Universal Estimation of Directed Information

Jiantao Jiao
Tsinghua/Stanford
Lei Zhao
Jump Operation
Haim Permuter
Ben-Gurion
Young-Han Kim
UCSD
Tsachy Weissman
Stanford

ISIT
July 2012, Boston
Definition of Directed Information (Discrete Time)

\[
I(X^n; Y^n) \triangleq H(Y^n) - H(Y^n | X^n) = \sum_{i=1}^{n} I(X^n; Y_i | Y^{i-1})
\]

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

\[
P(y^n | x^n) = \prod_{i=1}^{n} P(y_i | x^n, y^{i-1})
\]
Definition of Directed Information (Discrete Time)

**Directed Information**

\[
I(X^n \to Y^n) \triangleq H(Y^n) - H(Y^n \| X^n) \triangleq \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1})
\]

\[
I(X^n; Y^n) \triangleq H(Y^n) - H(Y^n | X^n) = \sum_{i=1}^{n} I(X^n; Y_i | Y^{i-1})
\]

**Causal Conditioning**

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

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

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

\[
P(y^n | x^n) = \prod_{i=1}^{n} P(y_i | x^n, y^{i-1})
\]
Directed information and causal conditioning characterize

- rate reduction in *lossless compression* due to *causal* side information at the decoder  [Simeone/P12]
Directed information and causal conditioning characterize

1. Rate reduction in \textit{lossless compression} due to \textit{causal} side information at the decoder \cite{Simeone/P12}
2. Gain in growth rate in \textit{horse-race gambling} due to \textit{causal} side information \cite{P/Kim/Weissman11}
Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
Directed information and causal conditioning characterize

1. Rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. Gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. Channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. Networks capacity with feedback [Kramer98]
Directed information and causal conditioning characterize

1. rate reduction in **lossless compression** due to **causal** side information at the decoder [Simeone/P12]
2. gain in growth rate in **horse-race gambling** due to **causal** side information [P/Kim/Weissman11]
3. **channel capacity with feedback** [Kim08, Tatakonda/Mitter09, P/Weissman/Goldsmith09]
4. **networks capacity with feedback** [Kramer98]
5. **rate distortion with feedforward** [Venkataramanan/Pradhan07]
1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]
5. rate distortion with feedforward [Venkataramanan/Pradhan07]
6. causal MMSE for additive Gaussian noise [Weissman/P/Kim11]
Directed information and causal conditioning characterize

1. rate reduction in **lossless compression** due to causal side information at the decoder  [Simeone/P12]
2. gain in growth rate in **horse-race gambling** due to causal side information  [P/Kim/Weissman11]
3. **channel capacity with feedback**  [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback  [Kramer98]
5. rate distortion with feedforward  [Venkataramanan/Pradhan07]
6. causal **MMSE** for additive Gaussian noise  [Weissman/P/Kim11]
7. **stock investment** with causal side information  [P/Kim/Weissman11]
Directed information and causal conditioning characterize

1. rate reduction in **lossless compression** due to **causal** side information at the decoder  [Simeone/P12]
2. gain in growth rate in **horse-race gambling** due to **causal** side information  [P/Kim/Weissman11]
3. **channel capacity with feedback**  [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. **networks capacity with feedback**  [Kramer98]
5. **rate distortion with feedforward**  [Venkataramanan/Pradhan07]
6. **causal MMSE for additive Gaussian noise**  [Weissman/P/Kim11]
7. **stock investment with causal side information**  [P/Kim/Weissman11]
8. **actions with causal constraint** such as “to feed or not to feed back”  [Asnani/P/Weissman11]
Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder  [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information  [P/Kim/Weissman11]
3. channel capacity with feedback  [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback  [Kramer98]
5. rate distortion with feedforward  [Venkataramanan/Pradhan07]
6. causal MMSE for additive Gaussian noise  [Weissman/P/Kim11]
7. stock investment with causal side information  [P/Kim/Weissman11]
8. actions with causal constraint such as “to feed or not to feed back”  [Asnani/P/Weissman11]

Can be optimized using convex optimization tools  [Naiss/P11]
In this talk

- Contribution: We present several estimator for estimation the directed information using universal compression algorithms.
- Idea: Use a universal compression algorithm to induce a universal probability assignment, then plug in the probability assignment into the estimator.
- Result: The proposed estimator are all consistent and provide different properties (such as range, smoothness, convergence guarantees).
Casual condition, the Chain rule

Causal Conditioning

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

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

Chain Rule

\[ P(x^n, y^n) = Q(x^n|y^{n-1})P(y^n|x^{n-1}) \]
Conservation Law

\[ I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n) \] [Massey06]

Recall \( P(x^n, y^n) = P(x^n \| y^{n-1}) P(y^n \| x^n) \)

\[
I(X^n; Y^n) = \mathbb{E} \left[ \ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)} \right]
\]

\[
= \mathbb{E} \left[ \ln \frac{P(Y^n \| X^n)P(X^n \| Y^{n-1})}{P(Y^n)P(X^n)} \right]
\]

\[
= \mathbb{E} \left[ \ln \frac{P(Y^n \| X^n)}{P(Y^n)} \right] + \mathbb{E} \left[ \ln \frac{P(X^n \| Y^{n-1})}{P(X^n)} \right]
\]

\[
= I(X^n \to Y^n) + I(Y^{n-1} \to X^n).
\]
Conservation Law

\[ I(X^n; Y^n) = I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n) \]  [Massey06]

Recall

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

\[
I(X^n; Y^n) = \mathbb{E} \left[ \ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)} \right] \\
= \mathbb{E} \left[ \ln \frac{P(Y^n \| X^n)P(X^n \| Y^{n-1})}{P(Y^n)P(X^n)} \right] \\
= \mathbb{E} \left[ \ln \frac{P(Y^n \| X^n)}{P(Y^n)} \right] + \mathbb{E} \left[ \ln \frac{P(X^n \| Y^{n-1})}{P(X^n)} \right] \\
= I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n). 
\]

In case that we have \( X_i - (X^{i-1}, Z^{i-1}) - Y^{i-1} \):

\[ I(X^n; Y^n) = I(X^n \rightarrow Y^n) + I(Z^{n-1} \rightarrow X^n). \]
Conservation Law

\[ I(X^n; Y^n) = I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n) \] \[\text{[Massey06]}\]

Recall \( P(x^n, y^n) = P(x^n||y^{n-1})P(y^n||x^n) \)

\[
I(X^n; Y^n) = \mathbb{E} \left[ \ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)} \right] \\
= \mathbb{E} \left[ \ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)} \right] \\
= \mathbb{E} \left[ \ln \frac{P(Y^n||X^n)}{P(Y^n)} \right] + \mathbb{E} \left[ \ln \frac{P(X^n||Y^{n-1})}{P(X^n)} \right] \\
= I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n). \\
\]

In case that we have \( X_i - (X^{i-1}, Z^{i-1}) - Y^{i-1} \):

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

If there is no feedback, \( z_i = \text{null} \), then

\[
I(X^n; Y^n) = I(X^n \rightarrow Y^n) + 0. \\
\]
Causal influence/relevance between two sequences

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

\[ P(x_i|x^{i-1}, y^{i-1}) \]

The model implies an order \( X_1, Y_1, X_2, Y_2, X_3, Y_3, \ldots \).

- The forward link exists if and only if \( I(X^n \rightarrow Y^n) > 0 \) (\( X_i \) is “causing” \( Y_i \)).
- The backward link exists if and only if \( I(Y^{n-1} \rightarrow X^n) > 0 \) (\( Y_i \) is “causing” \( X_i \)).
Previous work on Causality

- Granger Causality [Granger69]
- Bidirectional communication [Marko73]
- Gourieroux/Monfort/Renault87

\[
I(X^{n-1} \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n) + \sum_i I(X_i; Y_i | X^{i-1}, Y^{i-1})
\]

- “Measures of mutual and causal dependence between two time series” [Rissanen/Max 87]
- Relation between Granger Causality and directed information [Quinn/Coleman/Kiyavash/Hatsopoulos10] [Quinn/Coleman/Kiyavash11] [Amblard/Michel10]
- Directed information estimation has been applied to
  - Neurobiology [Quinn/Coleman/Kiyavash/Hatsopoulos10]
  - Gene Network [Rao/Hero/States/Engel08]
  - Video Indexing [Chen/Savarese/Hero12]
Universal sequential probability assignment

- A **sequential probability** assignment $Q$ consists of a set of conditional probabilities $\{Q_{X_i|x^{i-1}}(\cdot), \forall x^{i-1} \in \mathcal{X}^{i-1}\}_{i=1}^{\infty}$.

- A sequential probability assignment $Q$ is **universal** if

  $$\limsup_{n \to \infty} \frac{1}{n} D(P_{X^n} \| Q_{X^n}) = 0$$

  for any stationary probability measure $P$.

- A source code is **universal** if each code is uniquely decodable and

  $$\limsup_{n \to \infty} \frac{1}{n} \mathbb{E} [l_n(X^n)] = H(X).$$

  for every stationary ergodic source $X$.

- Univ. compressors $\iff$ Univ. sequential probability assign.
The idea of estimating information measures using universal compressor has been used

**LZ** Lempel-Ziv, [Wyner/Ziv89], [Ziv/Merhav93]

**BWT** Burrows-Wheeler Transform [Cai/Kulkarni/Verdú04,06]

**CTW** Context Tree Weighting [Yu/Verdú06]
Context Tree Weighting (CTW)

[Willems, Shtarkov, Tjalkens, 1995], [Willems, 1998]

- Universal compressor
- Optimal convergence rates
- Linear complexity
- Explicit sequential probability assignment

x=(000)11010010 with D=3
Estimator 1

\[ \hat{I}_1(X^n \rightarrow Y^n) \triangleq \hat{H}_1(Y^n) - \hat{H}_1(Y^n||X^n) \]

where

\[ \hat{H}_1(Y^n||X^n) \triangleq -\frac{1}{n} \log Q(Y^n||X^n) = -\frac{1}{n} \sum_{i=1}^{n} \log Q(Y_i|Y^{i-1}, X^i) \]

**Theorem**

Let \( Q \) be a universal sequential probability assignment, then

\[ \lim_{n \to \infty} \hat{I}_1(X^n \rightarrow Y^n) = I(X \rightarrow Y) \text{ in } L_1. \]

Further, if \( Q \) is a pointwise universal probability assignment, then the convergence of \( \hat{I}_1(X^n \rightarrow Y^n) \) to \( I(X \rightarrow Y) \) holds almost surely.
Theorem

Let $Q$ be the universal probability assignment induced by basic CTW method, if $(X, Y)$ then there exists a constant $C_1$ such that

$$\mathbb{E}\left|\hat{I}_1(X^n \rightarrow Y^n) - I(X \rightarrow Y)\right| \leq C_1 n^{-1/2} \log n$$

and $\forall \epsilon > 0$,

$$\hat{I}_1(X^n \rightarrow Y^n) - I(X \rightarrow Y) = o(n^{-1/2}(\log n)^{5/2+\epsilon}) \text{ a.s.}$$

Under a minimax criteria (Rissanen lower bound) this is the best one can do (One can not guarantee that the error can not decrease faster than $O(n^{-1/2})$).
Estimator 1

\[ \frac{1}{n} \hat{I}_1(Y^n \rightarrow X^n) \]

- merits: algorithmic and theoretical
- erratic for small \( n \)
- unbounded range including negative values.
Second approach

Consider an estimation of entropy rate

\[ \lim_{n \to \infty} \frac{1}{n} H(X^n) \]

First,

\[ -\frac{1}{n} \log Q(X^n) = -\frac{1}{n} \sum_{i=1}^{n} \log Q(X_i|X_{i-1}) \]
Second approach

Consider an estimation of entropy rate

\[
\lim_{n \to \infty} \frac{1}{n} H(X^n)
\]

First,

\[
-\frac{1}{n} \log Q(X^n) = -\frac{1}{n} \sum_{i=1}^{n} \log Q(X_i|X_i^{i-1})
\]

Second,

\[
\frac{1}{n} \sum_{i=1}^{n} h(Q(\cdot|X_i^{i-1}))
\]

where

\[
h(P(\cdot)) = \sum_{x} -P(x) \log P(x).
\]
Estimator 2

\[ \hat{I}_2(X^n \rightarrow Y^n) \triangleq \hat{H}_2(Y^n) - \hat{H}_2(Y^n \mid| X^n) \]

where

\[ \hat{H}_2(Y^n \mid| X^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(Q_{X_{i+1}, Y_{i+1} | X^i, Y^i}(\cdot, \cdot)) \]

\[ f(P_{X,Y}) \triangleq - \sum_{x,y} P_{X,Y}(x,y) \log P_{Y|X}(y|x) \]

performance guarantees for \( \hat{I}_2 \) similar to those for \( \hat{I}_1 \)
Estimator 2

\[ \frac{1}{n} \hat{I}_2(Y^n \rightarrow X^n) \]

- merits: algorithmic, theoretical, smooth, bounded range
- can be negative

Jiao/Zhao/Permuter/Kim/Weissman
Universal estimation of Directed Information
Third approach

Estimate directed information using divergence,

\[ I(X^i; Y_i) = D(P_{X^i,Y_i} \| P_{X^i,Y_i} \times P_{X^i,Y_i}) = D(P_{Y_i|X^i} \| P_{Y_i} P_{X^i}) \]

And the directed information

\[ I(X^i; Y_i|Y^{i-1}) = D(P_{Y_i|X^i,Y^{i-1}} \| P_{Y_i|Y^{i-1}} P_{X^i,Y^{i-1}}) \]

Estimator 3:

\[ \hat{I}_3(X^n \rightarrow Y^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} D(Q_{Y_i|X^i,Y^{i-1}}(\cdot) \| Q_{Y_i|Y^{i-1}}(\cdot)) \]

Similar performance guarantees, though weaker (Stationary ergodic Markov needed) than for the previous two.
Estimator 3

\[ \frac{1}{n} \hat{I}_3(Y^n \rightarrow X^n) \]

- merits: algorithmic, theoretical, bounded range, nonnegative
- smoothness can be improved for small \( n \)
Estimator 4

\[ \hat{I}_4(X^n \to Y^n) \]

\[ \triangleq \frac{1}{n} \sum_{i=1}^{n} D(Q_{X_{i+1},Y_{i+1}|X^i,Y^i(\cdot,\cdot)} \| Q_{Y_{i+1}|Y^i(\cdot)} Q_{X_{i+1}|X^i,Y^i(\cdot)}) \]
Estimator 4

\[ \hat{I}_4(X^n \rightarrow Y^n) \equiv \frac{1}{n} \sum_{i=1}^{n} D(Q_{X_{i+1},Y_{i+1}}|X^i,Y^i(\cdot,\cdot)) \| Q_{Y_{i+1}}|Y^i(\cdot)Q_{X_{i+1}}|X^i,Y^i(\cdot)) \]

Recall the other Estimators:

\[ \hat{H}_1(X^n||Y^n) \equiv -\frac{1}{n} \sum_{i=1}^{n} \log Q(Y_i|Y^i-1,X^i) \]

\[ \hat{H}_2(Y^n||X^n) \equiv \frac{1}{n} \sum_{i=1}^{n} f(Q_{X_{i+1},Y_{i+1}}|X^i,Y^i(\cdot,\cdot)) \]

\[ \hat{I}_3(X^n \rightarrow Y^n) \equiv \frac{1}{n} \sum_{i=1}^{n} D(Q_{Y_i}|X^i,Y^i-1(\cdot)) \| Q_{Y_i}|Y^i-1(\cdot)) \]
All estimators

\[ \hat{I}_1(Y^n \rightarrow X^n) \]

\[ \hat{I}_2(Y^n \rightarrow X^n) \]

\[ \hat{I}_3(Y^n \rightarrow X^n) \]

\[ \hat{I}_4(Y^n \rightarrow X^n) \]
Computation of directed and reverse-directed information

Alg. 1

Alg. 2

Alg. 3

Alg. 4

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

Jiao/Zhao/Permuter/Kim/Weissman  Universal estimation of Directed Information
Stock market example

The Hang Seng Index (HSI) and Dow Jones Index (DJI) indexes between 1990-2011

We would like to determine who is causally influencing whom.
Mutual influence of HSI, \( \{X_i\} \), and DJI, \( \{Y_i\} \).

![Graphs showing the mutual influence of HSI and DJI over years](image-url)
Summary and Future Work

- Universal compressor used to estimate the directed information via the assignment probability that it induces.
- Four different algorithms are suggested.
- All use the probability assignment $Q(X_i, Y_i \mid X_{i-1}, Y_{i-1})$ and $Q(Y_i \mid Y_{i-1})$.
- Different properties (smoothness, range, nonnegative).
- CTW is a good universal compressor.
- $L_1$ and a.s. convergence is guaranteed.
- Future work:
  - Continuous alphabet and larger alphabet.
  - Low number of samples.
  - Applications [page ranking, biology].
Summary and Future Work

- Universal compressor used to estimate the directed information via the assignment probability that it induces.
- Four different algorithms are suggested.
- All use the probability assignment $Q(X_i, Y_i | X_{i-1}, Y_{i-1})$ and $Q(Y_i | Y_{i-1})$.
- Different properties (smoothness, range, nonnegative).
- CTW is a good universal compressor.
- $L_1$ and $a.s.$ convergence is guaranteed.
- Future work:
  - Continuous alphabet and larger alphabet.
  - Low number of samples.
  - Applications [page ranking, biology].

Thank you very much!