Active Bayesian Causal Inference

A Bayesian Active Learning Framework for Integrated Causal Discovery and Reasoning

Julius von Kügelgen

Max Planck Institute for Intelligent Systems, Tübingen & University of Cambridge

June 29, 2022

Joint work with:

Active Bayesian Causal Inference

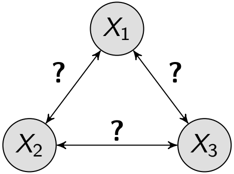
Christian Toth TU Graz
Lars Lorch ETH Zürich
Christian Knoll TU Graz
Andreas Krause ETH Zürich
Franz Pernkopf TU Graz

Robert Peharz* TU Graz
Julius von Kügelgen* MPI for Intelligent Systems, Tübingen, University of Cambridge

Outline

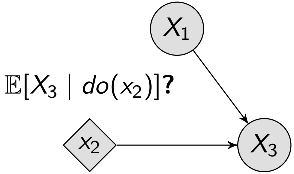
Causal Discovery vs Causal Reasoning

1. Causal Discovery



Infer the causal graph/SCM from data and assumptions.

2. Causal Reasoning



Assuming the causal model is known, (identify &) estimate some query.

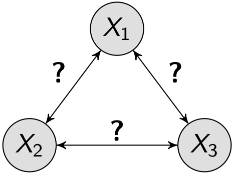
This work: What if we are interested in causal reasoning, but do not have access to a causal model a priori?

2-stage approach uneconomical for *actively-collected interventional data*:

- causal query of interest may not require a fully-specified causal model
- epistemic uncertainty in causal model should be taken into account

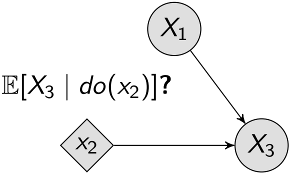
Causal Discovery vs Causal Reasoning

1. Causal Discovery



Infer the causal graph/SCM from data and assumptions.

2. Causal Reasoning



Assuming the causal model is known, (identify &) estimate some query.

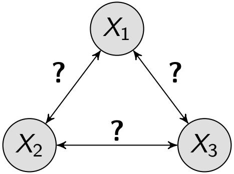
This work: What if we are interested in causal reasoning, but do not have access to a causal model a priori?

2-stage approach uneconomical for *actively-collected interventional data*:

- causal query of interest may not require a fully-specified causal model
- epistemic uncertainty in causal model should be taken into account

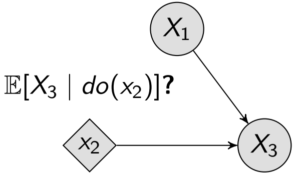
Causal Discovery vs Causal Reasoning

1. Causal Discovery



Infer the causal graph/SCM from data and assumptions.

2. Causal Reasoning



Assuming the causal model is known, (identify &) estimate some query.

This work: What if we are interested in causal reasoning, but do not have access to a causal model a priori?

2-stage approach uneconomical for *actively-collected interventional data*:

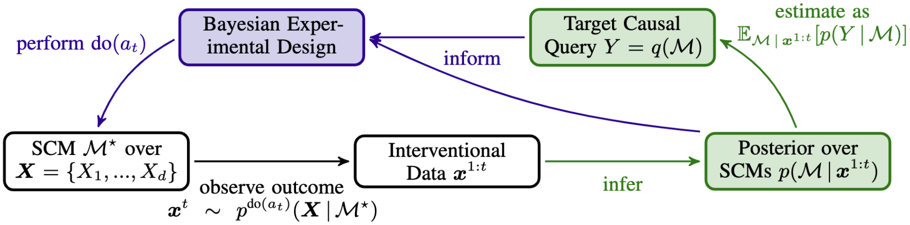
- causal query of interest may not require a fully-specified causal model
- epistemic uncertainty in causal model should be taken into account

Outline

Big Picture

To perform causal reasoning, we:

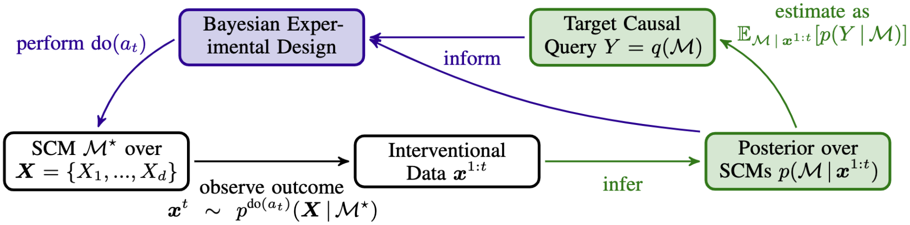
1. Postulate a mathematically well-defined causal model → SCMs.
2. Reduce causal queries to epistemic questions, i.e., what and how much is known about the causal model → Bayesian approach.
3. Collect interventional data to reduce our uncertainty in the causal query of interest → experimental design/active learning.



Big Picture

To perform causal reasoning, we:

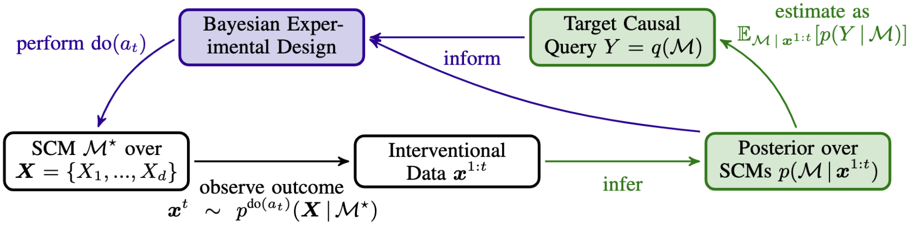
1. Postulate a mathematically well-defined causal model $\rightarrow$ SCMs.
2. Reduce causal queries to epistemic questions, i.e., what and how much is known about the causal model $\rightarrow$ Bayesian approach.
3. Collect interventional data to reduce our uncertainty in the causal query of interest $\rightarrow$ experimental design/active learning.



Big Picture

To perform causal reasoning, we:

1. Postulate a mathematically well-defined causal model $\rightarrow$ SCMs.
2. Reduce causal queries to epistemic questions, i.e., what and how much is known about the causal model $\rightarrow$ Bayesian approach.
3. Collect interventional data to reduce our uncertainty in the causal query of interest $\rightarrow$ experimental design/active learning.



Structural Causal Models (SCMs)

Definition (Pearl 2009)

An SCM $\mathcal{M}$ over endogenous (observed) variables $\boldsymbol{X} = \{X_1, \ldots, X_d\}$ and exogenous (latent) variables $\boldsymbol{U} = \{U_1, \ldots, U_d\}$ consists of:

1. structural equations, or mechanisms,

$$X_i := f_i(\mathbf{Pa}_i, U_i), \qquad \text{for} \qquad i \in \{1, \ldots, d\}, \tag{1}$$

   which assign the value of each X_i as a deterministic function f_i of its direct causes, or causal parents, $\mathbf{Pa}_i \subseteq \boldsymbol{X} \setminus \{X_i\}$ and U_i;
2. a joint distribution $p(\mathbf{U})$ over the exogenous variables.

The corresponding causal graph G is assumed to be acyclic.

$p(\boldsymbol{X} \mid \mathcal{M})$ = pushforward of $p(\boldsymbol{U})$ through the causal mechanisms (1).

Interventions: modify (1), $\mathrm{do}(X_i = \tilde{f}_i(\mathbf{Pa}_i, U_i))$, e.g., $\mathrm{do}(X_2 = 0)$

Structural Causal Models (SCMs)

Definition (Pearl 2009)

An SCM $\mathcal{M}$ over endogenous (observed) variables $\boldsymbol{X} = \{X_1, \ldots, X_d\}$ and exogenous (latent) variables $\boldsymbol{U} = \{U_1, \ldots, U_d\}$ consists of:

1. structural equations, or mechanisms,

$$X_i := f_i(\mathbf{Pa}_i, U_i), \qquad \text{for} \qquad i \in \{1, \ldots, d\}, \tag{1}$$

   which assign the value of each X_i as a deterministic function f_i of its direct causes, or causal parents, $\mathbf{Pa}_i \subseteq \boldsymbol{X} \setminus \{X_i\}$ and U_i;
2. a joint distribution $p(\mathbf{U})$ over the exogenous variables.

The corresponding causal graph G is assumed to be acyclic.

$p(\boldsymbol{X} \mid \mathcal{M})$ = pushforward of $p(\boldsymbol{U})$ through the causal mechanisms (1).

Interventions: modify (1), $\mathrm{do}(X_i = \tilde{f}_i(\mathbf{Pa}_i, U_i))$, e.g., $\mathrm{do}(X_2 = 0)$

Being Bayesian with Respect to Causal Models

Epistemic challenge: true causal model $\mathcal{M}^\star$ is not (completely) known.

Bayesian approach:

1. place a prior $p(\mathcal{M})$ over causal models,
2. collect data $\mathcal{D}$ from the true model $\mathcal{M}^\star$,
3. compute the posterior via Bayes rule:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{\int p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})\, \mathrm{d}\mathcal{M}}.$$

Computationally delicate, as we require a way to

- parametrise the class of models $\mathcal{M}$, and
- perform posterior inference over this model class.

Being Bayesian with Respect to Causal Models

Epistemic challenge: true causal model $\mathcal{M}^\star$ is not (completely) known.

Bayesian approach:

1. place a prior $p(\mathcal{M})$ over causal models,
2. collect data $\mathcal{D}$ from the true model $\mathcal{M}^\star$,
3. compute the posterior via Bayes rule:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{\int p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})\, \mathrm{d}\mathcal{M}}.$$

Computationally delicate, as we require a way to

- parametrise the class of models $\mathcal{M}$, and
- perform posterior inference over this model class.

Being Bayesian with Respect to Causal Models

Epistemic challenge: true causal model $\mathcal{M}^\star$ is not (completely) known.

Bayesian approach:

1. place a prior $p(\mathcal{M})$ over causal models,
2. collect data $\mathcal{D}$ from the true model $\mathcal{M}^\star$,
3. compute the posterior via Bayes rule:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{\int p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})\, \mathrm{d}\mathcal{M}}.$$

Computationally delicate, as we require a way to

- parametrise the class of models $\mathcal{M}$, and
- perform posterior inference over this model class.

Being Bayesian with Respect to Causal Models

Epistemic challenge: true causal model $\mathcal{M}^\star$ is not (completely) known.

Bayesian approach:

1. place a prior $p(\mathcal{M})$ over causal models,
2. collect data $\mathcal{D}$ from the true model $\mathcal{M}^\star$,
3. compute the posterior via Bayes rule:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{\int p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})\, \mathrm{d}\mathcal{M}}\,.$$

Computationally delicate, as we require a way to

- parametrise the class of models $\mathcal{M}$, and
- perform posterior inference over this model class.

Being Bayesian with Respect to Causal Models

Epistemic challenge: true causal model $\mathcal{M}^\star$ is not (completely) known.

Bayesian approach:

1. place a prior $p(\mathcal{M})$ over causal models,
2. collect data $\mathcal{D}$ from the true model $\mathcal{M}^\star$,
3. compute the posterior via Bayes rule:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{\int p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})\, \mathrm{d}\mathcal{M}}\,.$$

Computationally delicate, as we require a way to

- parametrise the class of models $\mathcal{M}$, and
- perform posterior inference over this model class.

Target Causal Query

Causal query function q specifies a *target causal query* $Y = q(\mathcal{M})$:

Causal Discovery: $Y = q_{\mathrm{CD}}(\mathcal{M}) = G$

Partial Causal Discovery: $Y = q_{\mathrm{PCD}}(\mathcal{M}) = \phi(G)$

Causal Model Learning: $Y = q_{\mathrm{CML}}(\mathcal{M}) = \mathcal{M}$

Causal Reasoning: $Y = q_{\mathrm{CR}}(\mathcal{M}) = \{p^{\mathrm{do}(\boldsymbol{X}_{\mathcal{I}(j)})}(X_j \,|\, \mathcal{M})\}_{j \in \mathcal{J}},$

Bayesian inference naturally extends to the *query posterior*:

$$p(Y \,|\, \mathcal{D}) = \int p(Y \,|\, \mathcal{M})\, p(\mathcal{M} \,|\, \mathcal{D})\, \mathrm{d}\mathcal{M} = \mathbb{E}_{\mathcal{M} \,|\, \mathcal{D}}[\, p(Y \,|\, \mathcal{M})],$$

Target Causal Query

Causal query function q specifies a *target causal query* $Y = q(\mathcal{M})$:

Causal Discovery: $\quad Y = q_{\mathrm{CD}}(\mathcal{M}) = G$

Partial Causal Discovery: $\quad Y = q_{\mathrm{PCD}}(\mathcal{M}) = \phi(G)$

Causal Model Learning: $\quad Y = q_{\mathrm{CML}}(\mathcal{M}) = \mathcal{M}$

Causal Reasoning: $\quad Y = q_{\mathrm{CR}}(\mathcal{M}) = \{p^{\mathrm{do}(\boldsymbol{X}_{\mathcal{I}(j)})}(X_j \mid \mathcal{M})\}_{j \in \mathcal{J}},$

Bayesian inference naturally extends to the *query posterior*:

$$p(Y \mid \mathcal{D}) = \int p(Y \mid \mathcal{M})\, p(\mathcal{M} \mid \mathcal{D})\, \mathrm{d}\mathcal{M} = \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}[\, p(Y \mid \mathcal{M})],$$

Active Learning with Sequential Interventions

At each time t, can perform an experiment a_t and observe outcome:

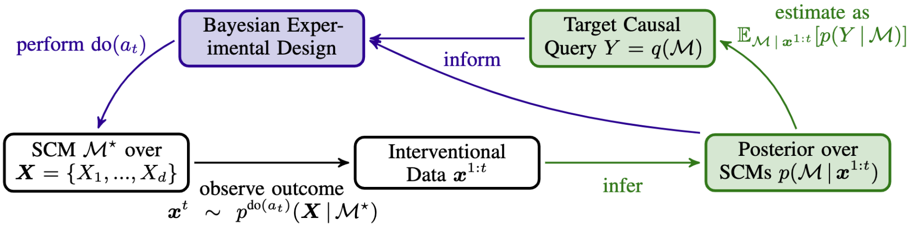
$$\boldsymbol{x}^t = \{\boldsymbol{x}^{t,n}\}_{n=1}^{N_t}, \qquad \boldsymbol{x}^{t,n} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M}^\star)$$

Design experiment a_t to be *maximally informative* about causal query Y:

$$\max_{a_t} \mathrm{I}(Y; \boldsymbol{X}^t \mid \boldsymbol{x}^{1:t-1})$$

where $\boldsymbol{X}^t$ follows the predictive interventional distribution:

$$\boldsymbol{X}^t \sim p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \boldsymbol{x}^{1:t-1}) \propto \int p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M})\, p(\mathcal{M} \mid \boldsymbol{x}^{1:t-1})\, \mathrm{d}\mathcal{M}.$$



Active Learning with Sequential Interventions

At each time t, can perform an experiment a_t and observe outcome:

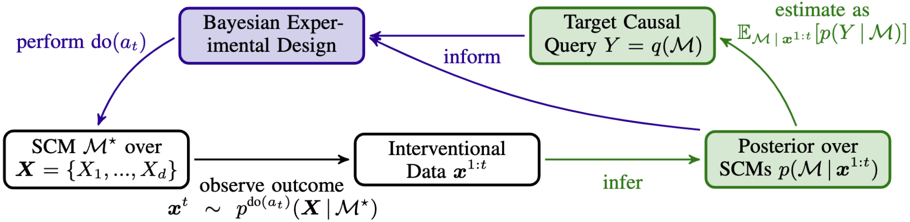
$$\boldsymbol{x}^t = \{\boldsymbol{x}^{t,n}\}_{n=1}^{N_t}, \qquad \boldsymbol{x}^{t,n} \overset{\text{i.i.d.}}{\sim} p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M}^\star)$$

Design experiment a_t to be *maximally informative* about causal query Y:

$$\max_{a_t} \mathrm{I}(Y; \boldsymbol{X}^t \mid \boldsymbol{x}^{1:t-1})$$

where $\boldsymbol{X}^t$ follows the predictive interventional distribution:

$$\boldsymbol{X}^t \sim p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \boldsymbol{x}^{1:t-1}) \propto \int p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M})\, p(\mathcal{M} \mid \boldsymbol{x}^{1:t-1})\, \mathrm{d}\mathcal{M}.$$



Outline

Model Class and Parametrisation

Nonlinear additive Gaussian noise models:

$$X_i := f_i(\mathbf{Pa}_i) + U_i, \quad \text{with} \quad U_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_i^2) \quad \text{for} \quad i \in \{1, \dots, d\}, \tag{2}$$

Mutually independent $U_i \rightarrow$ causal sufficiency/no hidden confounding.

Can parametrise such models $\mathcal{M}$ as triples $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$, where

- G is a causal DAG,
- $\boldsymbol{f} = (f_1, \dots, f_d)$ are functions over the parent sets implied by G,
- $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_d^2)$ are the Gaussian noise variances.

Interventional Likelihood

Consider hard interventions $\mathrm{do}(a_t) = \mathrm{do}(\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I})$ for $\boldsymbol{X}_\mathcal{I} \subseteq \boldsymbol{W}$.

Due to causal sufficiency and Gaussian noise:

$$p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \mathbf{1}_{\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I}} \prod_{j \notin \mathcal{I}} p(X_j \mid \mathbf{Pa}_j^G)$$
$$= \mathbf{1}_{\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I}} \prod_{j \notin \mathcal{I}} \mathcal{N}(f_j(\mathbf{Pa}_j^G), \sigma_j^2).$$

The likelihood of the entire dataset $\boldsymbol{x}^{1:t}$ collected up to time t is:

$$p(\boldsymbol{x}^{1:t} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \prod_{\tau=1}^{t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^\tau \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$$
$$= \prod_{\tau=1}^{t} \prod_{n=1}^{N_t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^{\tau,n} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2).$$

Interventional Likelihood

Consider hard interventions $\mathrm{do}(a_t) = \mathrm{do}(\boldsymbol{X}_{\mathcal{I}} = \boldsymbol{x}_{\mathcal{I}})$ for $\boldsymbol{X}_{\mathcal{I}} \subseteq \boldsymbol{W}$.

Due to causal sufficiency and Gaussian noise:

$$
\begin{aligned}
p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) &= \mathbf{1}_{\boldsymbol{X}_{\mathcal{I}}=\boldsymbol{x}_{\mathcal{I}}} \prod_{j \notin \mathcal{I}} p(X_j \mid \mathbf{Pa}_j^G) \\
&= \mathbf{1}_{\boldsymbol{X}_{\mathcal{I}}=\boldsymbol{x}_{\mathcal{I}}} \prod_{j \notin \mathcal{I}} \mathcal{N}(f_j(\mathbf{Pa}_j^G), \sigma_j^2).
\end{aligned}
$$

The likelihood of the entire dataset $\boldsymbol{x}^{1:t}$ collected up to time t is:

$$
\begin{aligned}
p(\boldsymbol{x}^{1:t} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) &= \prod_{\tau=1}^{t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^{\tau} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) \\
&= \prod_{\tau=1}^{t} \prod_{n=1}^{N_t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^{\tau,n} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2).
\end{aligned}
$$

Interventional Likelihood

Consider hard interventions $\mathrm{do}(a_t) = \mathrm{do}(\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I})$ for $\boldsymbol{X}_\mathcal{I} \subseteq \boldsymbol{W}$.

Due to causal sufficiency and Gaussian noise:

$$p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \mathbf{1}_{\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I}} \prod_{j \notin \mathcal{I}} p(X_j \mid \mathbf{Pa}_j^G)$$
$$= \mathbf{1}_{\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I}} \prod_{j \notin \mathcal{I}} \mathcal{N}(f_j(\mathbf{Pa}_j^G), \sigma_j^2).$$

The likelihood of the entire dataset $\boldsymbol{x}^{1:t}$ collected up to time t is:

$$p(\boldsymbol{x}^{1:t} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \prod_{\tau=1}^{t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^\tau \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$$
$$= \prod_{\tau=1}^{t} \prod_{n=1}^{N_t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^{\tau,n} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2).$$

Model Prior

For a given causal graph G, distinguish between

- root nodes $\mathbf{R}(G) = \{i \in [d] : \mathbf{Pa}_i^G = \emptyset\}$ with $f_i = \text{const}$
- non-root nodes $\mathbf{NR}(G) = [d] \setminus \mathbf{R}(G)$.

Place the following structured prior over SCMs $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$:

$$p(\mathcal{M}) = p(G) \prod_{i \in \mathbf{R}(G)} p(f_i, \sigma_i^2 \mid G) \prod_{j \in \mathbf{NR}(G)} p(f_j \mid G)\, p(\sigma_j^2 \mid G) \, .$$

Model Prior

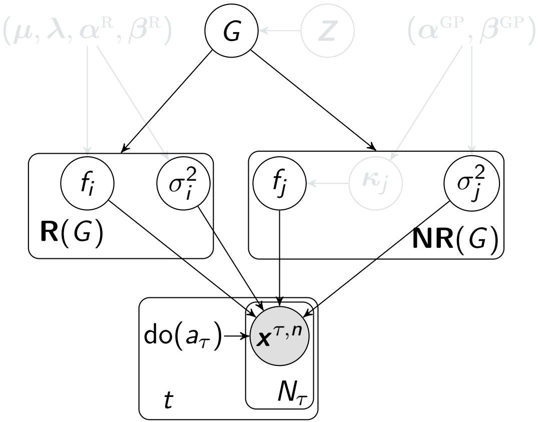
For a given causal graph G, distinguish between

- root nodes $\mathbf{R}(G) = \{i \in [d] : \mathbf{Pa}_i^G = \varnothing\}$ with $f_i = \text{const}$
- non-root nodes $\mathbf{NR}(G) = [d] \setminus \mathbf{R}(G)$.

Place the following structured prior over SCMs $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$:

$$p(\mathcal{M}) = p(G) \prod_{i \in \mathbf{R}(G)} p(f_i, \sigma_i^2 \mid G) \prod_{j \in \mathbf{NR}(G)} p(f_j \mid G)\, p(\sigma_j^2 \mid G)\,.$$

Graphical Model Representation



Model Posterior

Given $\boldsymbol{x}^{1:t}$, the posterior over SCMs $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$ can be written as

$$p(\mathcal{M} \mid \boldsymbol{x}^{1:t}) = p(G \mid \boldsymbol{x}^{1:t}) \prod_{i \in \mathbf{R}(G)} p(f_i, \sigma_i^2 \mid \boldsymbol{x}^{1:t}, G) \prod_{j \in \mathbf{NR}(G)} p(f_j, \sigma_j^2 \mid \boldsymbol{x}^{1:t}, G)\,.$$

For root nodes: conjugate N-$\Gamma^{-1}(\mu_i, \lambda_i, \alpha_i^{\mathrm{R}}, \beta_i^{\mathrm{R}})$ priors on $p(f_i, \sigma_i^2 \mid G)$
$\implies$ closed form for $p(f_i, \sigma_i^2 \mid \boldsymbol{x}^{1:t}, G)$.

The graph and non-root node posteriors are more tricky:

$$p(G \mid \boldsymbol{x}^{1:t}) = \frac{p(\boldsymbol{x}^{1:t} \mid G)\, p(G)}{p(\boldsymbol{x}^{1:t})}\,,$$

$$p(f_j, \sigma_j^2 \mid \boldsymbol{x}^{1:t}, G) = \frac{p(\boldsymbol{x}^{1:t} \mid G, f_j, \sigma_j^2)\, p(f_j, \sigma_j^2 \mid G)}{p(\boldsymbol{x}^{1:t} \mid G)}\,.$$

Challenge 1: Marginalising out the Functions

$$p(\boldsymbol{x}^{1:t} \mid G) = \int p(\boldsymbol{x}^{1:t} \mid G, f_j, \sigma_j^2)\, p(f_j \mid G)\, p(\sigma_j^2 \mid G)\, \mathrm{d}f_j\, \mathrm{d}\sigma_j^2$$

Gaussian processes (GPs)[1]: *nonlinear* functions + analytical expressions.

$$\begin{aligned} p(f_j \mid G, \boldsymbol{\kappa}_j) &= \mathcal{GP}(0, k_j^G(\cdot, \cdot; \boldsymbol{\kappa}_j)), \\ p(\sigma_j^2 \mid G) &= \Gamma(\alpha_j^\sigma, \beta_j^\sigma), \\ p(\kappa_j \mid G) &= \Gamma(\alpha_j^\kappa, \beta_j^\kappa) \end{aligned}$$

where $k_j^G(\cdot, \cdot; \boldsymbol{\kappa}_j)$ is a covariance function over $\mathbf{Pa}_j^G$ with length scales $\boldsymbol{\kappa}_j$.

$\implies$ closed-form GP-marginal likelihood $p(\boldsymbol{x}^{1:t} \mid G, \sigma_j^2, \boldsymbol{\kappa}_j)$, posteriors $p(f_j \mid \boldsymbol{x}^{1:t}, G, \sigma_j^2, \boldsymbol{\kappa}_j)$ and predictive posteriors $p(\boldsymbol{X} \mid \boldsymbol{x}^{1:t}, G, \boldsymbol{\sigma}^2, \boldsymbol{\kappa})$

[1] Williams and Rasmussen 2006.

Challenge 2: Marginalising out the GP-Hyperparameters

In general, no analytical expression for $p(\sigma_j^2, \kappa_j \mid \boldsymbol{x}^{1:t}, G)$.

Approximate expectations w.r.t. posterior with MAP estimate $(\hat{\sigma}_j^2, \hat{\boldsymbol{\kappa}}_j)$:

$$p(f_j \mid \boldsymbol{x}^{1:t}, G) \approx p(f_j \mid \boldsymbol{x}^{1:t}, G, \hat{\sigma}_j^2, \hat{\boldsymbol{\kappa}}_j)$$

obtained via gradient ascent on the log posterior:

$$\nabla \log p(\sigma_j^2, \boldsymbol{\kappa}_j \mid \boldsymbol{x}^{1:t}, G) = \nabla \log p(\boldsymbol{x}^{1:t} \mid G, \sigma_j^2, \boldsymbol{\kappa}_j) + \nabla \log p(\sigma_j^2, \boldsymbol{\kappa}_j \mid G)\,.$$

Challenge 2: Marginalising out the GP-Hyperparameters

In general, no analytical expression for $p(\sigma_j^2, \kappa_j \,|\, \boldsymbol{x}^{1:t}, G)$.

Approximate expectations w.r.t. posterior with MAP estimate $(\hat{\sigma}_j^2, \hat{\boldsymbol{\kappa}}_j)$:

$$p(f_j \,|\, \boldsymbol{x}^{1:t}, G) \approx p(f_j \,|\, \boldsymbol{x}^{1:t}, G, \hat{\sigma}_j^2, \hat{\boldsymbol{\kappa}}_j)$$

obtained via gradient ascent on the log posterior:

$$\nabla \log p(\sigma_j^2, \boldsymbol{\kappa}_j \,|\, \boldsymbol{x}^{1:t}, G) = \nabla \log p(\boldsymbol{x}^{1:t} \,|\, G, \sigma_j^2, \boldsymbol{\kappa}_j) + \nabla \log p(\sigma_j^2, \boldsymbol{\kappa}_j \,|\, G)\,.$$

Challenge 3: Marginalising out the Graphs

$$p(\boldsymbol{x}^{1:t}) = \sum_G p(\boldsymbol{x}^{1:t} \mid G)\, p(G)$$

Intractable for $d \geq 5$ (# DAGs grows super-exponentially in d).

DiBS (Lorch et al. 2021): continuous prior $p(\boldsymbol{Z})$ models G via $p(G \mid \boldsymbol{Z})$ and simultaneously enforces acyclicity of G.

$\rightarrow$ can efficiently infer expectations w.r.t. $p(G \mid \boldsymbol{x}^{1:t})$ via $p(\boldsymbol{Z} \mid \boldsymbol{x}^{1:t})$.

Stein Variational Gradient Descent[2] to approximately infer $p(\boldsymbol{Z} \mid \boldsymbol{x}^{1:t})$.

[2] Liu and Wang 2016.

Challenge 3: Marginalising out the Graphs

$$p(\boldsymbol{x}^{1:t}) = \sum_{G} p(\boldsymbol{x}^{1:t} \mid G)\, p(G)$$

Intractable for $d \geq 5$ ($\#$ DAGs grows super-exponentially in d).

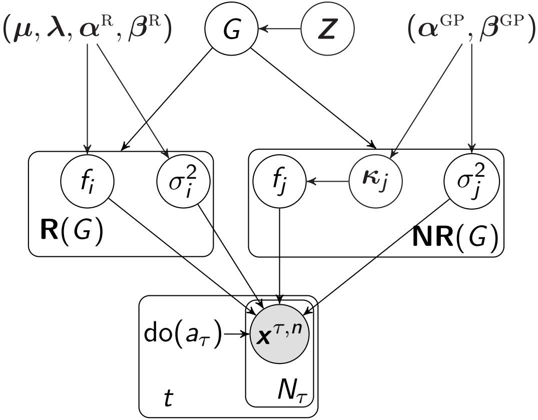
DiBS (Lorch et al. 2021): continuous prior $p(\boldsymbol{Z})$ models G via $p(G \mid \boldsymbol{Z})$ and simultaneously enforces acyclicity of G.

$\rightarrow$ can efficiently infer expectations w.r.t. $p(G \mid \boldsymbol{x}^{1:t})$ via $p(\boldsymbol{Z} \mid \boldsymbol{x}^{1:t})$.

Stein Variational Gradient Descent[2] to approximately infer $p(\boldsymbol{Z} \mid \boldsymbol{x}^{1:t})$.

[2] Liu and Wang 2016.

Graphical Model Representation



Experimental Design

Given:

- previously collected data $\mathcal{D} = \boldsymbol{x}^{1:t-1}$,
- target causal query Y,

choose optimal next intervention $a_t^* = (\mathcal{I}^*, \boldsymbol{x}_{\mathcal{I}}^*)$ by maximising

$$U_Y(a) = H(\boldsymbol{X}^t \,|\, \mathcal{D}) + \mathbb{E}_{\mathcal{M}\,|\,\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{X}^t, Y\,|\,\mathcal{M}}\left[\log \mathbb{E}_{\mathcal{M}'\,|\,\mathcal{D}}\left[p(\boldsymbol{X}^t \,|\, \mathcal{M}')\, p(Y \,|\, \mathcal{M}')\right]\right]\right]$$

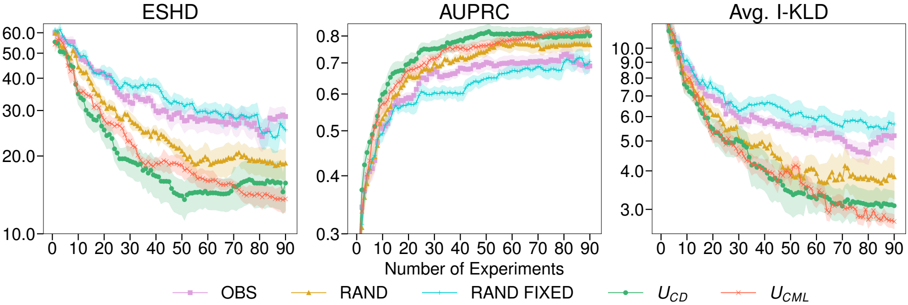
Nested, bi-level optimization scheme:

$$\begin{aligned} \forall \mathcal{I}: \quad & \boldsymbol{x}_{\mathcal{I}}^* \in \arg\max_{\boldsymbol{x}_{\mathcal{I}}} U_Y(\mathcal{I}, \boldsymbol{x}_{\mathcal{I}}), && \text{(Bayesian Optimisation)} \\ & \mathcal{I}^* \in \arg\max_{\mathcal{I}} U_Y(\mathcal{I}, \boldsymbol{x}_{\mathcal{I}}^*). && (|\mathcal{I}| \leq k, \text{ here: } k = 1) \end{aligned}$$

Outline

Experiment 1: Causal Discovery and Model Learning

Random scale-free graphs, 20 nodes, 5 ground truth SCMs, 6 runs each; initialise with 5 obs. samples, then 3 samples per experiment.

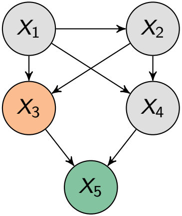


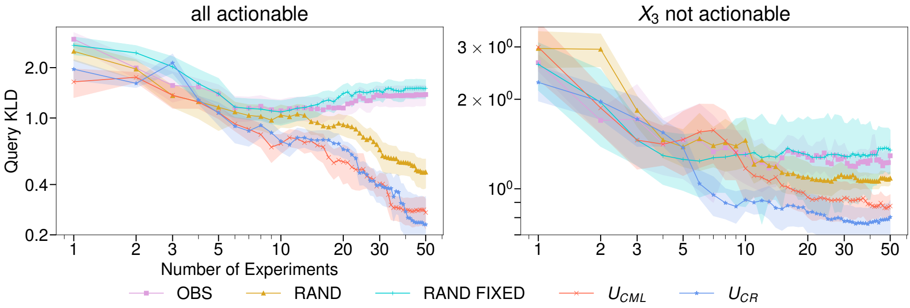
1. **ESHD:** Expected Structural Hamming Distance
2. **AUPRC:** Area Under Precision Recall Curve (for predicting edges)
3. **Average I-KLD:** Average KL between true and inferred single-node interventional distributions (proxy for SCM learning).

Experiment 2: Causal Reasoning

Unknown ground truth graph over 5 nodes:

Query: $p^{\mathrm{do}(X_3=\psi)}(X_5 \mid \mathcal{M})$ with $\psi \sim \mathcal{U}[4,7]$





Outline

Related Work on Active Bayesian Causal Discovery

Work	**Target Query**	**Model Class**
(Tong and Koller 2001), (Murphy 2001)	causal graph G	Conjugate Dirichlet-Multinomial
(Cho, Berger, and Peng 2016)	causal graph G	Conjugate linear Gaussian-inverse-Gamma
(Agrawal et al. 2019)	some function $\phi(G)$ of the causal graph G	Linear Gaussian
(Tigas et al. 2022)	causal graph G and parameters of f_i	Additive Gaussian noise with parametric neural network functions f_i
GP-DiBS-ABCI (ours)	some function $q(\mathcal{M})$ of the full SCM $\mathcal{M}$	Additive Gaussian noise with nonparametric functions f_i modeled by GPs

Limitations and Extensions

In our GP-DiBS-ABCI approach, we did not consider:

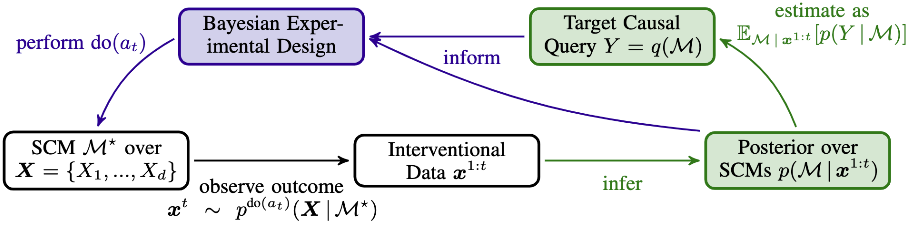
- hidden confounding
- cyclic causal relationships
- heteroscedastic noise
- soft interventions
- counterfactual queries
- causal models other than SCMs

Future work: implementations for richer model classes + extensions.

In principle, possible within the ABCI framework, but can be challenging with regard to model parametrisation and tractable inference.

Summary

Principled, flexible framework for active Bayesian causal inference:



Useful when actively collecting (some) interventional data is feasible, but expensive relative to compute (e.g., for biological applications).

References I

[1] Raj Agrawal et al. "ABCD-strategy: Budgeted experimental design for targeted causal structure discovery". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 3400–3409.

[2] Hyunghoon Cho, Bonnie Berger, and Jian Peng. "Reconstructing causal biological networks through active learning". In: *PloS one* 11.3 (2016), e0150611.

[3] Qiang Liu and Dilin Wang. "Stein variational gradient descent: A general purpose Bayesian inference algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by D Lee et al. Vol. 29. Curran Associates, Inc., 2016.

[4] Lars Lorch et al. "DiBS: Differentiable Bayesian Structure Learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[5] Kevin P Murphy. *Active learning of causal Bayes net structure*. 2001.

[6] Judea Pearl. *Causality*. 2nd. Cambridge University Press, 2009.

[7] Panagiotis Tigas et al. "Interventions, Where and How? Experimental Design for Causal Models at Scale". In: *arXiv preprint arXiv:2203.02016* (2022).

[8] Simon Tong and Daphne Koller. "Active learning for structure in Bayesian networks". In: *International Joint Conference on Artificial Intelligence*. Vol. 17. 2001, pp. 863–869.

[9] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. Vol. 2. MIT Press Cambridge, MA, 2006.