Directed Information

Nick Landolfi and Sanjay Lall
Stanford University

Background

Information theoretic quantities

- let $X, Y, Z \in \mathbf{R}^n$ random vectors
  - denote elements of $X = (X(1), \dots X(n))$
  - denote subvector $(X(1), \dots, X(s))$ by X^s with X^0 empty
- define *entropy*
$$H(X) := -\mathsf{E} \log P_X$$
- define *mutual information*
$$I(X, Y) := H(X) - H(X \mid Y)$$
- 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*

$$H(X \| Y) := \sum_{t=1}^{n} H(X(t) \mid X^{t-1}, Y^{t})$$

- define *directed information* from X to Y by

$$I(X \to Y) = H(Y) - H(Y \| X),$$

  and $I(X \to Y) \neq I(Y \to X)$ in general

- define directed information from X to Y *causally conditioned* on Z by

$$I(X \to Y \| Z) = H(Y \| Z) - H(Y \| X, Z)$$

Directed information (notation)

- suppose $X = (X_1, \ldots, X_m)$ is m stochastic processes over a time horizon n.
- so for $i = 1, \ldots, m$, $X_i = (X_i(1), \ldots, X_i(n)) \in \mathbf{R}^n$
- X is random object in $\mathbf{R}^{m \times n}$
- for $A \subset [m]$, X_A consists of $(X_i)_{i \in A} \in \mathbf{R}^{|A| \times n}$
- want to talk about causal relations between processes using directed information

Directed information (sum of informations)

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

$$\begin{aligned} I(X_i \to X_j \parallel X_{-\{i,j\}}) &= H(X_j \parallel X_{-\{i,j\}}) - H(X_j \parallel X_{-\{j\}}) \\ &= \sum_{t=1}^{n} H(X_j(t) \mid X_{-\{i\}}^{t-1}) - H(X_j(t) \mid X^{t-1}) \\ &= \sum_{t=1}^{n} I(X_j(t), X_i^{t-1} \mid X_{-\{i\}}^{t-1}) \end{aligned}$$

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

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

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

$$\inf_{q_t} \mathrm{E} \sum_{i=1}^{n} \ell(q_t(X^{t-1}_{-\{i\}}), X_j(t)) > \inf_{p_t} \mathrm{E} \sum_{i=1}^{n} \ell(p_t(X^{t-1}_{-\{i\}}), X_j(t))$$

Directed information (regret between predictors)

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

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

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

$$\mathsf{E} \sum_{t=1}^{n} \log \frac{P(X_j(t) \mid X^{t-1})}{P(X_j(t) \mid X_{-\{i\}}^{t-1})} \overset{(\star)}{=} I(X_i \to X_j) \parallel X_{-\{i,j\}})$$

  $(\star)$ 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

$$
\begin{aligned}
I(X_i \to X_j \,\|\, X_{-\{i,j\}}) &= H(X_j \,\|\, X_{-\{i,j\}}) - H(X_j \,\|\, X_{-\{j\}}) \\
&= \sum_{t=1}^{n} H(X_j(t) \mid X_{-\{i\}}^{t-1}) - \sum_{t=1}^{n} H(X_j(t) \mid X^{t-1}) \\
&= \mathsf{E} \sum_{t=1}^{n} -\log P(X_j(t) \mid X_{-\{i\}}^{t-1}) + \log P(X_j(t) \mid X^{t-1}) \\
&= \mathsf{E} \sum_{t=1}^{n} \log \frac{P(X_j(t) \mid X_{-\{i\}}^{t-1})}{P(X_j(t) \mid X^{t-1})}
\end{aligned}
$$

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 $I(X_i \to X_j \parallel X_{-\{i,j\}}) > 0$
- we call G the *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 $p : \mathbf{R}^{m-1} \to \Delta(\mathbf{R})$ and $q : \mathbf{R}^{m-2} \to \Delta(\mathbf{R})$
  - 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 $\ell : \Delta(\mathbf{R}) \times \mathbf{R} \to \mathbf{R}_+$
  - $\inf_q \mathsf{E}[\ell(q, x_j)] > \inf_p \mathsf{E}[\ell(p, x_j)]$
  - study expected regret of q with respect to p: $\mathsf{E}[\ell(q, x) - \ell(p, x)]$
  - 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 $P(X_j \mid X_{-\{i,j\}})$

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

- the regret of not knowing X_i in building a predictor for X_j is the *conditional mutual information*

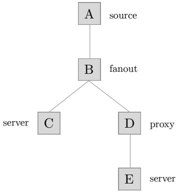
Random variable case: equivalence

- 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

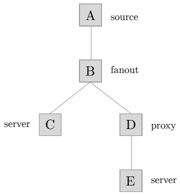
Server Tree

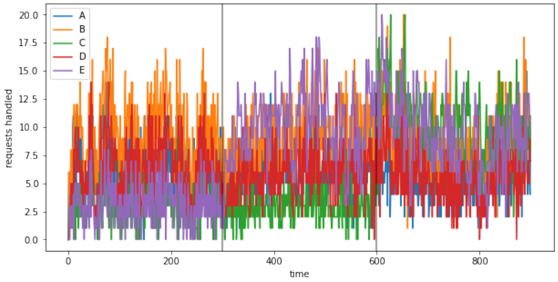
- consider a simple server tree with 5 nodes
- every node required to service requests at A
  - A is a source, new requests arrive from Poisson at rate λ
  - B sends one request to C and one to D for each it request from A
  - D proxies requests to E
  - C/E serve requests, complete request at time t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



Example Trajectory

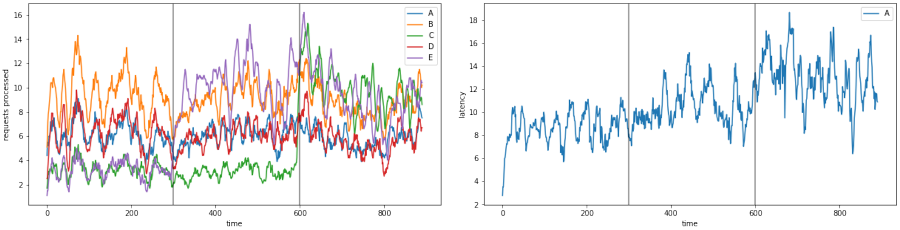
- system state is gross and complicated (origins, paths, blocking, destination)
- system output is simple and interpretable: number of requests processed and latency
- outputs over 900 time steps, $\lambda = 3$, $p = 1$
  - at time $t = 300$, E "breaks," *i.e.*, E goes to $p = 1/3$
  - at time $t = 600$,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





Smoothed 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