Probabilistic Active Learning of Structural Causal Models

Causality Workshop
Uncertainty in Artificial Intelligence, August 2017

Paul K. Rubenstein, Ilya Tolstikhin, Philipp Hennig, Bernhard Schölkopf

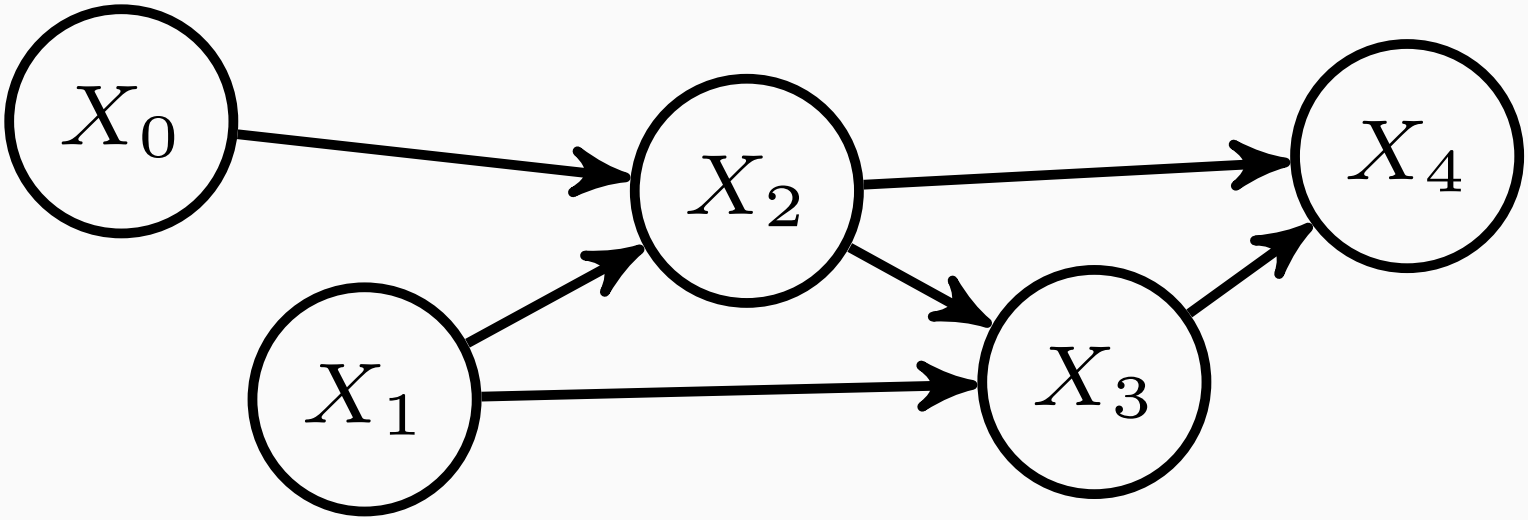
Max Planck Insititute for Intelligent Systems, Tübingen

Motivation

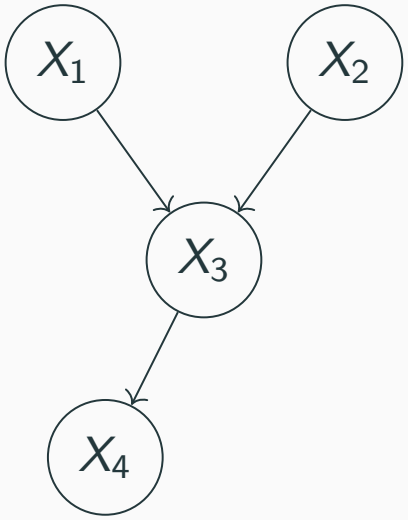
What should we do after causal discovery? Understand **relationships** between variables for control.

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

Motivation



Structural Equation Models



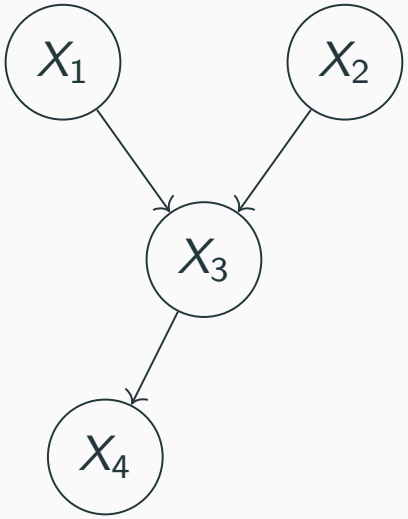
$$X_1 = f_1(E_1)$$
$$X_2 = f_2(E_2)$$
$$X_3 = f_3(X_1, X_2, E_3)$$
$$X_4 = f_4(X_3, E_4)$$
$$\mathbf{E} \sim \mathbb{P}_E$$

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 consider *perfect interventions*, e. g. $\text{do}(X_3 = 0)$

Desired problem setup



$$X_1 = f_1(E_1)$$
$$X_2 = f_2(E_2)$$
$$X_3 = f_3(X_1, X_2, E_3)$$
$$X_4 = f_4(X_3, E_4)$$
$$\mathbf{E} \sim \mathbb{P}_E$$

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. (= estimate f_i and $\mathbb{P}_E$)

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

How should we define distance between $\mathcal{M}$ **and** $\widehat{\mathcal{M}}$**?**

Open question

How should we define distance between $\mathcal{M}$ and $\widehat{\mathcal{M}}$?

$$L^i_{\mathrm{KL}}(\widehat{\mathcal{M}}||\mathcal{M}) = \mathrm{KL}\left[\mathbb{P}_X^{\mathsf{do}(i)}||\hat{\mathbb{P}}_X^{\mathsf{do}(i)}\right]$$

$$L^{\mathcal{I}}_{\mathrm{KL}}(\widehat{\mathcal{M}}||\mathcal{M}) = \sup_{i\in\mathcal{I}} L^i_{\mathrm{KL}}(\widehat{\mathcal{M}}||\mathcal{M})$$

$$L^i_{\mathrm{MMD}_l}(\widehat{\mathcal{M}}||\mathcal{M}) = \mathrm{MMD}_l\left[\mathbb{P}_X^{\mathsf{do}(i)}||\hat{\mathbb{P}}_X^{\mathsf{do}(i)}\right]$$

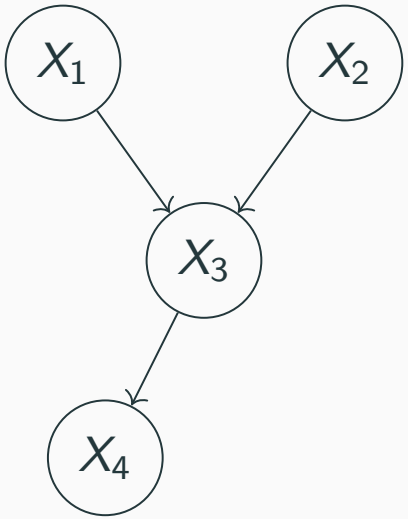
$$L^{\mathcal{I}}_{\mathrm{MMD}_l}(\widehat{\mathcal{M}}||\mathcal{M}) = \sup_{i\in\mathcal{I}} L^i_{\mathrm{MMD}_l}(\widehat{\mathcal{M}}||\mathcal{M})$$

Goal: make statements like

$$L_(\widehat{\mathcal{M}}||\mathcal{M}) < \epsilon \implies ???$$

Our problem setup

Assume **causal graph known** and **additive Gaussian noise with known diagonal covariance**



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

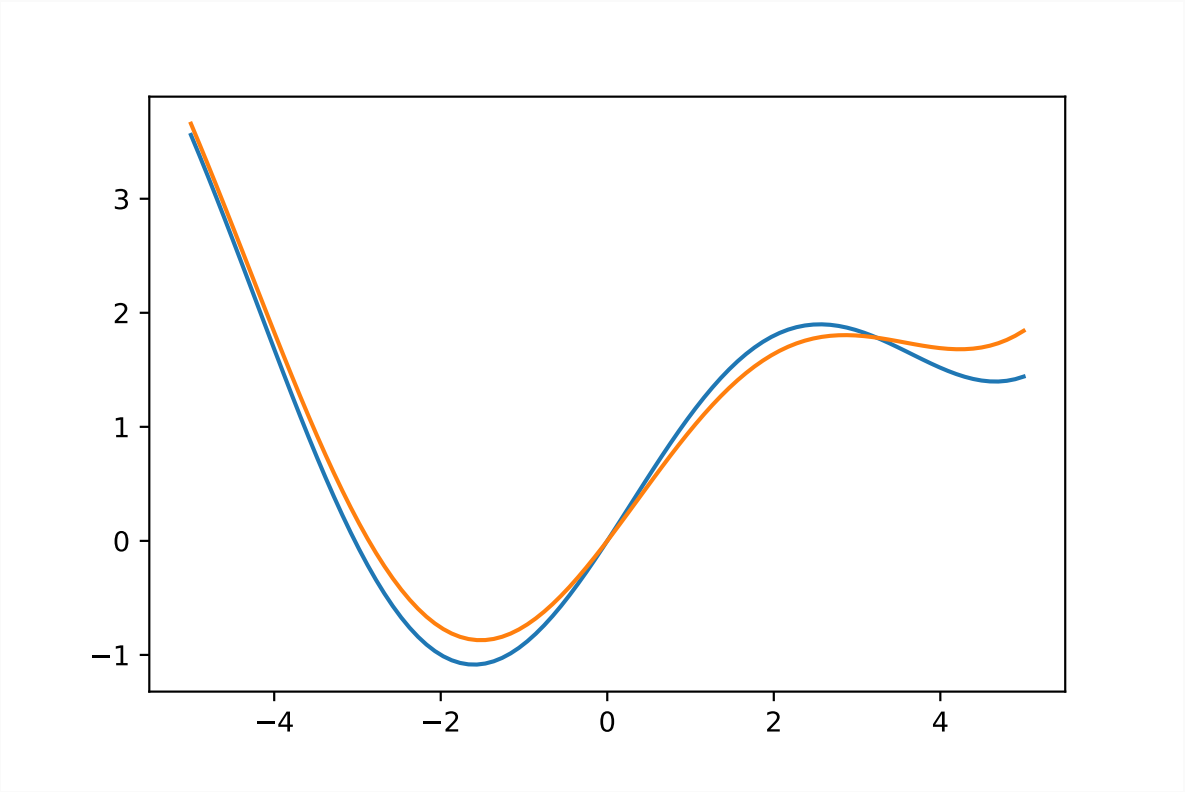
Only unknown quantities are functions $f_n : \mathcal{X}_{pa(n)} \to \mathcal{X}_n$

Assume we are given measures Π_n on $\mathcal{X}_{pa(n)}$

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

Our problem setup

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

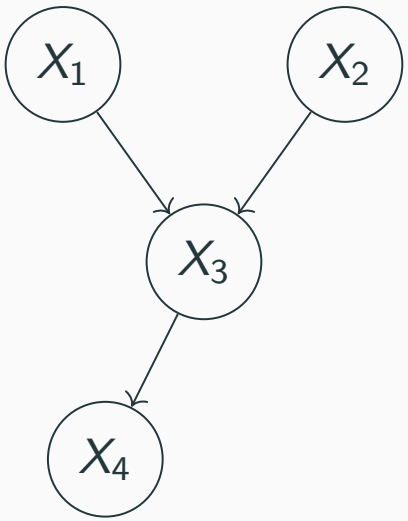


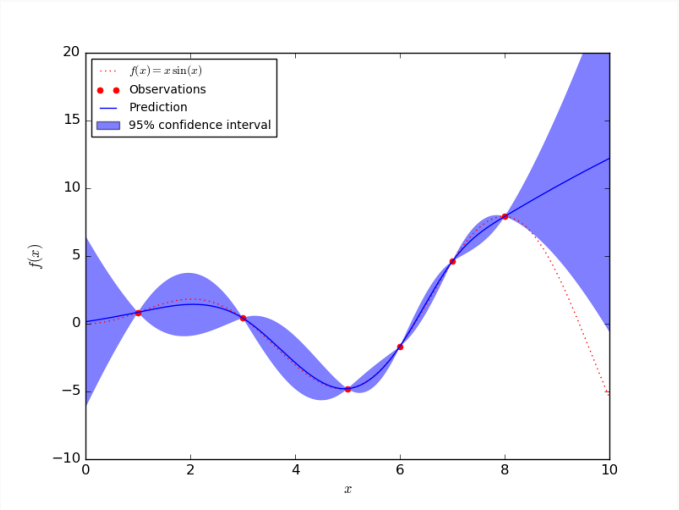
Problem 1: Estimating $\mathcal{M}$

Estimating $\mathcal{M}$ reduces to estimating each f_n.

Assume $f_n \sim \mathcal{GP}(0, k_n)$

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





Posterior:

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

Problem 1: Estimating $\mathcal{M}$

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 $(\mu_{f_n|\mathcal{D}_n})_{n=1,\ldots,N}$. Then

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

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

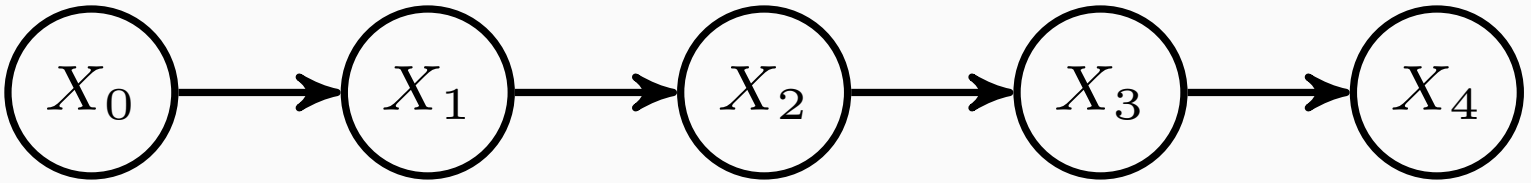
Problem 2: Active Learning

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

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

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

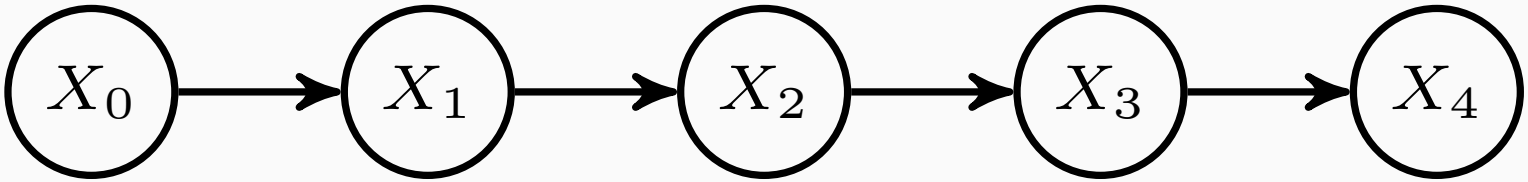
Problem 2: Active Learning



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

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

Problem 2: Active Learning



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

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

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