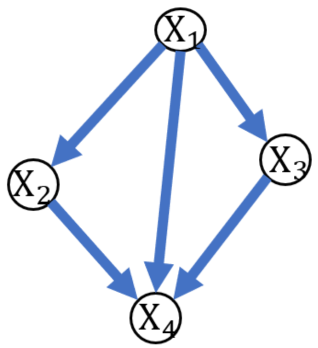
Sample Efficient Bayesian Learning of Causal Graphs from Interventions

Zihan Zhou*, Muhammad Qasim Elahi*, Murat Kocaoglu

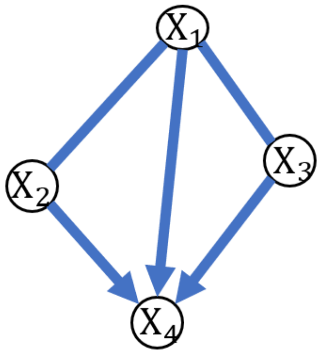
School of Electrical and Computer Engineering, Purdue University

Motivation

- **Causal Discovery**
  - Given observational data, we can only learn $\epsilon(D)$



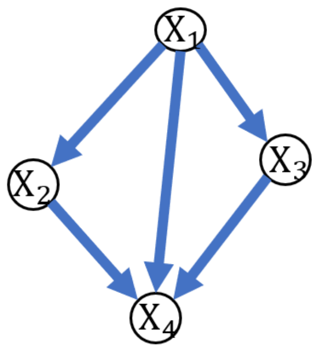
Ground Truth Graph D



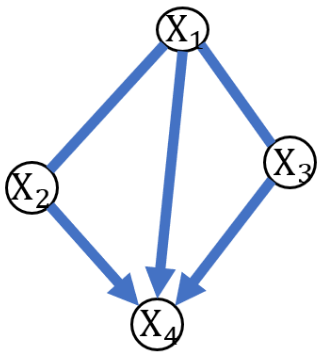
MEC of Causal Graph $\epsilon(D)$

Motivation

- **Causal Discovery**
  - Given observational data, we can only learn $\epsilon(D)$
  - Require further assumptions or interventional data to learn D



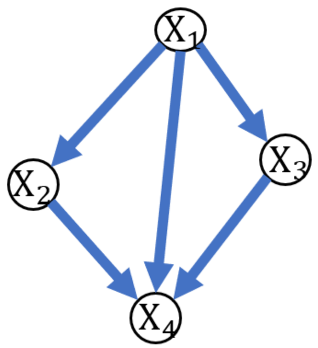
Ground Truth Graph D



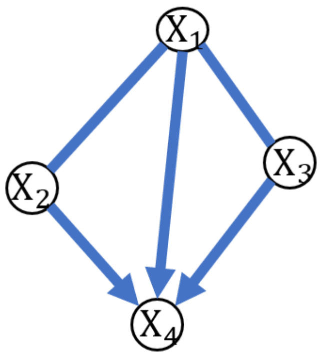
MEC of Causal Graph $\epsilon(D)$

Motivation

- **Causal Discovery**
  - Given observational data, we can only learn $\epsilon(D)$
  - Require further assumptions or interventional data to learn D
- **Challenges**
  - Experimental design methods usually assume access to infinite intervention data



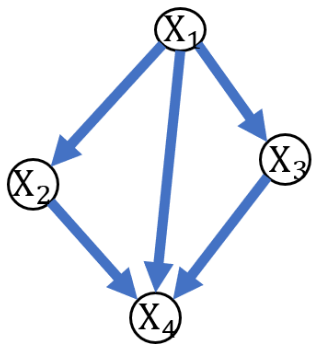
Ground Truth Graph D



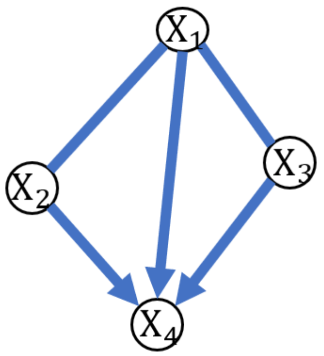
MEC of Causal Graph $\epsilon(D)$

Motivation

- **Causal Discovery**
  - Given observational data, we can only learn $\epsilon(D)$
  - Require further assumptions or interventional data to learn D
- **Challenges**
  - Experimental design methods usually assume access to infinite intervention data
  - Bayesian causal discovery methods have parametric assumptions on the SCM and noise



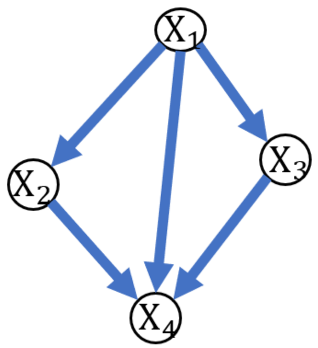
Ground Truth Graph D



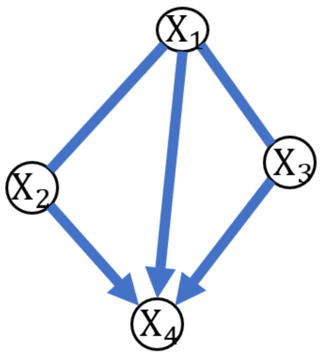
MEC of Causal Graph $\epsilon(D)$

Motivation

- **Causal Discovery**
  - Given observational data, we can only learn $\epsilon(D)$
  - Require further assumptions or interventional data to learn D
- **Challenges**
  - Experimental design methods usually assume access to infinite intervention data
  - Bayesian causal discovery methods have parametric assumptions on the SCM and noise
  - The only non-parametric Bayesian approach assume that the causal graph is a tree



Ground Truth Graph D



MEC of Causal Graph $\epsilon(D)$

Problem Formulation

- Assumptions:
  - Causal faithfulness

Problem Formulation

- Assumptions:
  - Causal faithfulness
  - Positive distributions

Problem Formulation

- Assumptions:
  - Causal faithfulness
  - Positive distributions
  - Hard interventions

Problem Formulation

- Assumptions:
  - Causal faithfulness
  - Positive distributions
  - Hard interventions
  - Causal sufficiency

Problem Formulation

- Assumptions:
  - Causal faithfulness
  - Positive distributions
  - Hard interventions
  - Causal sufficiency
  - Access to observational distribution

Problem Formulation

- Assumptions:
  - Causal faithfulness
  - Positive distributions
  - Hard interventions
  - Causal sufficiency
  - Access to observational distribution
- Task:
  - Given the UCCG G and N interventional samples, return a DAG D

Our Approach

- Separating System:

Lemma 5 ([Shanmugam et al., 2015]). *There exists a labeling procedure that gives distinct labels of length* l *for all elements in* $[n]$ *using letter from the integer alphabet* $\{0, 1, ..., a\}$, *where* $l = \lceil \log_a n \rceil$. *Furthermore, in every position, any integer letter is used at most* $\lceil \frac{n}{a} \rceil$ *times.*

Our Approach

- Separating System:

Lemma 5 ([Shanmugam et al., 2015]). *There exists a labeling procedure that gives distinct labels of length l for all elements in $[n]$ using letter from the integer alphabet $\{0, 1, ..., a\}$, where $l = \lceil \log_a n \rceil$. Furthermore, in every position, any integer letter is used at most $\lceil \frac{n}{a} \rceil$ times.*

- Cutting edge configuration and interventional distributions:

Lemma 1. *[Elahi et al., 2024] Assume that the faithfulness assumption holds and $\mathcal{D}^*$ is the true DAG. For any DAG $\mathcal{D}_1 \neq \mathcal{D}^*$, if $P_{\mathbf{s}}^{\mathcal{D}_1} = P_{\mathbf{s}}^{\mathcal{D}^*}$ for some $\mathbf{S} \subseteq \mathbf{V}$, they must share the same cutting edge orientation $\mathsf{C}(\mathbf{S})$.*

Our Approach

- Separating System:

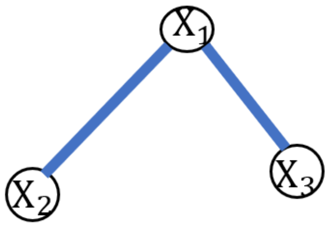
Lemma 5 ([Shanmugam et al., 2015]). *There exists a labeling procedure that gives distinct labels of length l for all elements in $[n]$ using letter from the integer alphabet $\{0, 1, ..., a\}$, where $l = \lceil \log_a n \rceil$. Furthermore, in every position, any integer letter is used at most $\lceil \frac{n}{a} \rceil$ times.*

- Cutting edge configuration and interventional distributions:

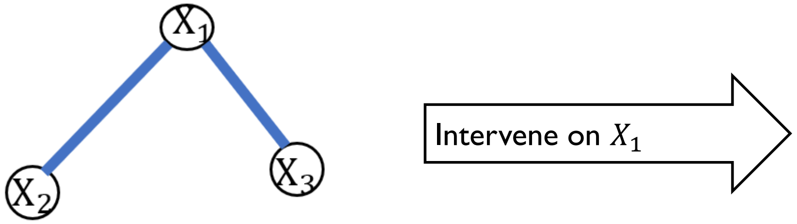
Lemma 1. *[Elahi et al., 2024] Assume that the faithfulness assumption holds and $\mathcal{D}^*$ is the true DAG. For any DAG $\mathcal{D}_1 \neq \mathcal{D}^*$, if $P_{\mathbf{S}}^{\mathcal{D}_1} = P_{\mathbf{S}}^{\mathcal{D}^*}$ for some $\mathbf{S} \subseteq \mathbf{V}$, they must share the same cutting edge orientation $\mathsf{C}(\mathbf{S})$.*

Shanmugam, Karthikeyan, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. "Learning causal graphs with small interventions." Advances in Neural Information Processing Systems 28 (2015).
Elahi, Muhammad Qasim, Lai Wei, Murat Kocaoglu, and Mahsa Ghasemi. "Adaptive Online Experimental Design for Causal Discovery." arXiv preprint arXiv:2405.11548 (2024).

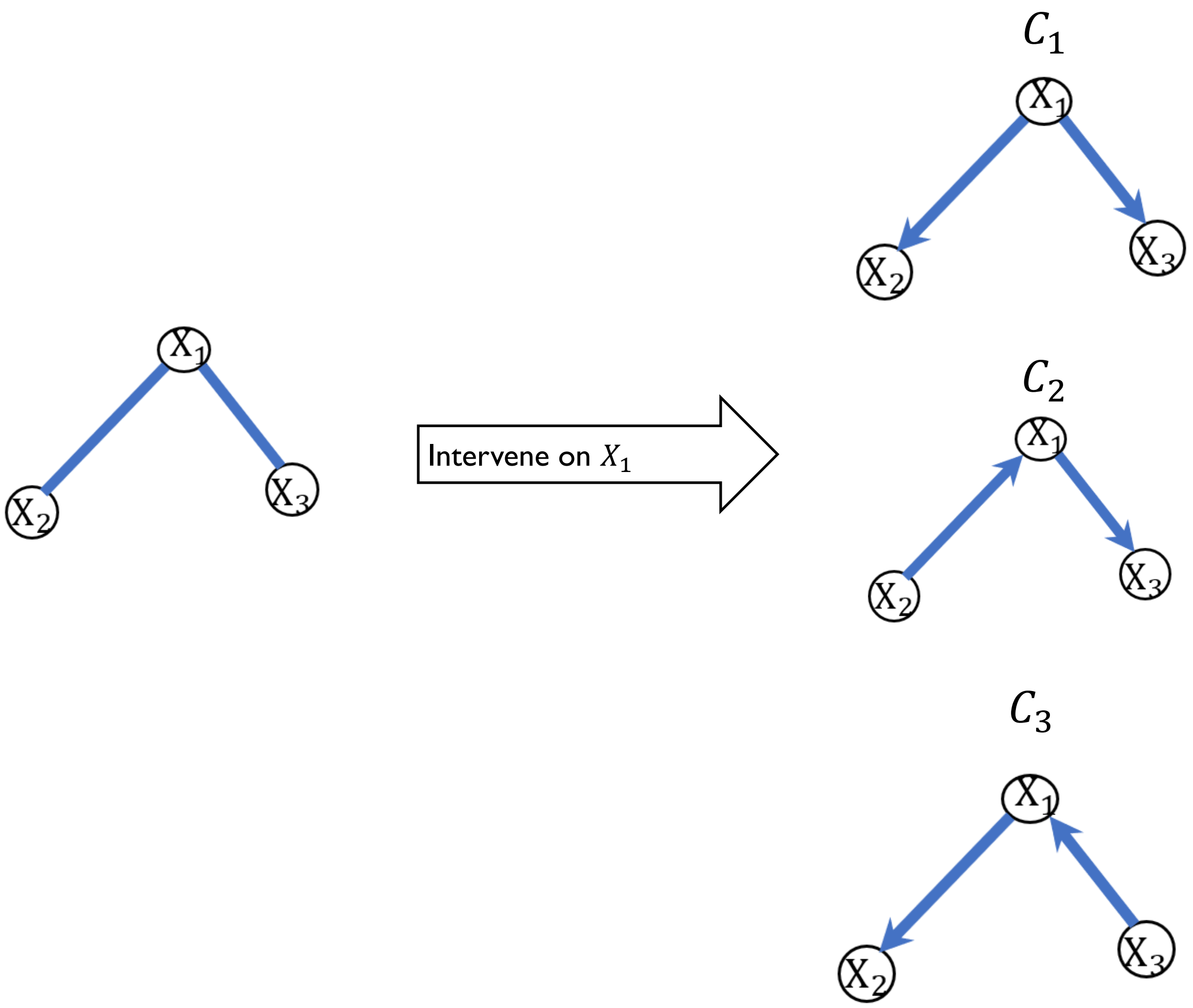
Our Approach



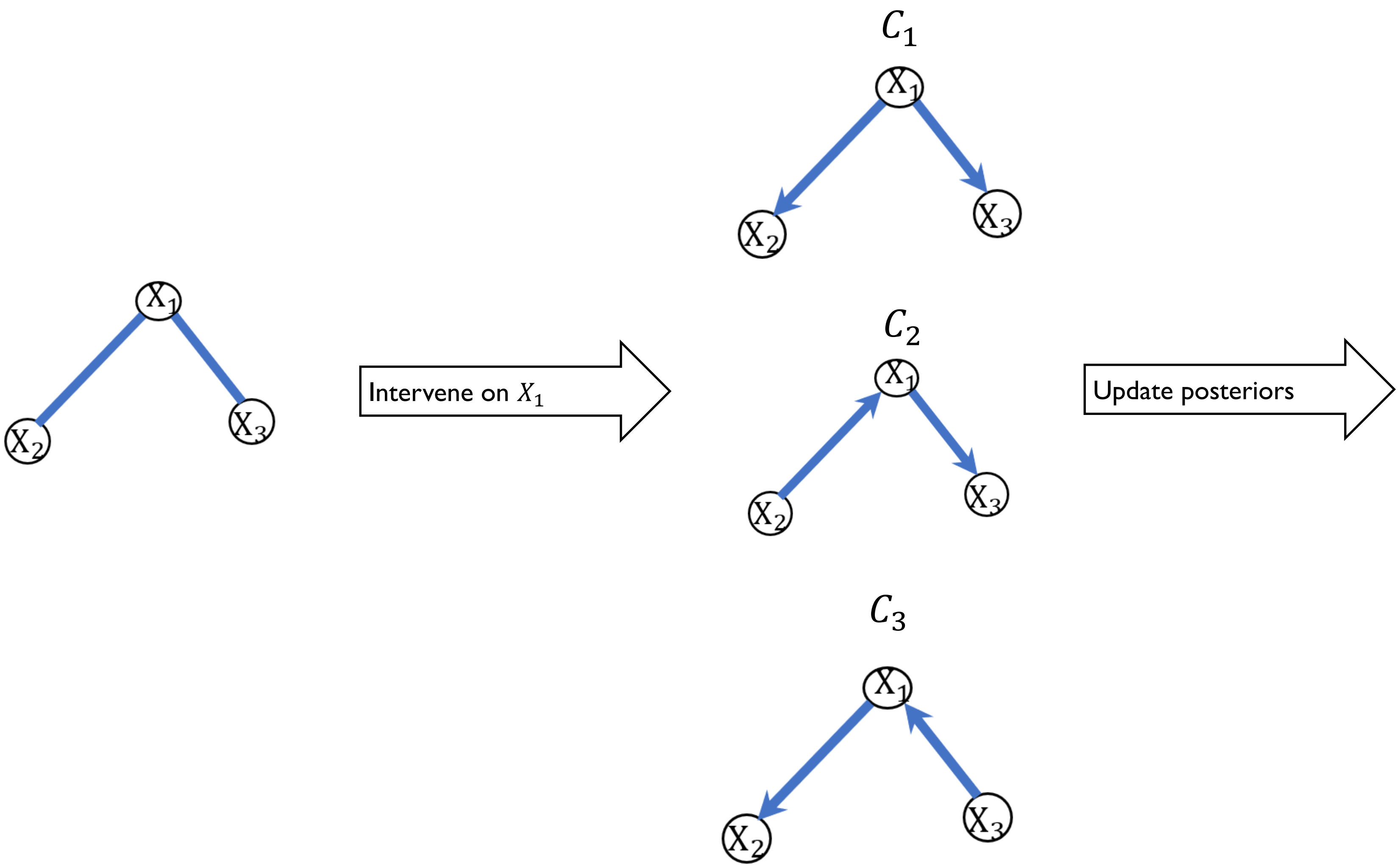
Our Approach



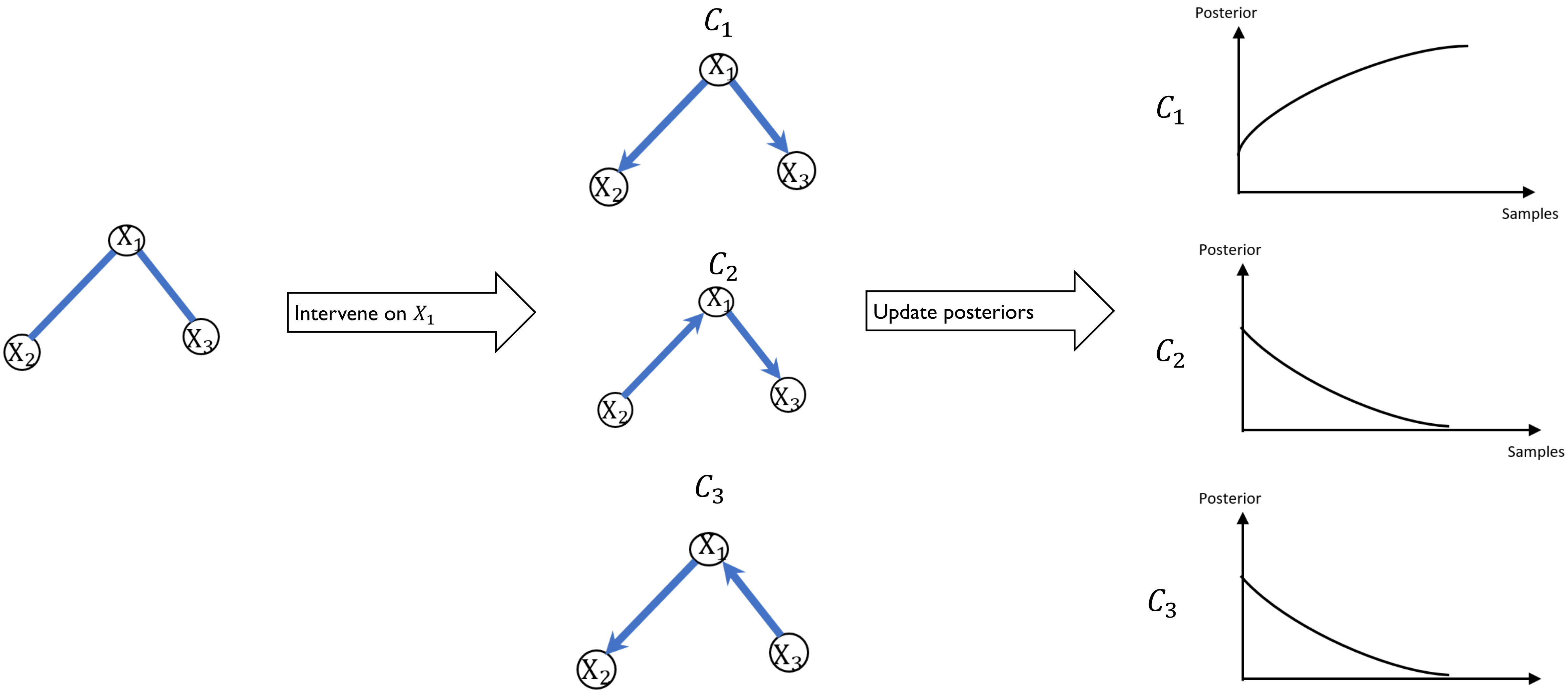
Our Approach



Our Approach



Our Approach



Theoretical Analysis

- Asymptotic results:

Lemma 2. *(Posterior Consistency) Consider an intervention target* $\mathbf{S}_i \in \mathcal{S}$ *and the corresponding true cutting-edge configuration* $\mathsf{C}^*(\mathbf{S}_i)$. *As the number of samples* $m_{\mathbf{s}_i} \to \infty$ *in* $\mathsf{Data}_{do(\mathbf{s}_i)} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{m_{\mathbf{s}_i}}\}$, *the posterior of the true cutting-edge configuration* $P(\mathsf{C}^*(\mathbf{S}_i) \mid \mathsf{Data}_{do(\mathbf{s}_i)})$ *converges to* 1 *with high probability. More precisely, we have the following high probability lower bound on the posterior probability of the true cutting-edge configuration.*

$$P(\mathsf{C}^*(\mathbf{S}_i) \mid \mathsf{Data}_{do(\mathbf{s}_i)}) \geq 1 - \frac{1}{1 + \alpha_1 \exp\left(\mathcal{O}(m_{\mathbf{s}_i}) - \alpha_2 \mathcal{O}\left(\sqrt{m_{\mathbf{s}_i} \ln \frac{1}{\delta}}\right)\right)} \quad w.p. \;\; at\ least\ 1 - \delta$$

Where α_1 *and* α_2 *are constants depending on the prior and the problem instance. Thus, for any small choice of the probability* δ, *with a sufficiently large number of samples* $m_{\mathbf{s}_i}$, *the posterior of the true cutting-edge configuration* $P(\mathsf{C}^*(\mathbf{S}_i) \mid \mathsf{Data}_{do(\mathbf{s}_i)})$, *converges to* 1 *with a probability at least* $1 - \delta$.

Theoretical Analysis

- Non-asymptotic results:

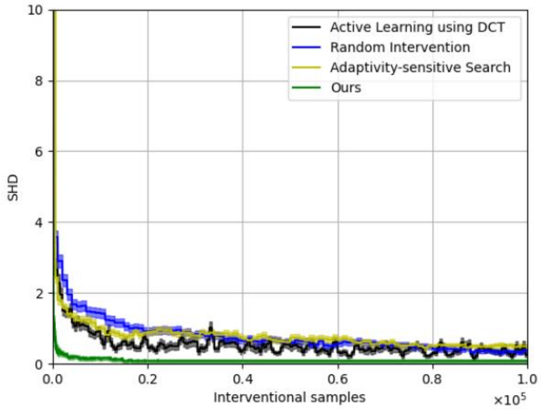
Theorem 3. *Given that the Assumption 1 hold, consider an intervention target* $\mathbf{S}_i \in \mathcal{S}$ *such that* $|\mathbf{S}_i| \leq k$ *and the corresponding true cutting edge configuration* $\mathsf{C}^*(\mathbf{S}_i)$. *If the number of samples* $m_{\mathbf{s}i}$ *in* $\mathsf{Data}_{do(\mathbf{s}_i)} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{m_{\mathbf{s}_i}}\}$ *satisfies the following:*

$$m_{\mathbf{s}_i} = \frac{2\beta^2}{(D^{\mathbf{s}_i})^2} \ln \frac{2^{(k+1)d_m}}{\delta} + \frac{2}{D^{\mathbf{s}_i}} \ln \frac{2^{kd_m}(1-\gamma)(1-p^*)}{p^*\gamma}$$

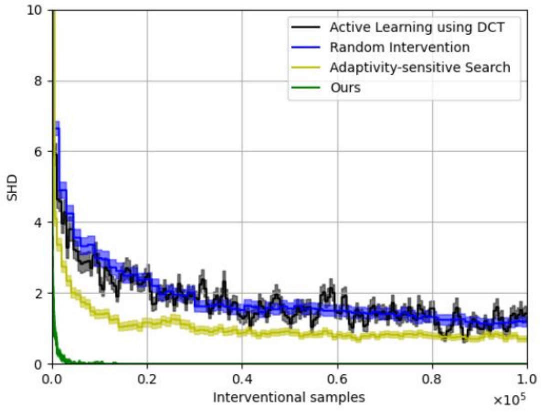
where p^* *is the prior assigned to the true cutting-edge configuration* $\mathsf{C}^*(\mathbf{S}_i)$, *then we have* $P(\mathsf{C}^*(\mathbf{S}_i) \mid \mathsf{Data}_{do(\mathbf{s}_i)}) \geq 1-\gamma$ *with a probability at least* $1-\delta$.

Experimental Results

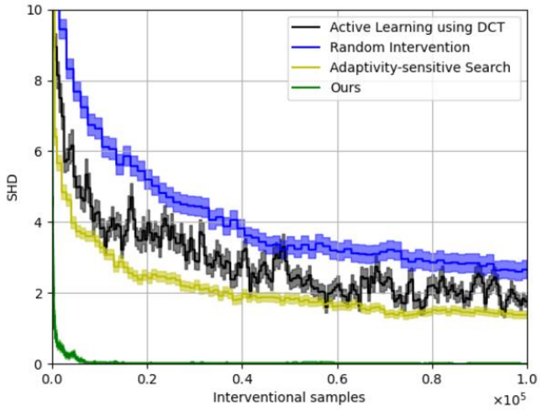
- Baselines:
  - Random
  - DCT
  - Adaptivity-sensitive search



(a) $n = 5, \rho = 1$



(b) $n = 6, \rho = 1$



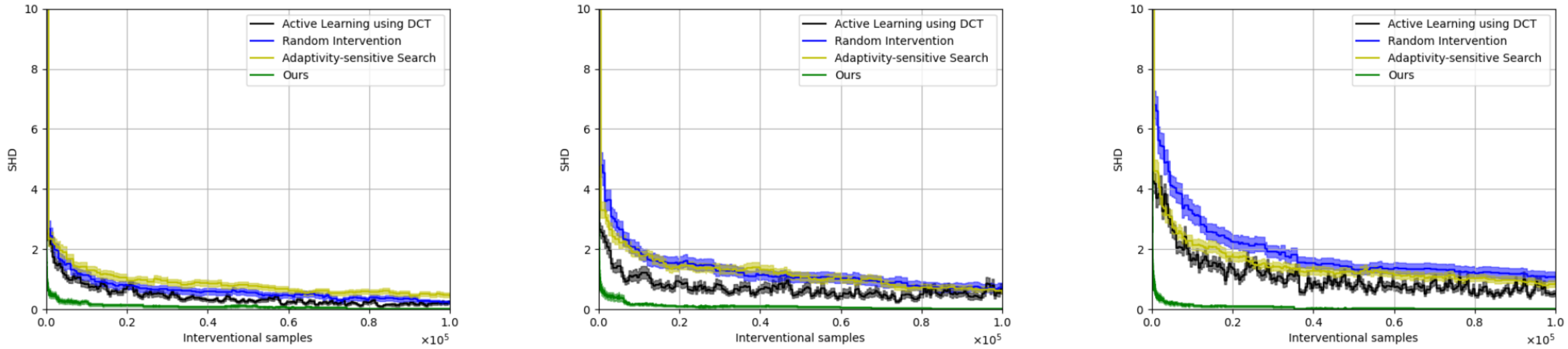
(c) $n = 7, \rho = 1$

Squires, Chandler, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. "Active structure learning of causal DAGs via directed clique trees." Advances in Neural Information Processing Systems 33 (2020): 21500-21511

Choo, Davin, and Kirankumar Shiragur. "Adaptivity complexity for causal graph discovery." In Uncertainty in Artificial Intelligence, pp. 391-402. PMLR, 2023..

Experimental Results

- Baselines:
  - Random
  - DCT
  - Adaptivity-sensitive search



(a) $n = 10, \rho = 0.1$ (b) $n = 10, \rho = 0.15$ (c) $n = 10, \rho = 0.2$

Squires, Chandler, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. "Active structure learning of causal DAGs via directed clique trees." Advances in Neural Information Processing Systems 33 (2020): 21500-21511
Choo, Davin, and Kirankumar Shiragur. "Adaptivity complexity for causal graph discovery." In Uncertainty in Artificial Intelligence, pp. 391-402. PMLR, 2023..

Case Study

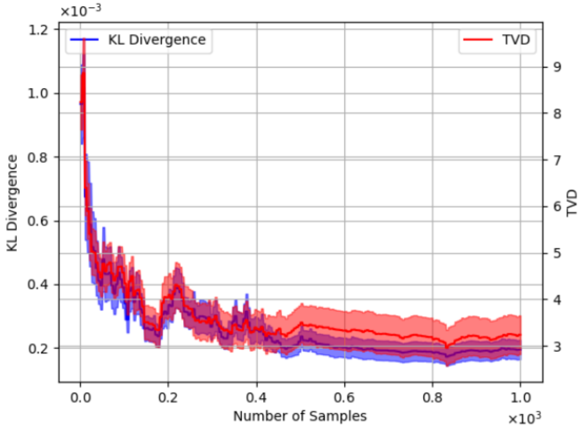
- Modify the main algorithm to solve more general causal queries

Case Study

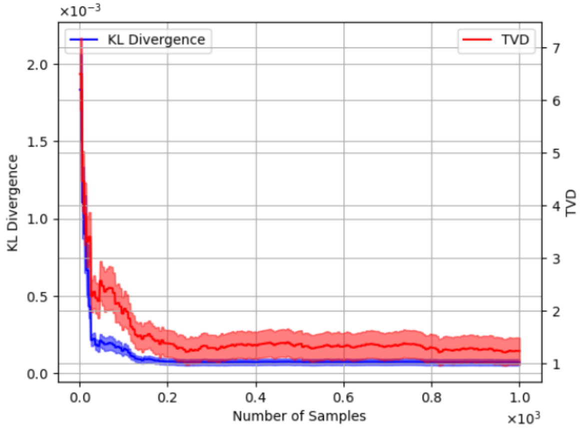
- Modify the main algorithm to solve more general causal queries
- Estimating the causal effect of non-intervenable vertices

Case Study

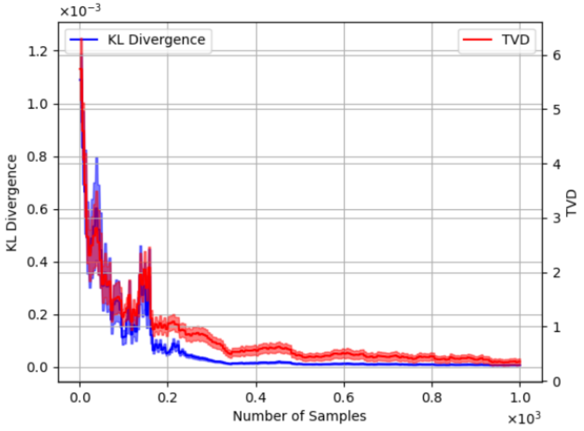
- Modify the main algorithm to solve more general causal queries
- Estimating the causal effect of non-intervenable vertices



(a) $n = 5, \rho = 0.3$



(b) $n = 6, \rho = 0.3$



(c) $n = 7, \rho = 0.3$

Conclusion

- We develop the sample-efficient Bayesian learning causal discovery algorithm without parametric assumptions on the SCM

Conclusion

- We develop the sample-efficient Bayesian learning causal discovery algorithm without parametric assumptions on the SCM
- We show in theory that our proposed algorithm will learn the true causal graph with a high probability given enough interventional samples and the convergence rate

Conclusion

- We develop the sample-efficient Bayesian learning causal discovery algorithm without parametric assumptions on the SCM
- We show in theory that our proposed algorithm will learn the true causal graph with a high probability given enough interventional samples and the convergence rate
- We demonstrate the performance of our algorithm with simulated experiments and show how to modify the algorithm to answer general causal queries with case study