UNIVERSITY OF AMSTERDAM
Faculty of Science

QUVA Deep Vision Lab

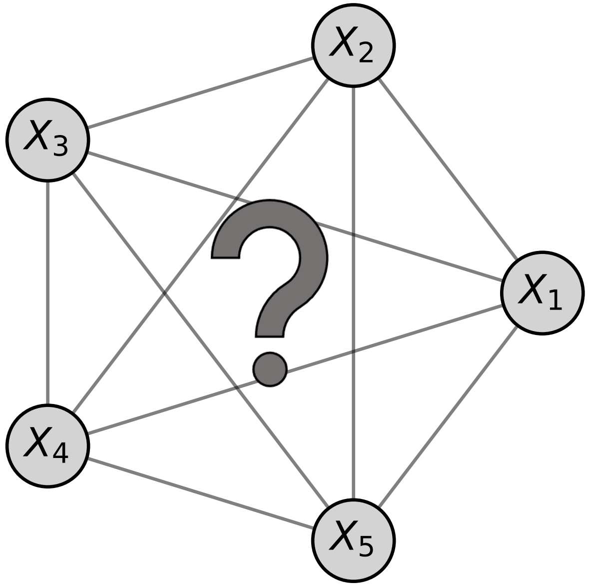
Qualcomm AI research

Efficient Neural Causal Discovery without Acyclicity Constraints

Phillip Lippe, Taco Cohen, Efstratios Gavves

Background: Neural Causal Discovery

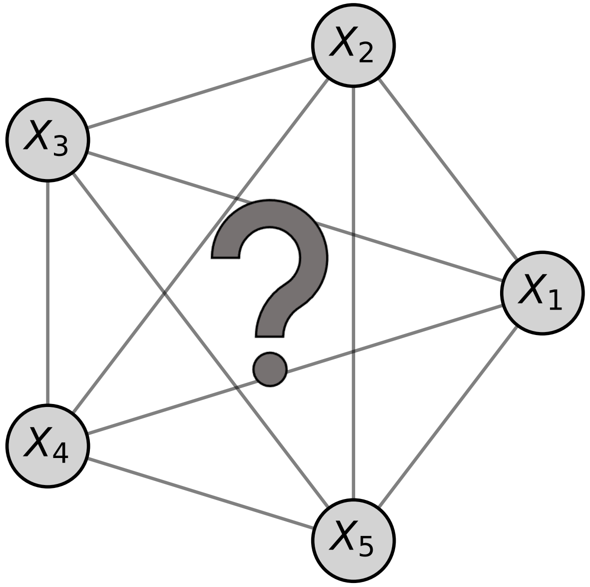
- *Causal structure learning*: find directed acyclic graph from observational and interventional data
- Recent work: continuous-optimization score-based causal discovery
  - Search the space of possible graphs with gradient based methods
  - Adjacency matrix parameterized by independent probabilities per edge
- Main problem: limit the search space to directed acyclic graphs
  - Constraint-based optimization: $h(W) = \mathrm{tr}\left(e^{W \circ W}\right) - d = 0$
    ⇒ Slow and hyperparameter sensitive
  - Regularization: penalize cyclic graphs
    ⇒ Hyperparameter sensitive and limited guarantees



How can we reliably perform causal discovery with gradient-based methods on large scales?

Scope and Assumptions

- Find the DAG of a causal graphical model (CGM) from observational and interventional samples
  - Variables can be discrete, continuous, or mixed
- CGM is causally sufficient
  - Extension to latent confounders possible
- Interventions are:
  - Sparse (single variable)
  - Perfect (independent of original parents)
  - Available for all observable variables



ENCO: Efficient Neural Causal Discovery

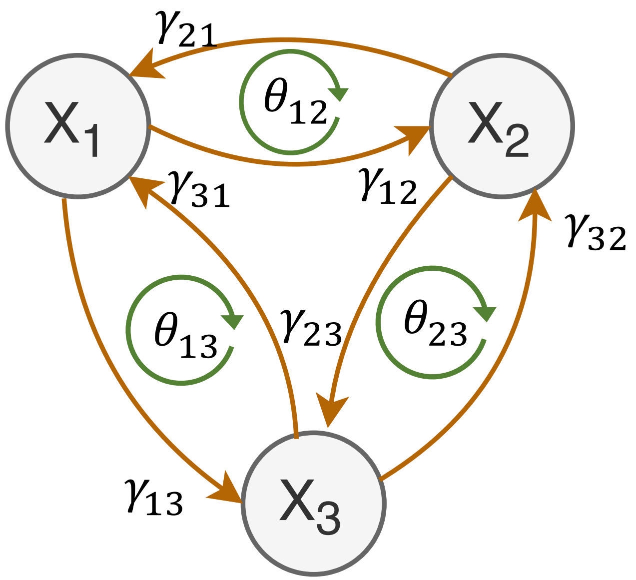
- Central idea: learn distributions $p(X_1| \dots)$ from observational data, test generalization to interventional data

- Parameterize graph with edge existence and orientation parameters
  - Probability of an edge: $\sigma(\gamma_{ij}) \cdot \sigma(\theta_{ij})$, with $\theta_{ij} = -\theta_{ji}$

- Benefits of two-variable parameterisation:

  ⇒ More control over gradient updates

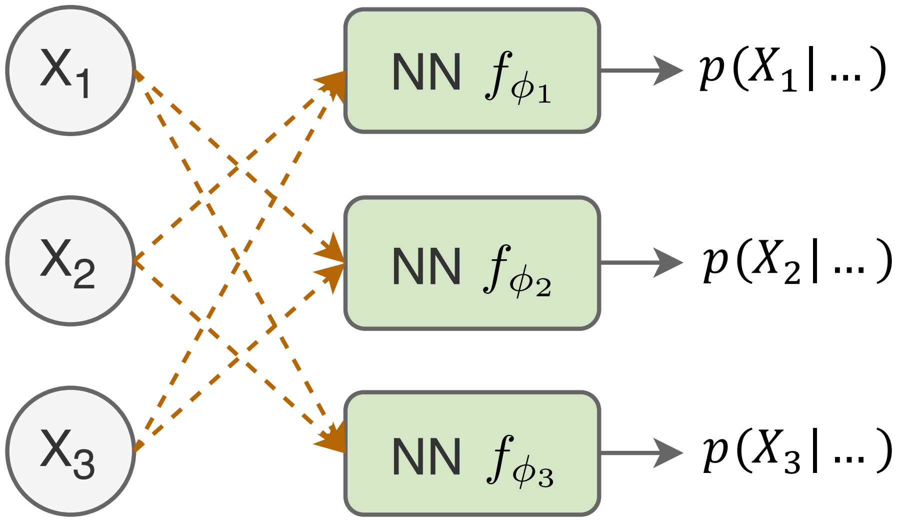
  ⇒ No constraint or regularization for acyclicity needed!



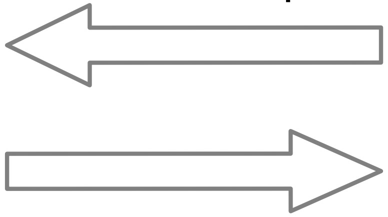
ENCO: Efficient Neural Causal Discovery

Overview

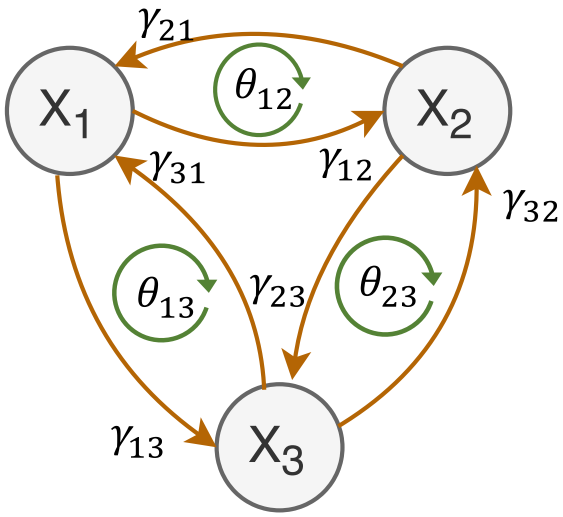
Distribution fitting



Alternate between both steps



Graph fitting



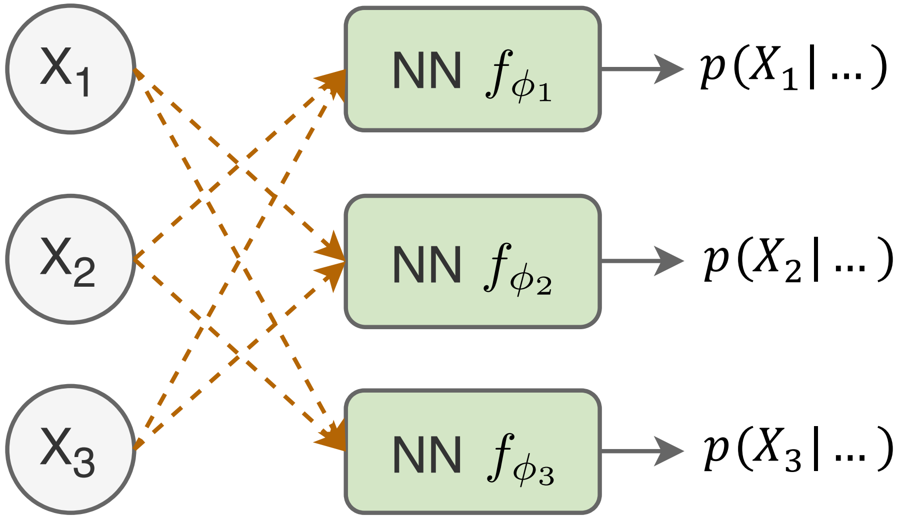
➜ Learn neural networks fitting conditional distributions on observational data

➜ Learn edge and orientation parameters based on fitted distributions

ENCO: Efficient Neural Causal Discovery

Objectives

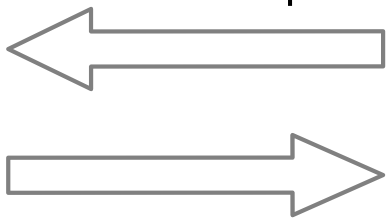
Distribution fitting



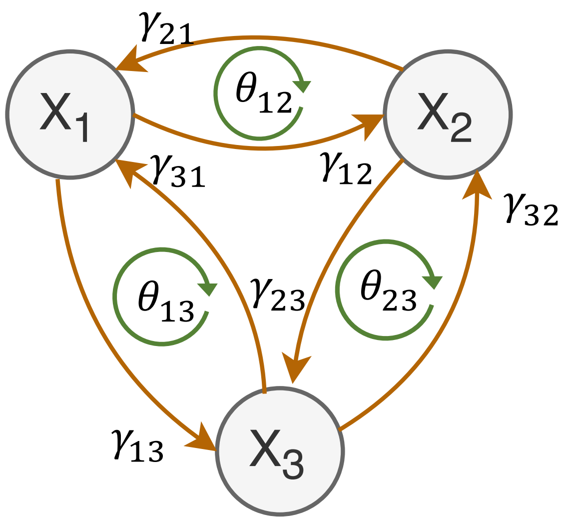
$$\min_{\phi_i} \mathbb{E}_{\boldsymbol{X}} \mathbb{E}_{\boldsymbol{M}} \left[-\log f_{\phi_i}(X_i; \boldsymbol{M}_{-i} \odot \boldsymbol{X}_{-i}) \right]$$

$$M_j \sim \mathrm{Ber}(p(X_j \to X_i))$$

Alternate between both steps



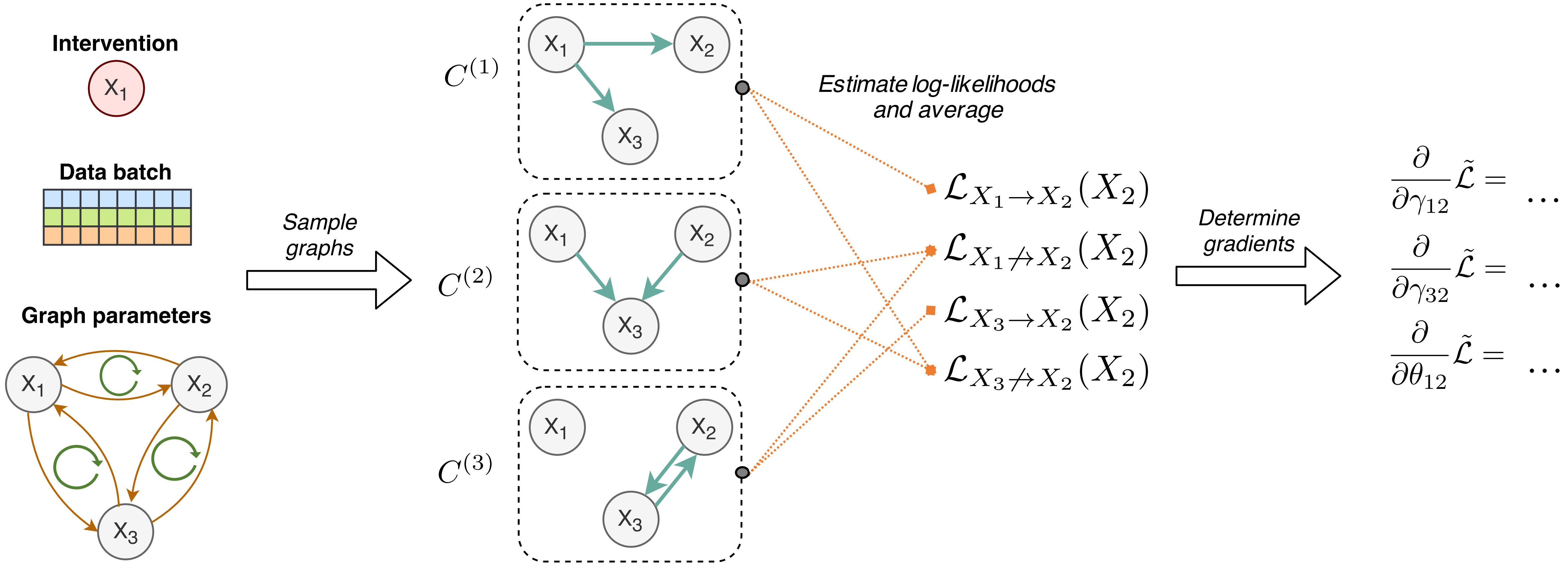
Graph fitting



$$\tilde{\mathcal{L}} = \mathbb{E}_{\hat{I} \sim p_I(I)} \mathbb{E}_{\tilde{p}_{\hat{I}}(\boldsymbol{X})} \mathbb{E}_{p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)} \left[\sum_{i=1}^{N} \mathcal{L}_C(X_i) \right]$$

$$+ \lambda_{\text{sparse}} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma(\gamma_{ij}) \cdot \sigma(\theta_{ij})$$

Graph fitting

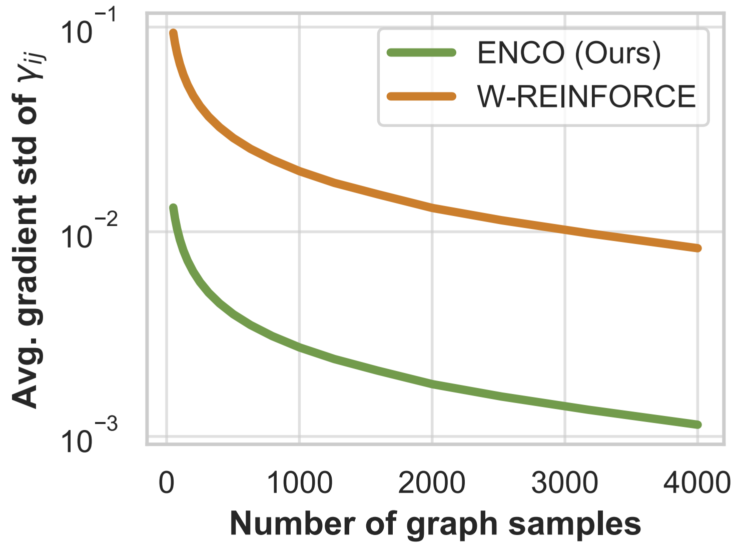


Gradient estimators

- Efficient low-variance, unbiased gradient estimators for edge and orientation parameters
- Edge gradients:

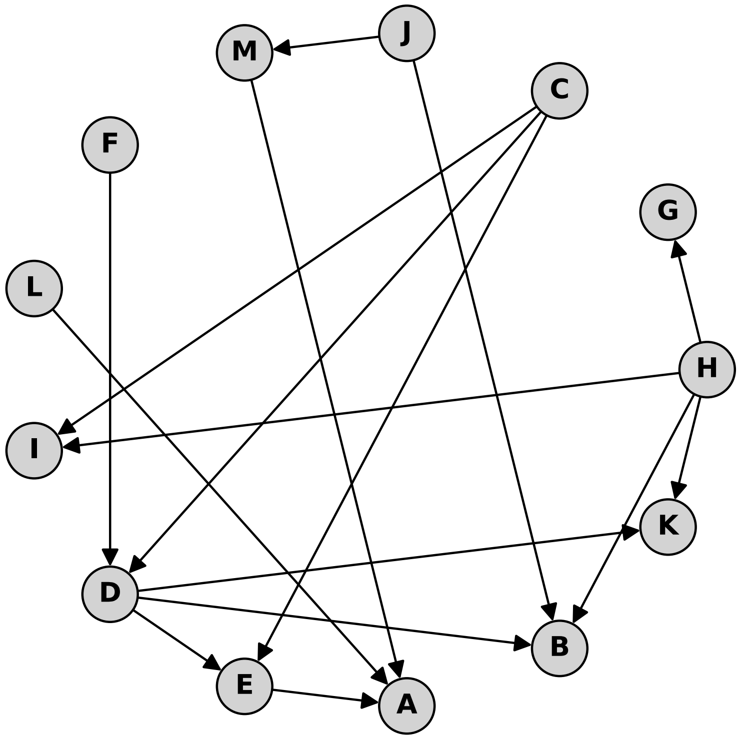
$$\frac{\partial}{\partial \gamma_{ij}}\mathcal{L} = \alpha \cdot \underbrace{\mathbb{E}_{\boldsymbol{X},C_{-ij}}}_{\text{Graph/Data samples}}\Big[\underbrace{\mathcal{L}_{X_i\to X_j}(X_j) - \mathcal{L}_{X_i\not\to X_j}(X_j)}_{\text{Log likelihood w/o edge}} + \underbrace{\lambda_{\text{sparse}}}_{\text{Sparsity regularizer}}\Big]$$

- Sample and evaluate K graphs to estimate whether an edge is "beneficial" or not
- Similar idea for orientation parameters, but only with adjacent interventional data

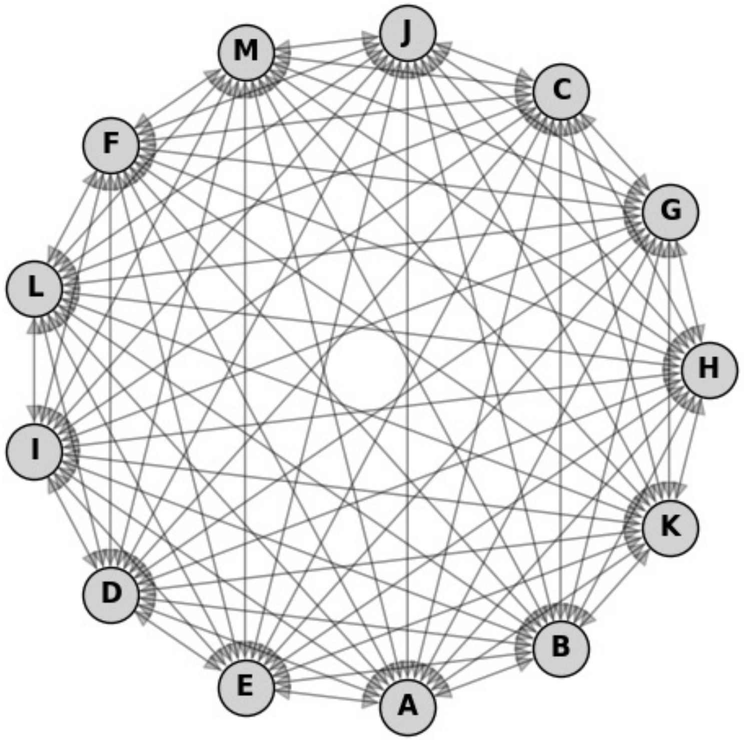


Learning causal graphs

Ground truth causal graph

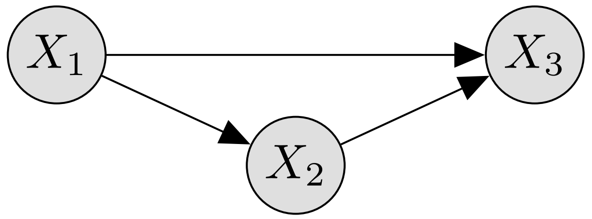


Learned edge probabilities



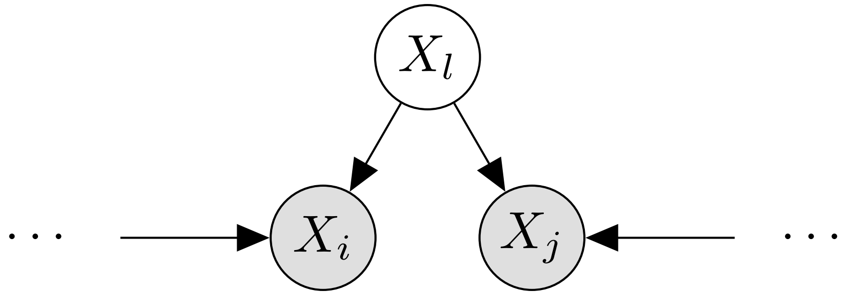
Convergence

- Theoretical guarantees can be given for ENCO converging to the true causal graph
- **Main conditions**: for every edge $X_i \rightarrow X_j$ in the causal graph,
  - the edge $X_i \rightarrow X_j$ must not be disadvantegous for the log likelihood estimate of X_j under interventions on X_i
  - the edge $X_i \rightarrow X_j$ must have a greater impact on the log likelihood estimate than the sparsity regularizer λ_{sparse}
- If the conditions are not fulfilled, local minima can exist



Latent confounders

- A latent confounder on two variables causes a unique pattern
  - On interventions on X_i and X_j, an edge is disadvantegous in both directions
  - On interventions on other variables, edges are beneficial



- Find confounders by tracking γ-parameters on adjacent interventions and other interventions
  - Score pairs of variables on pattern:

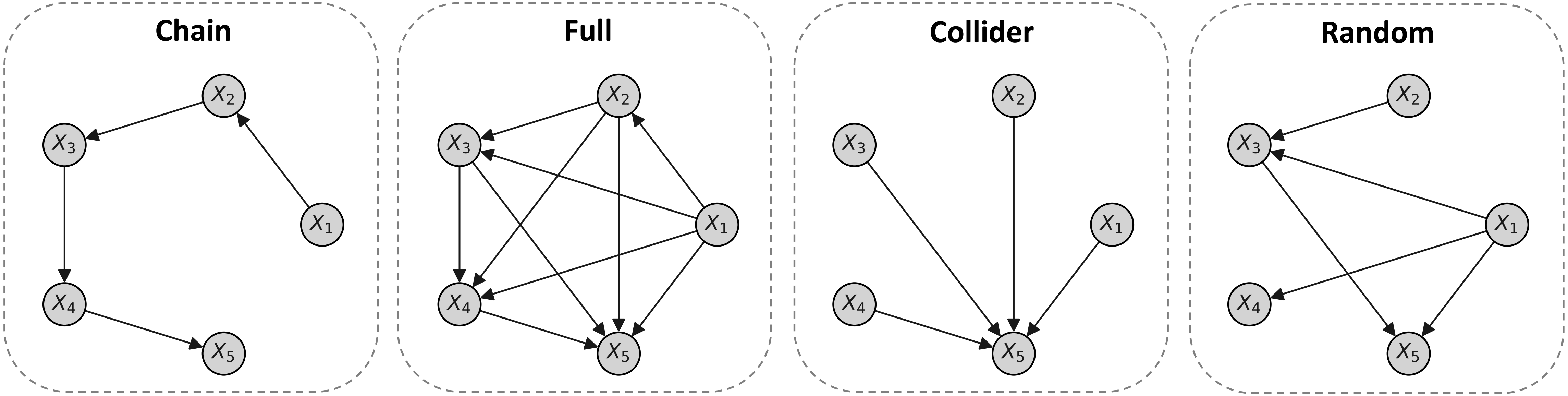
$$\mathrm{lc}(X_i, X_j) = \sigma\left(\gamma_{ij}^{(O)}\right) \cdot \sigma\left(\gamma_{ji}^{(O)}\right) \cdot \left(1 - \sigma\left(\gamma_{ij}^{(I)}\right)\right) \cdot \left(1 - \sigma\left(\gamma_{ji}^{(I)}\right)\right)$$

  - $\mathrm{lc}(X_i, X_j)$ goes to 1 if X_i, X_j share a confounder

Experiments

Synthetic graphs

- Recover syntheticly generated graphs
- Testing various common graph forms to find weaknesses
- Graph size: 25 nodes
- Metric: Structural Hamming Distance (SHD) = FP + FN + wrongly orientated edges



Experiments

Synthetic graphs

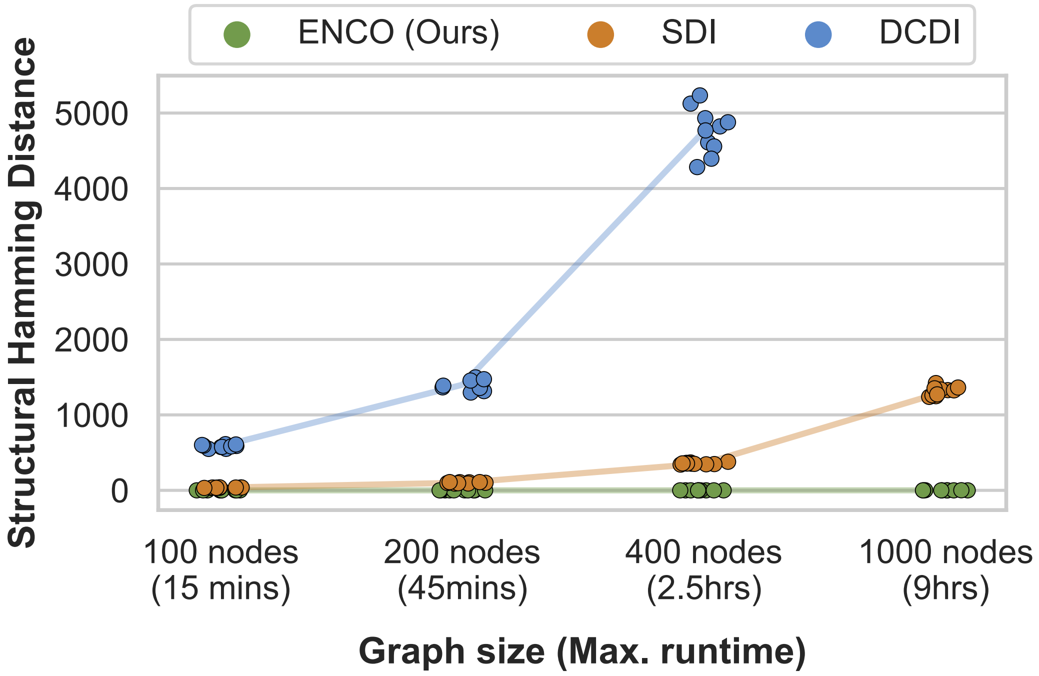
- Recover syntheticly generated graphs
- Testing various common graph forms to find weaknesses
- Graph size: 25 nodes
- Metric: Structural Hamming Distance (SHD) = FP + FN + wrongly orientated edges

Graph type	bidiag	chain	collider	full	jungle	random
GIES [Hauser and Bühlmann, 2012]	47.4 (±5.2)	22.3 (±3.5)	13.3 (±3.0)	152.7 (±12.0)	53.9 (±8.9)	86.1 (±12.0)
IGSP [Wang et al., 2017]	33.0 (±4.2)	12.0 (±1.9)	23.4 (±2.2)	264.6 (±7.4)	38.6 (±5.7)	76.3 (±7.7)
SDI [Ke et al., 2019]	2.1 (±1.5)	0.8 (±0.9)	14.7 (±4.0)	121.6 (±18.4)	1.8 (±1.6)	1.8 (±1.9)
DCDI [Brouillard et al., 2020]	3.7 (±1.5)	4.0 (±1.3)	**0.0** (±0.0)	2.8 (±2.1)	1.2 (±1.5)	2.2 (±1.5)
ENCO (Ours)	**0.0** (±0.0)	**0.0** (±0.0)	**0.0** (±0.0)	**0.3** (±0.9)	**0.0** (±0.0)	**0.0** (±0.0)

Experiments

Scalability

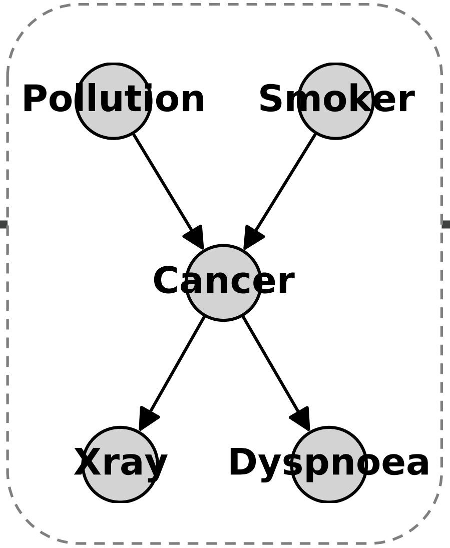
- Testing scalability of the approach with synthetic graphs of up to 1000 nodes
- All baselines got the same computational resources
- On average, less than 1 mistake among 1 million edges for largest graph



Experiments

BnLearn Repository

- Experiments on real-world inspired causal graphs from BnLearn repository [Scutari, 2010]
- Deterministic variables and very rare events



Graph "cancer"

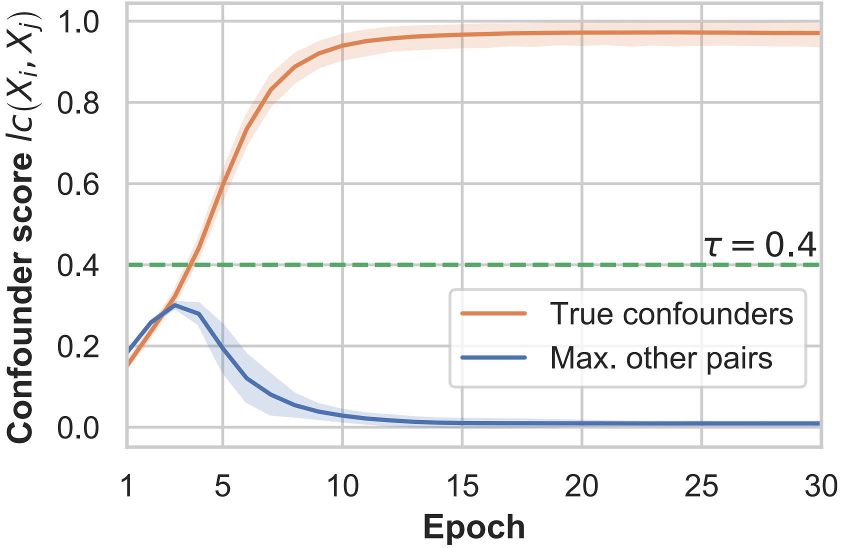
Dataset	cancer (5 nodes)	asia (8 nodes)	sachs (11 nodes)	child (20 nodes)	alarm (37 nodes)	diabetes (413 nodes)	pigs (441 nodes)
SDI [Ke et al., 2019]	3.0	4.0	7.0	11.8	24.6	422.4	18.0
ENCO (Ours)	**0.0**	**0.0**	**0.0**	**0.0**	**1.0**	**2.0**	**0.0**

Experiments

Latent confounders

- Synthetic, random graphs with 5 additional latent confounders
- Detecting confounders by thresholding pairwise scores

Metrics	ENCO
SHD	0.0 (±0.0)
Confounder recall	96.8% (±9.5%)
Confounder precision	100.0% (±0.0%)



Conclusion

- ENCO: method for finding causal relations from observational and interventional data
- Main characteristics of approach:
  - Score function unconstrained in terms of acyclicity
  - Scalable in both dataset and graph size
  - Guarantees for finding the correct graph
- Future work:
  - Extension to imperfect/incomplete intervention sets
  - Encoding transitivity: if $X_1 \succ X_2$ and $X_2 \succ X_3$, then $X_1 \succ X_3$

Code available at: https://github.com/phlippe/ENCO