} = y_{t+1}$
- If the output $y_{t+1} = s_t$ repeat the last input

\[ D: \quad \begin{align*}
D: & \quad p(x_t = 0|y^t) = 0 \\
E: & \quad p(x_{t+1} = x_t) = a
\end{align*} \]

\[ E: \quad \begin{align*}
D: & \quad p(x_t = 0|y^t) = 1 \\
E: & \quad p(x_{t+1} = x_t) = a
\end{align*} \]

\[ D: \quad \begin{align*}
E: & \quad x_{t+1} = x_t \\
D: & \quad \text{Waits}
\end{align*} \]
Alternate between 0 and 1 with prob. $1 - a$.

If the output $y_{t+1} \neq s_t$, then decode $x_{t+1} = y_{t+1}$

If the output $y_{t+1} = s_t$ repeat the last input

$$C = \frac{H(1 - a)}{a \cdot 2 + (1 - a) \cdot (2 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2})} = \frac{H(a)}{\frac{3}{2} + \frac{a}{2}}$$

**DP states**

- **$p_0$**: $\begin{cases} D: p(x_t = 0|y^t) = 0 \\ E: p(x_{t+1} = x_t) = a \end{cases}$

- **$p_1, p_2$**: $\begin{cases} D: Waits \\ E: x_{t+1} = x_t \end{cases}$

- **$p_3$**: $\begin{cases} D: p(x_t = 0|y^t) = 1 \\ E: p(x_{t+1} = x_t) = a \end{cases}$
Convexity can be exploited to calculate

\[
\max_{p(x^n||y^{n-1})} I(X^n \rightarrow Y^n)
\]

using alternating maximization procedure.

DP can be formulated for Unifilar channel and numerically calculated.

For some cases, such as Trapdoor-Channel and Ising-Channel the DP can be solved analytically.

DP solution can lead to an optimal and concrete coding scheme.
Convexity can be exploited to calculate
\[
\max_{p(x^n||y^{n-1})} I(X^n \rightarrow Y^n)
\]
using alternating maximization procedure.

DP can be formulated for Unifilar channel and numerically calculated.

For some cases, such as Trapdoor-Channel and Ising-Channel the DP can be solved analytically.

DP solution can lead to an optimal and concrete coding scheme.

*Thank you very much!*