Continuous-Time Directed Information

Tsachy Weissman
Stanford

Young-Han Kim
UCSD

Haim Permuter
Ben-Gurion University

ITW
Oct 2011, Paraty

Outline

1. definition and properties of discrete-time directed information
2. review of results involving directed information in
   - feedback capacity
   - gambling
   - lossless compression
   - feedforwad rate distortion
   - multi-user information theory
   - causality measure
3. definition and properties of continuous-time directed information
4. continuous-time directed information in
   - estimation
   - continuous time communication

Definitions (Discrete Time)

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

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

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

Definitions (Discrete Time)

Directed Information [Massey90]

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

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

Causal Conditioning [Kramer98]

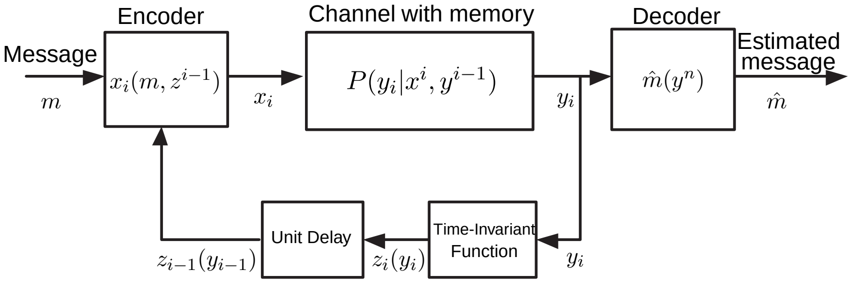
$$H(Y^n||X^n) \triangleq E[-\log P(Y^n||X^n)]$$

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

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

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

Channels with feedback



Directed information characterizes the channel capacity.

$$C = \lim_{n\to\infty} \max_{Q(x^n||z^{n-1})} \frac{1}{n} I(X^n \to Y^n)$$

[Massey90, Kramer98, Tatikonda/Mitter10, Kim10, P/Goldsmith/Weissman10]

Key Property: 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 $P(x^n, y^n) = P(x^n||y^{n-1})P(y^n||x^n)$

$$\begin{aligned}
I(X^n;Y^n) &= \mathbf{E}\left[\ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)}{P(Y^n)}\right] + \mathbf{E}\left[\ln \frac{P(X^n||Y^{n-1})}{P(X^n)}\right] \\
&= I(X^n \to Y^n) + I(Y^{n-1} \to X^n).
\end{aligned}$$

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

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

$$\begin{aligned} I(X^n;Y^n) &= \mathbf{E}\left[\ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)}\right] \\ &= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)}\right] \\ &= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)}{P(Y^n)}\right] + \mathbf{E}\left[\ln \frac{P(X^n||Y^{n-1})}{P(X^n)}\right] \\ &= I(X^n \to Y^n) + I(Y^{n-1} \to X^n). \end{aligned}$$

In case that we have deterministic feedback $z_i(y_i)$:

$$I(X^n;Y^n) = I(X^n \to Y^n) + I(Z^{n-1} \to X^n).$$

Conservation Law

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

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

$$\begin{aligned} I(X^n;Y^n) &= \mathbf{E}\left[\ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)}\right] \\ &= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)}\right] \\ &= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)}{P(Y^n)}\right] + \mathbf{E}\left[\ln \frac{P(X^n||Y^{n-1})}{P(X^n)}\right] \\ &= I(X^n \to Y^n) + I(Y^{n-1} \to X^n). \end{aligned}$$

In case that we have deterministic feedback $z_i(y_i)$:

$$I(X^n;Y^n) = I(X^n \to Y^n) + I(Z^{n-1} \to X^n).$$

If there is no feedback, $z_i = null$, then

$$I(X^n;Y^n) = I(X^n \to Y^n) + 0.$$

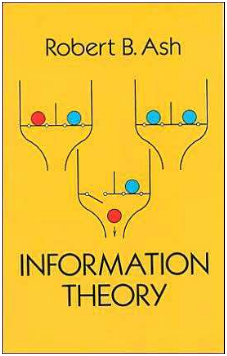
Directed Information as a functional of causal conditioning

$$\begin{aligned}
&\mathcal{I}(Q(x^n||y^{n-1}), P(y^n||x^n)) \\
&= \sum_{x^n, y^n} Q(x^n||y^{n-1}) P(y^n||x^n) \ln \frac{P(y^n||x^n)}{\sum_{x^n} Q(x^n||y^{n-1}) P(y^n||x^n)} \\
&= H(Y^n) - H(Y^n||X^n) \\
&= I(X^n \to Y^n)
\end{aligned}$$

- Causal-conditioning $Q(x^n||y^{n-1})$ and $P(y^n||x^n))$ are convex sets.
- Directed information is concave in $Q(x^n||y^{n-1})$ and convex in $P(y^n||x^n)$
- Blaut-Arimoto or Geometric programming can be used for computing $\max_{Q(x^n||y^{n-1})} I(X^n \to Y^n)$ [Naiss/P11]

Example: Blackwell's Channel/Trapdoor Channel/Chemocal Channel

Introduced by David Blackwell in 1961. [Ash65], [Ahlswede & Kaspi 87], [Ahlswede 98], [Kobayashi 02] [Berger91].



(a) Ash book

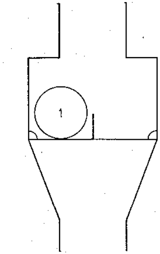


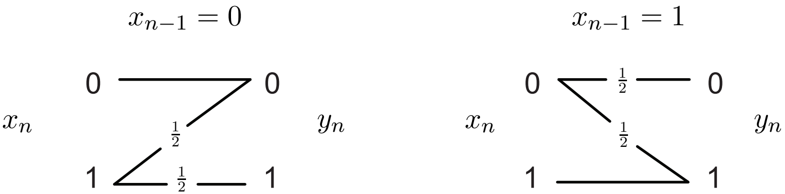
Fig. 7.1 A simple two-state channel.

(b) D. Blackwell

$C_{fb} = \log \frac{\sqrt{1+5}}{2}$, simple scheme [P/Cuff/Weissman/Van-Roy09].

Example: Ising Channel

- Introduced by Berger in 1990 and it models inter-symbol interference.
- The channel behaves as a Z-channel or S-channel, depending on the previous input.



- $C = \max_{0 \le x \le 1} \frac{2H_b(x)}{3+x}$, simple scheme [P/Elischo11].

Portfolio Theory and Gambling

Consider a horse-race market

- X_i - the horse that wins at time i.
- Y_i - side information available at time i.

(X_i, Y_i), i.i.d Kelly[56]

The optimal strategy is to invest the capital proportional to $P(x|y)$. The increase in the growth rate due to side information Y is

$$\Delta W = nI(X;Y).$$

(X_i, Y_i) general processes [P.&Kim&Weissman ISIT08]

The optimal strategy is to invest the capital proportional to $P(x_i|x^{i-1}, y^i)$. The increase in the growth rate due to *causal* side information is

$$\Delta W = I(Y^n \to X^n).$$

Lossless compression

- X_i - stationary source.
- Y_i - side information available at the encoder and *causally* at the decoder.

lossless compression with causal side information [Osvaldo/P 11]

The minimum rate needed to reconstruct the process $\{X_i\}$

$$R = \lim_{n\to\infty} \frac{1}{n} H(X^n||Y^n)$$

The reduction in the compression rate due to *causal* side information is

$$\Delta R = \lim_{n\to\infty} \frac{1}{n} (H(X^n) - H(X^n||Y^n)) = \lim_{n\to\infty} \frac{1}{n} I(Y^n \to X^n)$$

More results involving directed information

- Memoryless networks with feedback [Gerhard98]
- Rate distortion with feedforward [Weissman/Merhav03][Venkataramanan/Pradhan07][Naiss/P11]
- Broadcast with feedback [Deborah/Goldsmith10]
- Finite state MAC with feedback[P/Weissman/Chen10]
- Compound channels with feedback [Shrader/P. 10]
- Quantity causality [Colman et. al10][Zhao et al 10][Rao et al 08][Mathai et all07]...

Continuous time

In discrete time

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

How to define directed information in continuous-time?

Continuous time

In discrete time

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

How to define directed information in continuous-time?

Recall continuous-value mutual information is defined as

$$I(X;Y) = \sup_{\mathcal{Q},\mathcal{P}} I([X]_{\mathcal{Q}}; [Y]_{\mathcal{P}}),$$

where $\mathcal{Q}$, and $\mathcal{P}$ are partitions.

Main idea: use time-partition!

Time-partition

- For a continuous-time process $\{X_t\}$, let $X_a^b = \{X_s : a \leq s < b\}$.
- Let $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ denote an n-dimensional vector

$$0 \leq t_1 \leq t_2 \leq \cdots \leq t_n < T.$$

- Let $X_0^{T,\mathbf{t}}$ denote

$$X_0^{T,\mathbf{t}} = \left(X_0^{t_1}, X_{t_1}^{t_2}, \ldots, X_{t_{n-1}}^{t_n}, X_{t_n}^{T} \right).$$

- Directed information as a function of the time-partition

$$\begin{aligned} I_{\mathbf{t}} \left(X_0^T \to Y_0^T \right) &\triangleq I \left(X_0^{T,\mathbf{t}} \to Y_0^{T,\mathbf{t}} \right) \\ &= \sum_{i=1}^{n} I \left(X_0^{t_i}; Y_{t_{i-1}}^{t_i} | Y_0^{t_{i-1}} \right) \end{aligned}$$

Directed-information in continuous-time

Definition

Directed information between X_0^T *and* Y_0^T is defined as

$$I\left(X_0^T \to Y_0^T\right) \triangleq \inf_{\mathbf{t}} I_{\mathbf{t}}\left(X_0^T \to Y_0^T\right),$$

where the infimum is over all partitions $\mathbf{t}$.

Note that for continuous-value we had suprimum, and for continuous-time we use infimum.

Directed-information in continuous-time

Definition

Directed information between X_0^T *and* Y_0^T is defined as

$$I\left(X_0^T \to Y_0^T\right) \triangleq \inf_{\mathbf{t}} I_{\mathbf{t}}\left(X_0^T \to Y_0^T\right),$$

where the infimum is over all partitions $\mathbf{t}$.

Note that for continuous-value we had suprimum, and for continuous-time we use infimum.

Lemma

If $\mathbf{t}'$ *is a refinement of* $\mathbf{t}$*, then* $I_{\mathbf{t}'}\left(X_0^T \to Y_0^T\right) \leq I_{\mathbf{t}}\left(X_0^T \to Y_0^T\right)$.

Directed-information in continuous-time

Definition

Directed information between X_0^T and Y_0^T is defined as

$$I\left(X_0^T \to Y_0^T\right) \triangleq \inf_{\mathbf{t}} I_{\mathbf{t}}\left(X_0^T \to Y_0^T\right),$$

where the infimum is over all partitions $\mathbf{t}$.

Note that for continuous-value we had suprimum, and for continuous-time we use infimum.

Lemma

If $\mathbf{t}'$ is a refinement of $\mathbf{t}$, then $I_{\mathbf{t}'}\left(X_0^T \to Y_0^T\right) \leq I_{\mathbf{t}}\left(X_0^T \to Y_0^T\right)$.

For different durations we zero-pad at the beginning

$$I(X_0^{T-\delta} \to Y_0^T) := I((0_0^\delta X_0^{T-\delta}) \to Y_0^T).$$

Properties

- $I\left(X_0^T \to Y_0^T\right) \leq I\left(X_0^T; Y_0^T\right)$
- Monotonicity: $I(X_0^t \to Y_0^t)$ is monotone non-decreasing in t
- Invariance to time scaling: If ϕ is monotone strictly increasing and continuous, and $(\tilde{X}_{\phi(t)}, \tilde{Y}_{\phi(t)}) = (X_t, Y_t)$, then $I(X_0^T \to Y_0^T) = I\left(\tilde{X}_{\phi(0)}^{\phi(T)} \to \tilde{Y}_{\phi(0)}^{\phi(T)}\right)$
- Coincidence of directed and mutual information: If the Markov relation $Y_0^t - X_0^t - X_t^T$ (no feedback) holds then

$$I\left(X_0^T \to Y_0^T\right) = I\left(X_0^T; Y_0^T\right)$$

- Equivalent to discrete-time if the process is piecewise constant.

Conservation law

- If the continuity condition

$$\lim_{\delta \to 0^+} [I(X_0^\delta; Y_0^\delta) + I(X_\delta^T \to Y_\delta^T | Y_0^\delta)] = I(X_0^T \to Y_0^T)$$

holds then the directed information

$$I\left(Y_0^{T-} \to X_0^T\right) \triangleq \lim_{\delta \to 0^+} I\left(Y_0^{T-\delta} \to X_0^T\right)$$

exists and

$$I\left(X_0^T \to Y_0^T\right) + I\left(Y_0^{T-} \to X_0^T\right) = I\left(X_0^T; Y_0^T\right)$$

(this is continuous-time analogue of the conservation law$I(U^n \to V^n) + I(V^{n-1} \to U^n) = I(U^n; V^n)$)

Estimation: Duncan's Theorem

Theorem (Duncan 1970)

Let X_0^T *be a signal of finite average power* $\int_0^T E[X_t^2]dt < \infty$ *, independent of the standard Brownian motion* $\{B_t\}$ *, and let* Y_0^T *satisfy* $dY_t = X_t dt + dB_t$ *. Then*

$$\frac{1}{2}\int_0^T E\left[(X_t - E[X_t|Y_0^t])^2\right]dt = I\left(X_0^T; Y_0^T\right)$$

- relationship holds regardless of distribution of X_0^T
- The GSV theorem, $\frac{1}{2}\int_0^{\mathsf{snr}} mmse(\mathsf{snr}')\mathsf{dsnr}' = \mathsf{I}(\mathsf{snr})$, is related.

Breakdown of the Duncan Relationship

- Duncan stipulates independence between X_0^T and channel noise $\{B_t\}$
- excludes scenarios where evolution of X_t is affected by channel noise.
- for an extreme example, consider case $X_{t+\epsilon} = Y_t$
- in this case the causal MMSE

$$E\left[(X_t - E[X_t|Y_0^t])^2\right] = 0$$

while the mutual information

$$I\left(X_0^T; Y_0^T\right) = \infty$$

An Extension of Duncan's Theorem

Theorem

Let $\{B_t\}$ *be a standard Brownian motion, let* $\{W_t\}$ *be independent of* $\{B_t\}$ *, and let* $\{(X_t, Y_t)\}$ *satisfy*

$$X_t \in \sigma(Y_0^{t-}, W_0^T) \quad \text{and} \quad dY_t = X_t dt + dB_t.$$

Then, provided $\{X_t\}$ *has finite average power* $\int_0^T E[X_t^2]dt < \infty$ *,*

$$\frac{1}{2}\int_0^T E\left[(X_t - E[X_t|Y_0^t])^2\right] dt = I\left(X_0^T \to Y_0^T\right).$$

The Poisson Channel

The following is the analogue of Duncan's theorem for Poisson noise.

Theorem

Let X_0^T *be a non-negative signal satisfying* $E\int_0^T |X_t \log X_t| dt < \infty$ *and, conditioned on* X_0^T*, let* Y_0^T *be a Poisson point process with rate function* X_0^T*. Then*

$$\int_0^T E\left[\phi(X_t) - \phi\left(E[X_t|Y_0^t]\right)\right] dt = I\left(X_0^T; Y_0^T\right),$$

where $\phi(\alpha) = \alpha \log \alpha$

The Poisson Channel

The following is the analogue of Duncan's theorem for Poisson noise.

Theorem

Let X_0^T *be a non-negative signal satisfying* $E\int_0^T |X_t \log X_t| dt < \infty$ *and, conditioned on* X_0^T *, let* Y_0^T *be a Poisson point process with rate function* X_0^T *. Then*

$$\int_0^T E\left[\phi(X_t) - \phi\left(E[X_t|Y_0^t]\right)\right] dt = I\left(X_0^T; Y_0^T\right),$$

where $\phi(\alpha) = \alpha \log \alpha$

Note similarly as in Gaussian case:

- breaks down in presence of feedback
- in above theorem can replace $I\left(X_0^T; Y_0^T\right)$ by $I\left(X_0^T \to Y_0^T\right)$

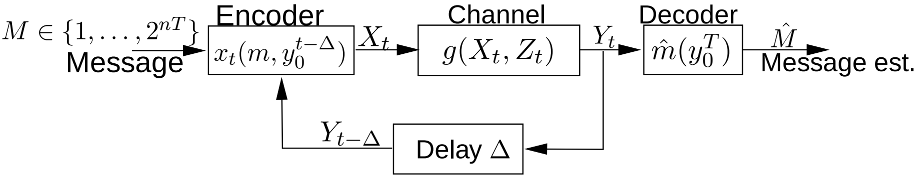
An Extended 'Duncan's Theorem' for the Poisson Channel

Theorem

Let Y_t be a point process and X_t be its $\mathcal{F}_t^Y$-predictable intensity. Then, provided $E\int_0^T |X_t \log X_t| dt < \infty$,

$$\int_0^T E\left[\phi(X_t) - \phi\left(E[X_t|Y_0^t]\right)\right] dt = I\left(X_0^T \to Y_0^T\right)$$

Continuous-time communication



- The channel of the form $Y_t = g(X_t, Z_t)$, where Z_t is a block ergodic process.
- The encoder assigns a symbol $x_t(m, y_0^{t-\Delta})$
- Message M independent of the noise process $\{Z_t\}$.

Definition

$$C(\Delta) = \sup\{R : R \text{ is achievable with feedback delay } \Delta\} \quad (1)$$

Capacity result

$$C^I(\Delta) \triangleq \lim_{T\to\infty} \frac{1}{T} \sup_{\mathcal{S}_\Delta} I(X_0^T \to Y_0^T),$$

where the supremum in is over

$$X_t = \begin{cases} g_t(U_t, Y_0^{t-\Delta}) & t \geq \Delta, \\ g_t(U_t) & t < \Delta, \end{cases}$$

The limit is shown to exist due to super-additivity.

Theorem

For this channel

$$C(\Delta) \leq C^I(\Delta),$$
$$C(\Delta) \geq C^I(\Delta') \quad \textit{for all } \Delta' > \Delta.$$

Since $C^I(\Delta)$ is a decreasing function in Δ, $C(\Delta) = C^I(\Delta)$ for any $\Delta \geq 0$ except of a set of points of measure zero.

summary

- $I(X^n; Y^n)$ amount of uncertainty about Y^n reduced by knowing X^n
- $I(X^n \to Y^n)$ amount of uncertainty about Y^n reduced by knowing X^n causally.
- Important role for discrete-time directed information in network information theory with feedback, feed-forward rate distortion, causality measure, horse-race market and lossless compression with causal side information.
- For continuous-time the idea of time-partition is useful.
- We saw an important role of directed information in continuous-time estimation and feedback-capacity.

summary

- $I(X^n; Y^n)$ amount of uncertainty about Y^n reduced by knowing X^n
- $I(X^n \to Y^n)$ amount of uncertainty about Y^n reduced by knowing X^n causally.
- Important role for discrete-time directed information in network information theory with feedback, feed-forward rate distortion, causality measure, horse-race market and lossless compression with causal side information.
- For continuous-time the idea of time-partition is useful.
- We saw an important role of directed information in continuous-time estimation and feedback-capacity.

Thank you very much!