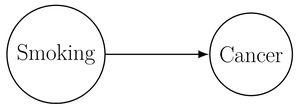
Probabilistic Active Learning of Functions in Structural Causal Models

Paul K. Rubenstein [1,2] Ilya Tolstikhin [1] Philipp Hennig [1] Bernhard Schölkopf [1]

[1]Max Planck Institute for Intelligent Systems, Tübingen [2]Machine Learning Group, University of Cambridge

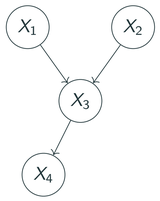
Motivation

What should we do after causal discovery? Understand **relationships** between variables for control.



Would it be better to reduce each smoker's consumption by 50%, or stop 50% of smokers completely?

Structural Causal Models (SCMs)



$$X_1 = f_1 + E_1$$
$$X_2 = f_2 + E_2$$
$$X_3 = f_3(X_1, X_2) + E_3$$
$$X_4 = f_4(X_3) + E_4$$
$$\mathbf{E} \sim \mathcal{N}(0, \Lambda)$$

An SCM $\mathcal{M}_X = (\mathcal{S}_X, \mathbb{P}_E)$ consists of

- $\mathcal{S}_X$ a set of structural equations $X_i = f_i(X, E_i)$;
- $\mathbb{P}_E$ is a distribution over the exogenous variables E;

We **further assume** additive Gaussian noise with diagonal covariance.
We consider *perfect interventions*, e.g. $\mathrm{do}(X_3 = 0)$

Problem Setup

We assume the **causal graph is known**.

Problem 1: Estimating $\mathcal{M}$ from data $\mathcal{D}$ consisting of draws $X^{(n)} \sim \mathbb{P}_X^{\mathrm{do}(i_n)}$ from various interventions i_n.

Problem 2: Active learning. Choose informative intervention so that next estimate $\widehat{\mathcal{M}}$ is closer.

How to define closeness between $\mathcal{M}$ and $\widehat{\mathcal{M}}$?

$$L_n(\hat{f}_n || f_n) = \int_{\mathcal{X}_{\mathrm{pa}(n)}} \left(f_n(x) - \hat{f}_n(x)\right)^2 d\Pi_n(x), \quad L(\hat{f}||f) = \sum_{n=1}^{N} \alpha_n L_n(\hat{f}_n||f_n)$$

$$L^i_{\mathrm{KL}}(\hat{f}||f) = \mathrm{KL}\left[\mathbb{P}_X^{\mathrm{do}(i)} || \hat{\mathbb{P}}_X^{\mathrm{do}(i)}\right], \qquad L^{\mathcal{I}'}_{\mathrm{KL}}(\hat{f}||f) = \sup_{i \in \mathcal{I}'} L^i_{\mathrm{KL}}(\hat{f}||f)$$

$$L^i_{\mathrm{MMD}_l}(\hat{f}||f) = \mathrm{MMD}_l\left[\mathbb{P}_X^{\mathrm{do}(i)} || \hat{\mathbb{P}}_X^{\mathrm{do}(i)}\right], \qquad L^{\mathcal{I}'}_{\mathrm{MMD}_l}(\hat{f}||f) = \sup_{i \in \mathcal{I}'} L^i_{\mathrm{MMD}_l}(\hat{f}||f)$$

How can we efficiently learn functional relationships between variables, once causal structure is known?

A Probabilistic Approach to Function Estimation

Assume Gaussian Process prior over each $f_n \sim \mathcal{GP}(0, k_n)$.

Let $\mathcal{D}_n \subseteq \mathcal{D}$ be observations of $(X_{\mathrm{pa}(n)}, X_n)$ for i in which X_n is not intervened.
Posterior:

$$f_n | \mathcal{D}_n \sim \mathcal{GP}(\mu_{f_n|\mathcal{D}_n}, k_{f_n|\mathcal{D}_n})$$

Which $\hat{f}$ should be chosen, given the posterior over f?

Lemma 1:

$$\mathbb{E}_{f|\mathcal{D}}\left[L(\hat{f}||f)\right] = \sum_{n=1}^{N} \alpha_n \int_{\mathcal{X}_{pa(n)}} \left(\hat{f}_n(x) - \mu_{f_n|\mathcal{D}_n}(x)\right)^2 + k_{n|\mathcal{D}_n}(x,x) d\Pi_n(x)$$

Lemma 2: Let $\mu_{f|\mathcal{D}}$ be the tuple of functions $(\mu_{f_n|\mathcal{D}_n})_{n=1,\dots,N}$. Then

$$\mu_{f|\mathcal{D}} = \arg\min_{\hat{f}} \mathbb{E}_{f|\mathcal{D}}\left[L(\hat{f}||f)\right]$$

Expected total risk: $\mathcal{R}(\mathcal{D}) = \mathbb{E}_{f|\mathcal{D}}\left[L(\mu_{f|\mathcal{D}}||f)\right]$

Active Learning

Goal: choose intervention i so that $\mathcal{R}(\mathcal{D} \cup (i,x))$ is minimised, taking into account cost of intervention.

$$V(i|\mathcal{D}) = \frac{\mathcal{R}(\mathcal{D}) - \mathbb{E}_{x \sim \mathbb{P}_X^{\mathrm{do}(i)}} \mathcal{R}(\mathcal{D} \cup \{(i,x)\})}{c(i)}$$

Problem: $\mathbb{P}_X^{\mathrm{do}(i)}$ is unknown! Solution: replace with *estimate* $\widetilde{\mathbb{P}}_X^{\mathrm{do}(i)}$, from which we can sample.

Algorithm 1: Estimate $\mathbb{E}_{x \sim \mathbb{P}_X^{\mathrm{do}(i)}} \mathcal{R}(\mathcal{D} \cup \{(i,x)\})$ as

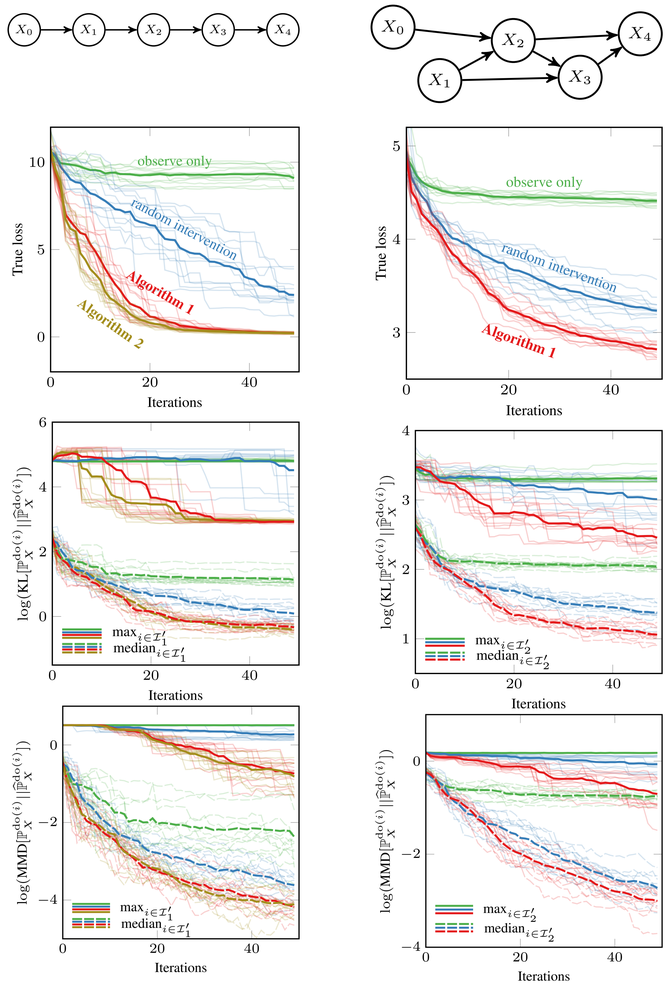
$$\frac{1}{T} \sum_{t=1}^{T} \mathcal{R}(\mathcal{D} \cup \{(i, x_t)\}), \quad x_t \sim \widetilde{\mathbb{P}}_X^{\mathrm{do}(i)}\}$$

Algorithm 2: Dynamic programming (need 'nice' graph and interventions).

$$U_n(x_{n-1}) = \mathcal{R}_n(\mathcal{D}_n \cup \{(x_{n-1}, x_n)\}) \qquad \text{for any value } x_n$$

$$\mathbb{E}_{x \sim \widetilde{\mathbb{P}}_X^{\mathrm{do}(i)}}\left[\mathcal{R}(\mathcal{D} \cup \{(i,x)\})\right] = \sum_{n=1}^{m} U_n^{\mathrm{curr}} + \mathbb{E}_{x \sim \widetilde{\mathbb{P}}_X^{\mathrm{do}(i)}}\left[\sum_{n=m+1}^{N} U_n(x_{n-1})\right].$$

Experiments



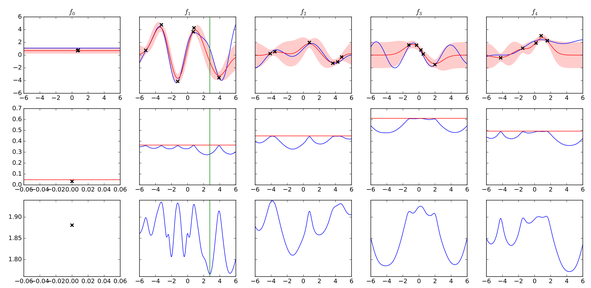


Figure: A visualisation of actively choosing an intervention n SCM with chain structure.