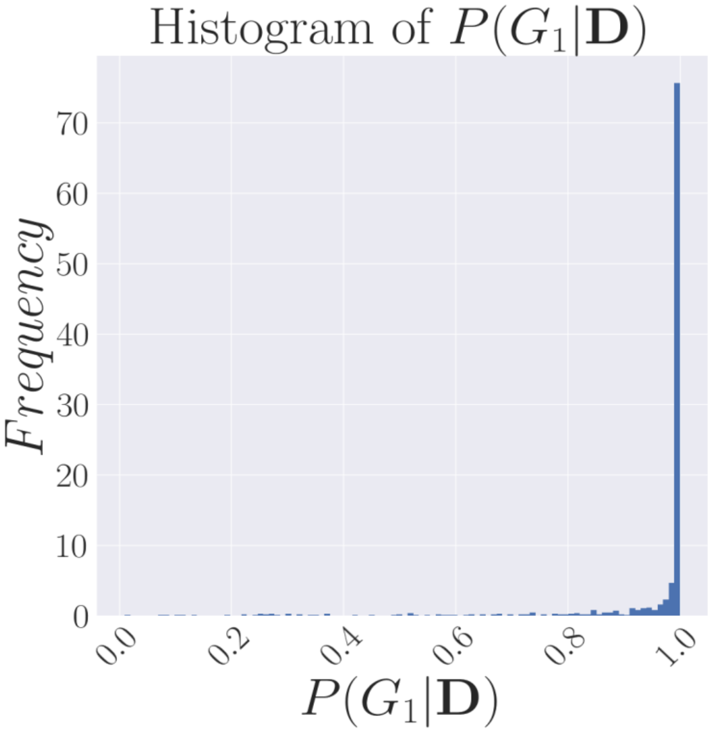
Active Bayesian Causal Discovery for Gaussian Process Networks



Presentation of Master's Thesis by Stefan Kienle

Munich, 26. October 2022

Main reference:

von Kügelgen, Julius, et al. "Optimal experimental design via Bayesian optimization: active causal structure learning for Gaussian process networks." arXiv preprint arXiv:1910.03962 (2019).

Outline

1. Active Bayesian approach for Causal Discovery
2. Bayesian Optimization / Experimental Design
3. Numerical Results for the Bivariate Case
4. Generalization to Four Variables
   - ➢ Computational Challenge
   - ➢ Proposal to overcome computational challenge
5. Conclusion

Active Bayesian approach for Causal Discovery

We consider a setting where we have:

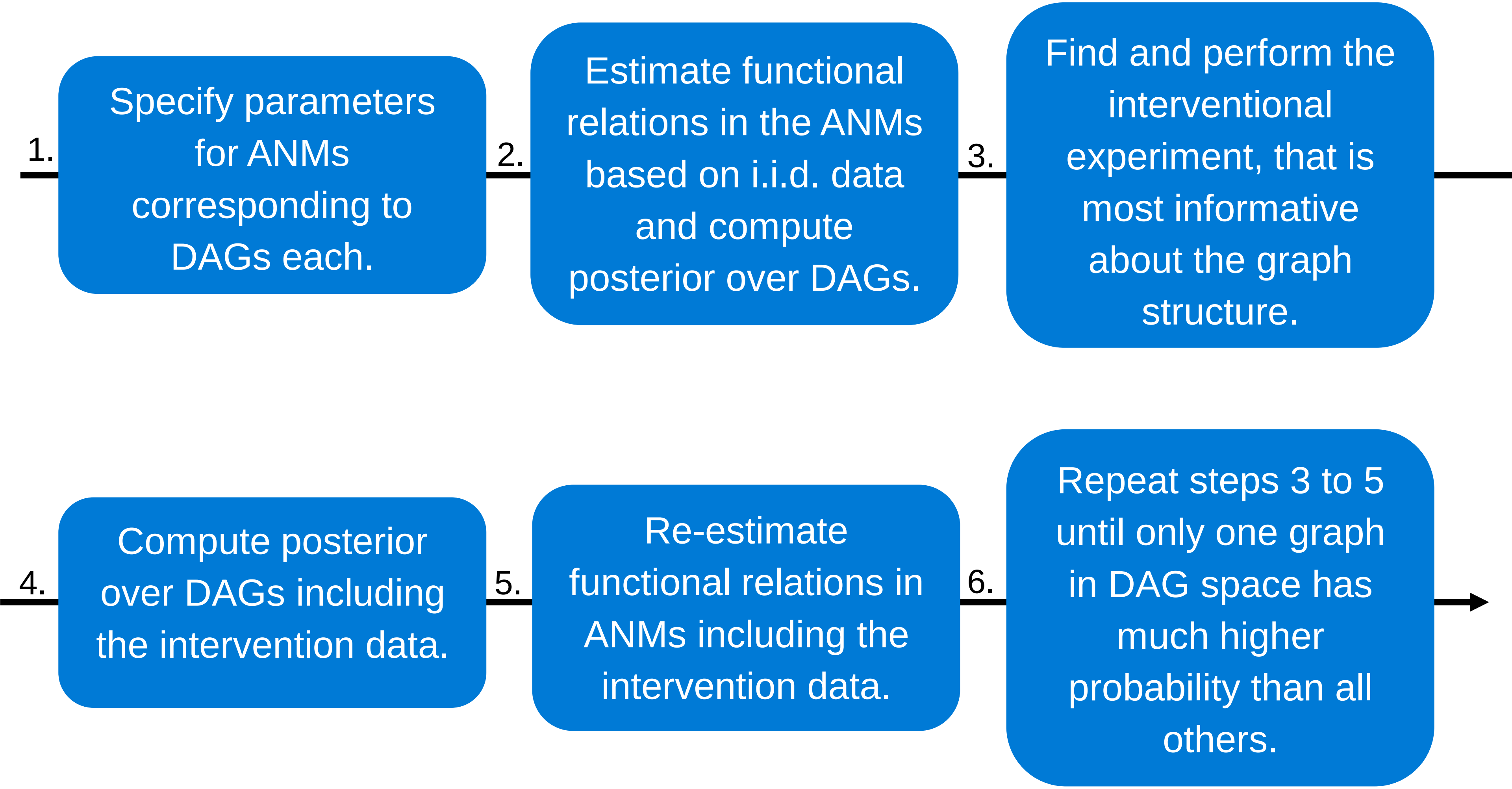
- a **low number of initial (i.i.d.) observations** of system variables and
- the **possibility to perform (perfect) interventions on every variable** in the system under consideration.

Goal: Discover the systems causal structure (DAG) using as few interventions as possible.

Modelling assumptions:

- System of random variables is modeled by an **additive noise model (ANM)** with Gaussian noise term.
- We assume the **intervention data is collected sequentially**.

Schematic overview of the procedure



1. Specify parameters for ANMs corresponding to DAGs each.

2. Estimate functional relations in the ANMs based on i.i.d. data and compute posterior over DAGs.

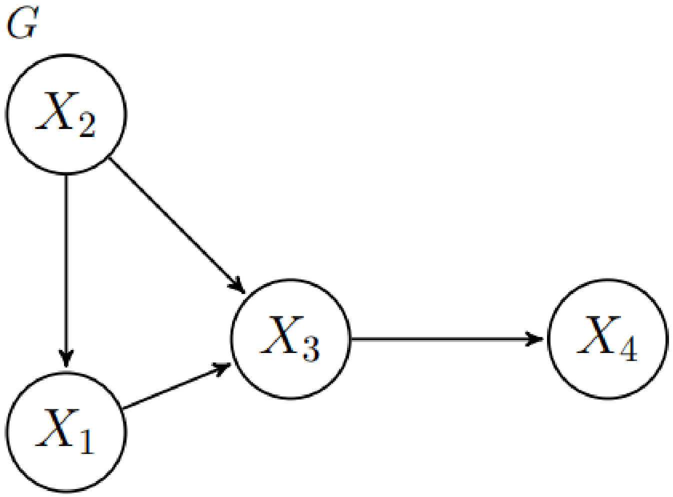
3. Find and perform the interventional experiment, that is most informative about the graph structure.

4. Compute posterior over DAGs including the intervention data.

5. Re-estimate functional relations in ANMs including the intervention data.

6. Repeat steps 3 to 5 until only one graph in DAG space has much higher probability than all others.

Active Bayesian approach for Causal Discovery



ANM

$$\begin{aligned} X_1 &:= f^{(1)}(X_2) + \epsilon_1 \\ X_2 &:= \epsilon_2 \\ X_3 &:= f^{(3)}(X_1, X_2) + \epsilon_3 \\ X_4 &:= f^{(4)}(X_3) + \epsilon_4 \end{aligned}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$p(x_1, x_2, x_3, x_4|G) = \underbrace{p(x_2|G)}_{\mathcal{N}(0,\sigma_2^2)} \ \underbrace{p(x_1|x_2, G)}_{\mathcal{N}(0,k_{X_2X_2}+\sigma_1^2 I_N)} \ \underbrace{p(x_3|x_1, x_2, G)}_{\mathcal{N}(0,k_{(X_1,X_2)(X_1,X_2)}+\sigma_3^2 I_N)} \ \underbrace{p(x_4|x_3, G)}_{\mathcal{N}(0,k_{X_3X_3}+\sigma_4^2 I_N)}$$

$$f^{(1)}_{\tilde{X}_2}(x_2^2)|\tilde{x}_1, \tilde{x}_2, x_2^2 \sim \mathcal{N}(\underbrace{k_{x_2^2\tilde{X}_2}(k_{\tilde{X}_2\tilde{X}_2} + \sigma_1^2 I_{\tilde{N}})^{-1}\tilde{x}_1}_{:=\tilde{\mu}_{G^{(1)}}(x_2^2)}, \underbrace{k_{x_2^2x_2^2} - k_{x_2^2\tilde{X}_2}(k_{\tilde{X}_2\tilde{X}_2} + \sigma_1^2 I_{\tilde{N}})^{-1}k_{\tilde{X}_1x_2^2}}_{:=\tilde{\sigma}^2_{G^{(1)}}(x_2^2)})$$

$$p(x^2, \tilde{x}|G) = \underbrace{p(x_1^2|\tilde{x}_1, \tilde{x}_2, x_2^2, G)}_{\mathcal{N}\left(\tilde{\mu}_{G^{(1)}}(x_2^2),\tilde{\sigma}^2_{G^{(1)}}(x_2^2)+\sigma_1^2\right)} \ \underbrace{p(x_3^2|\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, x_2^2, x_1^2, G)}_{\mathcal{N}\left(\tilde{\mu}_{G^{(3)}}(x_1^2,x_2^2),\tilde{\sigma}^2_{G^{(3)}}(x_1^2,x_2^2)+\sigma_3^2\right)} \ \underbrace{p(x_4^2|\tilde{x}_3, \tilde{x}_4, x_3^2, G)}_{\mathcal{N}\left(\tilde{\mu}_{G^{(4)}}(x_3^2),\tilde{\sigma}^2_{G^{(4)}}(x_3^2)+\sigma_4^2\right)} \ p(\tilde{x}|G)$$

Bayesian Experimental Design

- We want to choose the intervention experiment that contains (on average) most information about the true distribution over the DAG space.

$$\begin{aligned} g(\mathcal{E}, p(\theta)) &= \mathbb{E}_{X_{\mathcal{E}}}\left[\int p(\theta|x)\log(p(\theta|x))\,d\theta - \int p(\theta)\log(p(\theta))\,d\theta\right] \\ &= D_{KL}\left(p_{\mathcal{E}}(\theta, x)\,||\,p(\theta)p_{\mathcal{E}}(x)\right) \end{aligned}$$

- In the considered approach we end up with the following optimization problem:

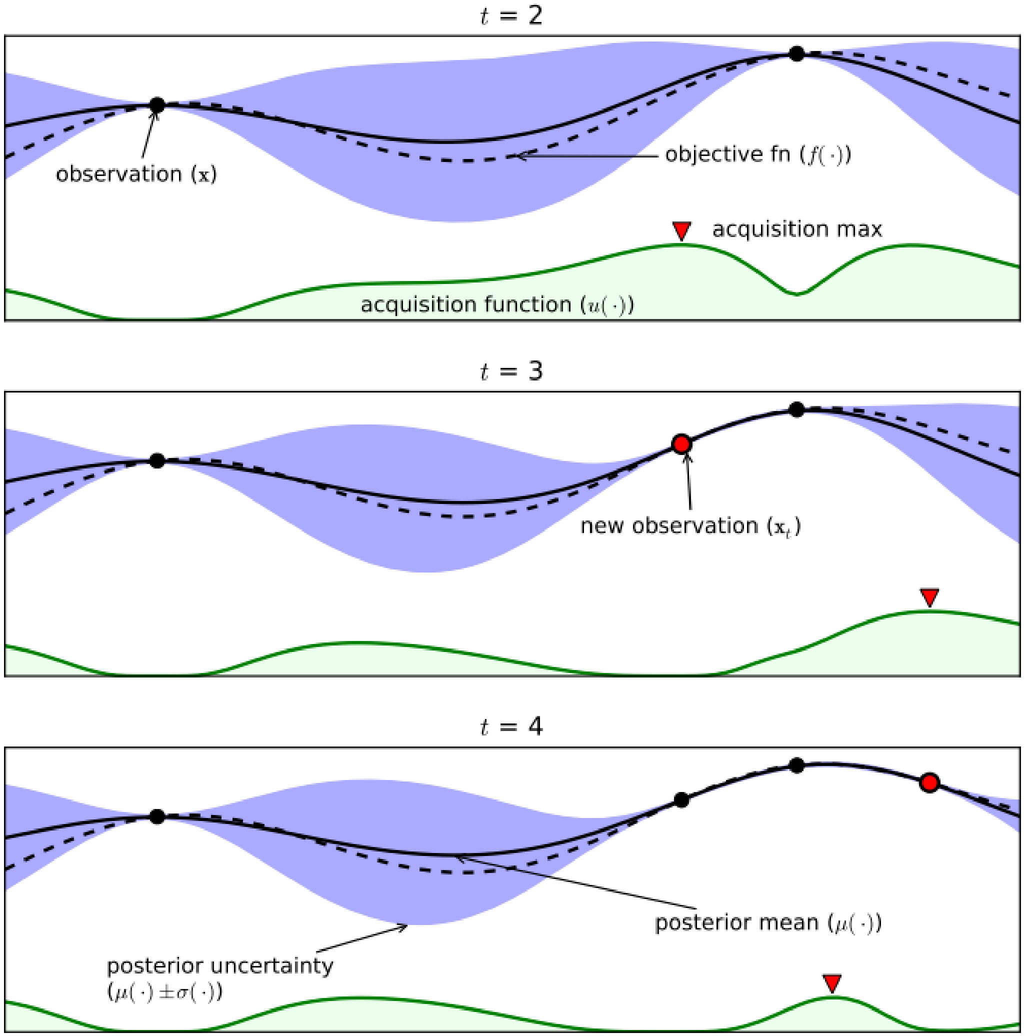
$$\begin{aligned} (j^*, x^*) &= \operatorname*{argmax}_{j\in\{1,\dots,d\},\; x\in\mathcal{X}_j} \mathbb{E}_G\left[\mathbb{E}_{\mathbf{x}_{-j}|G,do(X_j=x)}\left[\log\left(p_{G|\mathbf{x}_{-j},do(X_j=x)}(G|\mathbf{x}_{-j}, do(X_j=x))\right)\right]\right] \\ &\approx \operatorname*{argmax}_{j\in\{1,\dots,d\},\; x\in\mathcal{X}_j} \sum_{\tilde{G}\in\mathcal{G}} p_G(\tilde{G})\frac{1}{M}\sum_{m=1}^{M}\log(p_{G|\mathbf{x}_{-j},do(X_j=x)}(\tilde{G}|\mathbf{x}_{-j}^{(m)}, do(X_j=x))) \end{aligned}$$

Bayesian Optimization

- Bayesian Optimization is a derivative free **optimization procedure** for possibly noisy black box functions.

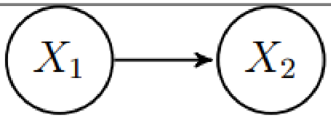
- The acquisition function is based on the **information contained in the GP fit** of the data available up to a certain iteration.

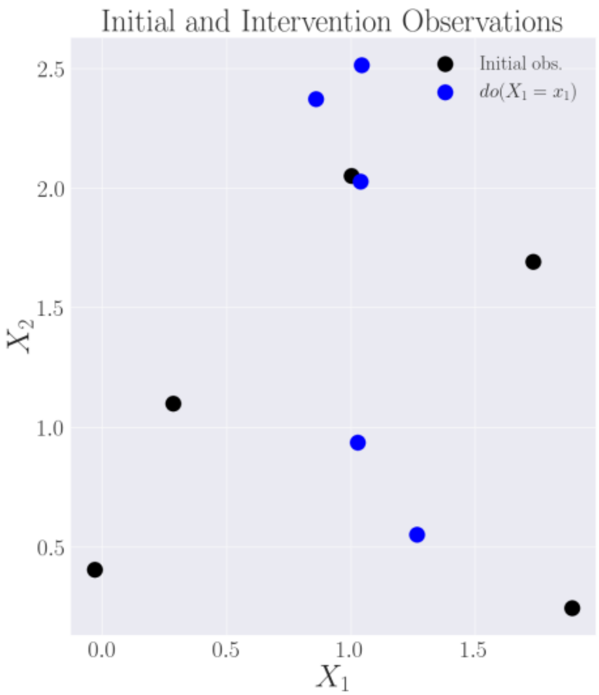
- If the underlying objective function is sufficiently smooth, we have convergence results with high probability.

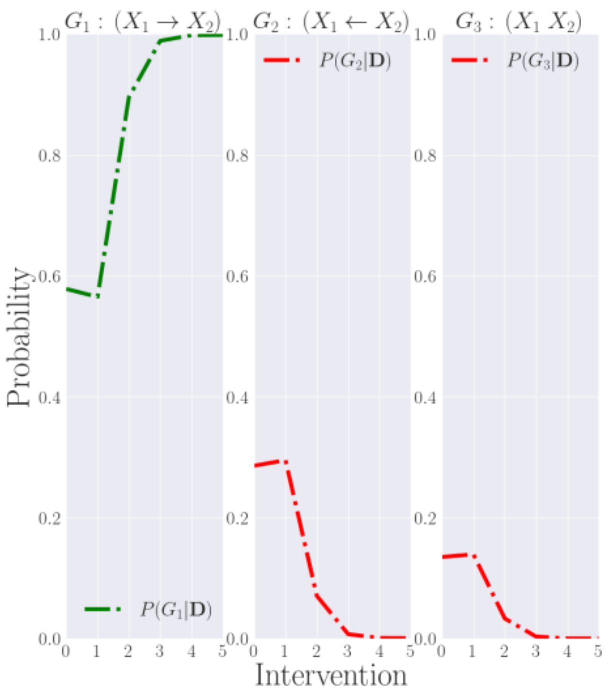


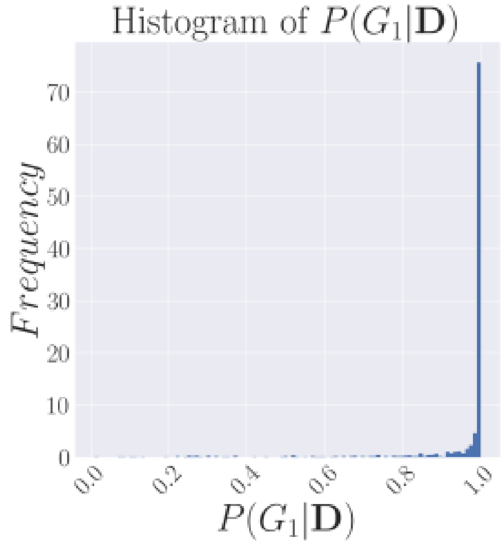
https://arxiv.org/pdf/1012.2599.pdf (Figure 1)

Numerical Results for the Bivariate Case

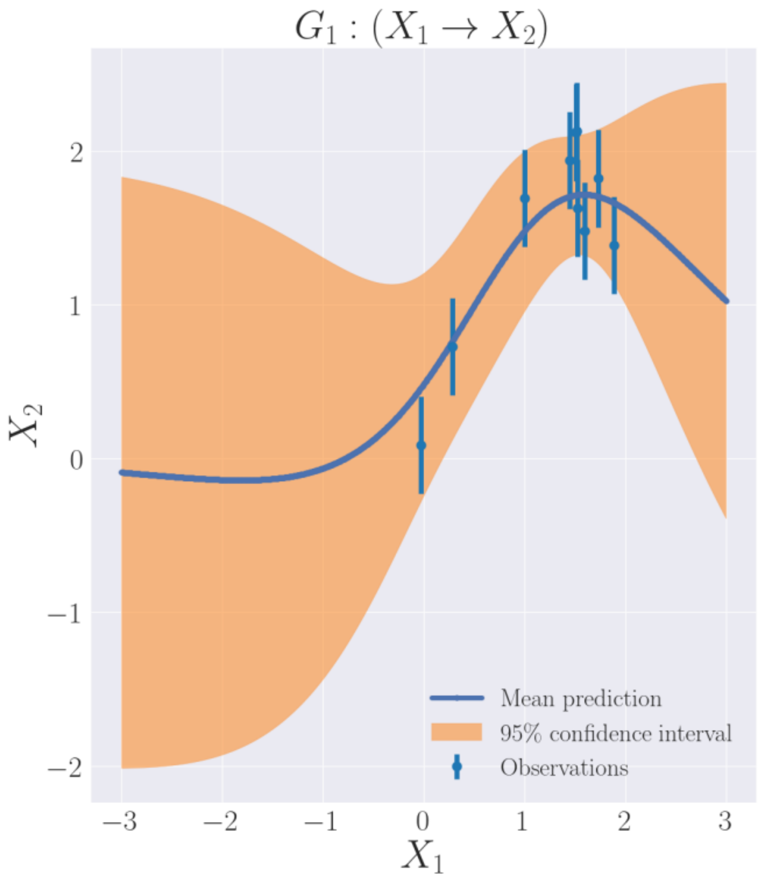
	Graph	ANM
G_{true}		$X_1 := \epsilon_1,$ $X_2 := 2\tanh(X_1) + \epsilon_2,$







Numerical Results for the Bivariate Case



- The algorithm chooses intervention values **mostly on one and the same variable in a small domain region**, where the regression function has **nonlinear curvature**.
- The algorithm chooses intervention values at **locations where we already have at least one observation**. But it does not necessarily choose the location where we have most observations.
- Whenever there is a valid backward model for a parent child relation of two random variables, it can be most beneficial to intervene on both.

Generalization to four variables

- The **computation of the likelihood scales well if we increase the system size**.
- But the optimization causes not acceptable increase in computational burden. Recall that we want to solve

$$(j^*, x^*) = \underset{j\in\{1,\ldots,d\},\; x\in\mathcal{X}_j}{\operatorname{argmax}} \mathbb{E}_G \left[\mathbb{E}_{\mathbf{x}_{-j}|G,do(X_j=x)} \left[\log\left(p_{G|\mathbf{x}_{-j},do(X_j=x)}(G|\mathbf{x}_{-j}, do(X_j=x))\right)\right]\right].$$

- And we approximate the inner expectation via

$$\frac{1}{M}\sum_{m=1}^{M} \log(p_{G|\mathbf{x}_{-j},do(X_j=x)}(\tilde{G}|D_j^{(m)},\tilde{D})) = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{p(D_j^{(m)},\tilde{D}|\tilde{G})p(\tilde{G})}{\sum_{\hat{G}\in\mathcal{G}} p(D_j^{(m)},\tilde{D}|\hat{G})p(\hat{G})}\right).$$

Proposal to overcome computational difficulties

- **Idea**: Mimic the behavior of the algorithm observed in the bivariate case based on the information nested in the GP fits.

  - To **confirm good functional estimates in areas where we already have information**, we can minimize the variance of a GP prediction at some point in the input domain

$$\tilde{\sigma}^2_{G^{(j)}}(x) = \underbrace{k(x,x)}_{=\text{constant}} - k_{x\tilde{X}_{pa(j)}}(k_{\tilde{X}_{pa(j)}\tilde{X}_{pa(j)}} + \sigma_j^2 I)^{-1} k_{\tilde{X}_{pa(j)}x}, \quad x \in \mathcal{X}_{pa(j)}$$

  - To **include the curvature information**, we aim to maximize the absolute value of the second derivative of the inferred regression function. In total we get the following objective

$$f^{(j)}_{\text{obj.}}(x) := \sigma^2_{G^{(j)}}(x) - \left|\frac{\partial^2}{\partial x^2}\tilde{\mu}_{G^{(j)}}(x)\right|$$

Conclusion

- The strength of this approach is that we have **closed forms for the likelihoods without imposing too restrictive assumptions**.

- **Interventional data has great potential** for causal inference.

- In the **four-variable case**, the super exponentially growing number of DAGs already made the optimization **too time consuming**.

- Based on the observations on the behavior in the bivariate case and the theory on ANMs it may be possible to identify a procedure that is approximately equivalent to optimizing the information gain that is computationally much more tractable.

Thank you for your attention!

Questions?