Directed Information

Nick Landolfi and Sanjay Lall
Stanford University
Background
Information theoretic quantities

- Let $X, Y, Z \in \mathbb{R}^n$ random vectors
  - Denote elements of $X = (X(1), \ldots, X(n))$
  - Denote subvector $(X(1), \ldots, X(s))$ by $X^s$ with $X^0$ empty

- Define **entropy**
  \[ H(X) := - \mathbb{E} \log P_X \]

- Define **mutual information**
  \[ I(X, Y) := H(X) - H(X | Y) \]

- Fact: $I(X, Y) = I(Y, X)$
Information theoretic quantities

- Chain rule for entropy
  \[ H(X \mid Y) = \sum_{t=1}^{n} H(X(t) \mid X^{t-1}, Y) \]

- Define causally conditioned entropy
  \[ H(X \parallel Y) := \sum_{t=1}^{n} H(X(t) \mid X^{t-1}, Y^t) \]

- Define directed information from \( X \) to \( Y \) by
  \[ I(X \to Y) = H(Y) - H(Y \parallel X), \]
  and \( I(X \to Y) \neq I(Y \to X) \) in general

- Define directed information from \( X \) to \( Y \) causally conditioned on \( Z \) by
  \[ I(X \to Y \parallel Z) = H(Y \parallel Z) - H(Y \parallel X, Z) \]
Directed information (notation)

- Suppose $X = (X_1, \ldots, X_m)$ is $m$ stochastic processes over a time horizon $n$.
- So for $i = 1, \ldots, m$, $X_i = (X_i(1), \ldots, X_i(n)) \in \mathbb{R}^n$.
- $X$ is random object in $\mathbb{R}^{m \times n}$.
- For $A \subseteq [m]$, $X_A$ consists of $(X_i)_{i \in A} \in \mathbb{R}^{|A| \times n}$.
- Want to talk about causal relations between processes using directed information.
Directed information (sum of informations)

- $I(X_i \rightarrow X_j \parallel X_{-\{i,j\}})$ is a sum of informations

\[
I(X_i \rightarrow X_j \parallel X_{-\{i,j\}}) = H(X_j \parallel X_{-\{i,j\}}) - H(X_j \parallel X_{-\{j\}})
\]

\[
= \sum_{t=1}^{n} H(X_j(t) \mid X_{-\{j\}}^{t-1}) - H(X_j(t) \mid X_{-\{i\}}^{t-1})
\]

\[
= \sum_{t=1}^{n} I(X_j(t), X_{i}^{t-1} \mid X_{-\{i\}}^{t-1})
\]

- directed information is sum over horizon of information between $X_j$ at current time and history of $X_i$

- if informations on right hand side are large, so is directed information

- condition on histories of all other processes
Directed information (regret between predictors)

- build sequence of predictors $p_t : \mathbb{R}^{m \times (t-1)} \rightarrow \Delta(R)$.
  - map signals histories to distributions over $X_j(t)$
  - have access to all signals
- build sequence of predictors $q_t : \mathbb{R}^{(m-1) \times (t-1)} \rightarrow \Delta(R)$
  - have access to all signals except $X_i$
- measure quality of predictor by loss $\ell : \Delta(R) \times \mathbb{R} \rightarrow \mathbb{R}_+$
- measure regret with respect to loss between $p_t$ and $q_t$

$$E \left[ \sum_{i=1}^{n} \ell(q_t(X_{\{i\}}^{t-1}), X_j(t)) - \ell(p_t(X_{\{i\}}^{t-1}), X_j(t)) \right]$$

- class of predictors $q_t$ has more information than predictors $p_t$, so

$$\inf_{q_t} E \sum_{i=1}^{n} \ell(q_t(X_{\{i\}}^{t-1}), X_j(t)) > \inf_{p_t} E \sum_{i=1}^{n} \ell(p_t(X_{\{i\}}^{t-1}), X_j(t))$$
Directed information (regret between predictors)

- consider $\ell(p_t, \alpha) = -\log p_t(\alpha)$, the *negative log likelihood*

- the regret is

$$
E \sum_{t=1}^{n} \log \frac{p_t(X_j(t) | X_{-\{i\}}^{t-1})}{q_t(X_j(t) | X_{-\{i\}}^{t-1})}
$$

- select predictors $p_t = P(X_j(t) | X_{-\{i\}}^{t-1})$ and $q_t = P(X_j(t) | X_{-\{i\}}^{t-1})$, the true conditionals, regret is

$$
E \sum_{t=1}^{n} \log \frac{P(X_j(t) | X_{-\{i\}}^{t-1})}{P(X_j(t) | X_{-\{i\}}^{t-1})} \overset{(*)}{=} I(X_i \rightarrow X_j \parallel X_{\{i,j\}})
$$

$(*)$ requires proof, next slide

- directed information quantifies how much the history of $X_i$ helps to predict $X_j$
Directed information (regret between predictors)

- expanding directed information according to the definition yields

\[ I(X_i \rightarrow X_j \mid X_{-\{i,j\}}) = H(X_j \mid X_{-\{i,j\}}) - H(X_j \mid X_{-\{j\}}) \]

\[ = \sum_{t=1}^{n} H(X_j(t) \mid X_{-\{i\}}^{t-1}) - \sum_{t=1}^{n} H(X_j(t) \mid X^{t-1}) \]

\[ = E \sum_{t=1}^{n} - \log P(X_j(t) \mid X_{-\{i\}}^{t-1}) + \log P(X_j(t) \mid X^{t-1}) \]

\[ = E \sum_{t=1}^{n} \log \frac{P(X_j(t) \mid X_{-\{i\}}^{t-1})}{P(X_j(t) \mid X^{t-1})} \]

as desired
Directed information graph

- let $X$ a set of $m$ stochastic processes of length $n$
- let $G = (V, E)$ a directed graph where
  - $V = [m]$
  - and $(i, j) \in E$ if $I(X_i \rightarrow X_j \mid X_{\{i,j\}}) > 0$
- we call $G$ the \textit{directed information graph} of $X$
- generalization of linear dynamical graph
  - edge from $i$ to $j$ if z-transform of linear response has non-zero entry $j, i$
Random variable case

- let $X$ a random vector $(X_1, \ldots, X_m)$
- build predictors $p : \mathbb{R}^{m-1} \to \Delta(\mathbb{R})$ and $q : \mathbb{R}^{m-2} \to \Delta(\mathbb{R})$
  - $p$ is a distribution for $X_j$ as function of $x_{-\{j\}}$, $q$ is a distribution for $X_j$ as function of $x_{-\{i,j\}}$
- measure quality of predictor via loss $\ell : \Delta(\mathbb{R}) \times \mathbb{R} \to \mathbb{R}_+$
  - $\inf_q E[\ell(q, x_j)] > \inf_p E[\ell(p, x_j)]$
  - study expected regret of $q$ with respect to $p$: $E[\ell(q, x) - \ell(p, x)]$
  - use $\ell(p, \alpha) = -\log p(\alpha)$, the negative log likelihood
- consider regret between ideal predictors, the true marginals $P(X_j \mid X_{-\{j\}})$ and $P(X_j \mid X_{-\{i,j\}})$

$$E \left[ \log \frac{P(X_j \mid X_{-\{j\}})}{P(X_j \mid X_{-\{i,j\}})} \right] = I(X_i, X_j \mid X_{-\{i,j\}})$$

- the regret of not knowing $X_i$ in building a predictor for $X_j$ is the conditional mutual information
Random variable case: equivalence

- let $X$ a random vector $(X_1, \ldots, X_m)$
- the information graph has a node for each random variable and an edge if $I(X_i, X_j \mid X_{\{i,j\}}) > 0$.
- sparsity coincides with undirected graphical model which has edge if $X_i \perp X_j \mid X_{\{i,j\}}$
- sparsity coincides with the mmse advantage
Example Application: Simple Server Model
Server Tree

- consider a simple server tree with 5 nodes
- every node required to service requests at A
  - A is a source, new requests arrive from Poisson at rate $\lambda$
  - B sends one request to C and one to D for each it request from A
  - D proxies requests to E
  - C/E serve requests, complete request at time $t$ w.p. $p \in (0, 1]$
Example Trajectory

- system state is gross and complicated (origins, paths, blocking, destination)
- system output is simple and interpretable: number of requests processed and latency
- outputs over 900 time steps, $\lambda = 3$, $p = 1$
  - at time $t = 300$, E "breaks," i.e., E goes to $p = 1/3$
  - at time $t = 600$, C "breaks," i.e., C goes to $p = 1/3$
Smoothened Output

- left: smoothed request load, can see E go up, then D go up
- right: smoothed latency of requests arriving at A
- computing the directed information on empirical data yields server tree