Efficient Neural Causal Discovery without Acyclicity Constraints

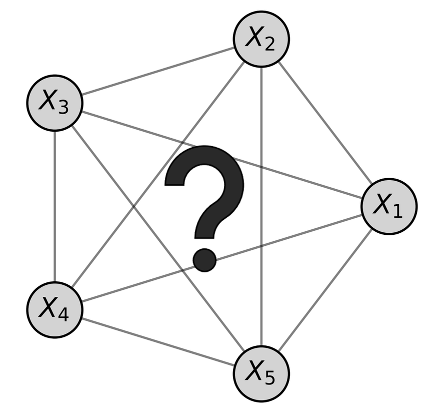
Phillip Lippe[1], Taco Cohen[2], Efstratios Gavves[1]

[1]University of Amsterdam, QUVA lab; [2]Qualcomm AI Research, The Netherlands

Paper and code

Problem statement

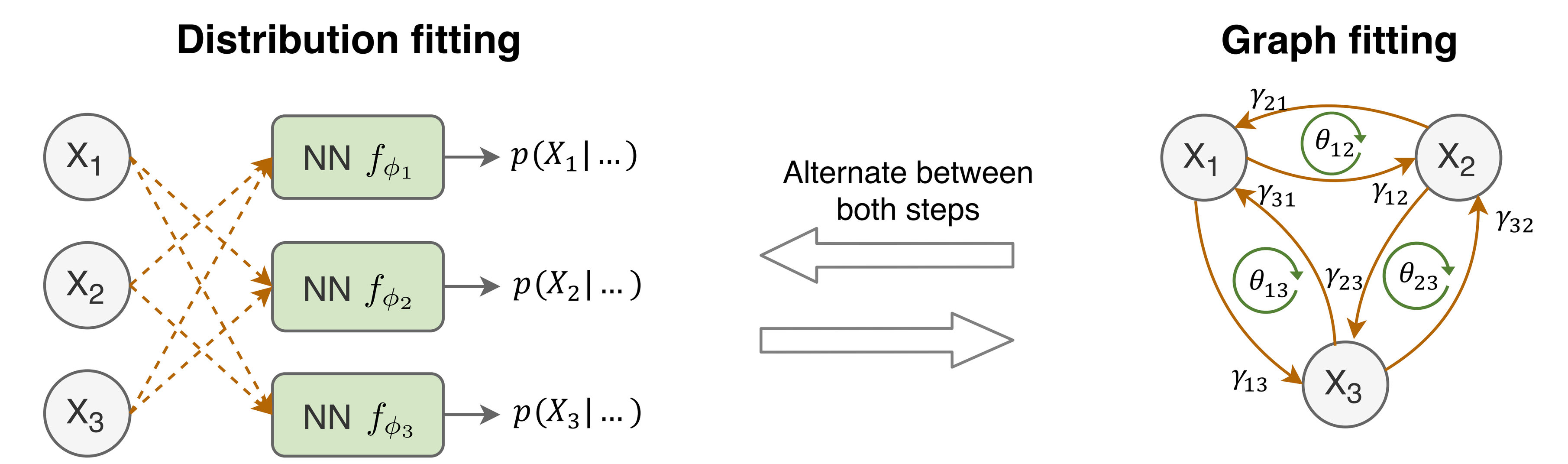
- Learn causal relations between variables as a directed, acyclic graph (DAG) from observational and interventional data
- Assumptions: interventions are sparse (only one variable at a time), soft (distribution over values), perfect (new distribution independent of original parents), and available for all variables.
- Continuous-optimization methods are promising due to their efficiency, but acyclicity needs to be ensured by constrained optimization methods which are slow and sensitive to hyperparameters, or regularizers without guarantess



ENCO – Efficient Neural Causal Discovery

- Optimize likelihoods of edges based on how well graphs generalize from observations to interventions
- Split parameters into two groups: edge existence γ and orientation θ for better control on gradients
- Probability of edges: $p(X_i \rightarrow X_j) = \sigma(\gamma_{ij}) \cdot \sigma(\theta_{ij})$
- Fit observational distributions by neural networks, and evaluate different graphs on likelihood-based score function without acyclicity constraint:

$$\tilde{\mathcal{L}} = \mathbb{E}_{\hat{I}\sim p_I(I)} \mathbb{E}_{\tilde{p}_{\hat{I}}(\mathbf{X})} \mathbb{E}_{p_{\gamma,\theta}(C)} \left[\sum_{i=1}^{N} \mathcal{L}_C(X_i)\right] + \lambda_{\text{sparse}} \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma(\gamma_{ij}) \cdot \sigma(\theta_{ij})$$



➔ Fit NNs to conditional, observational distributions

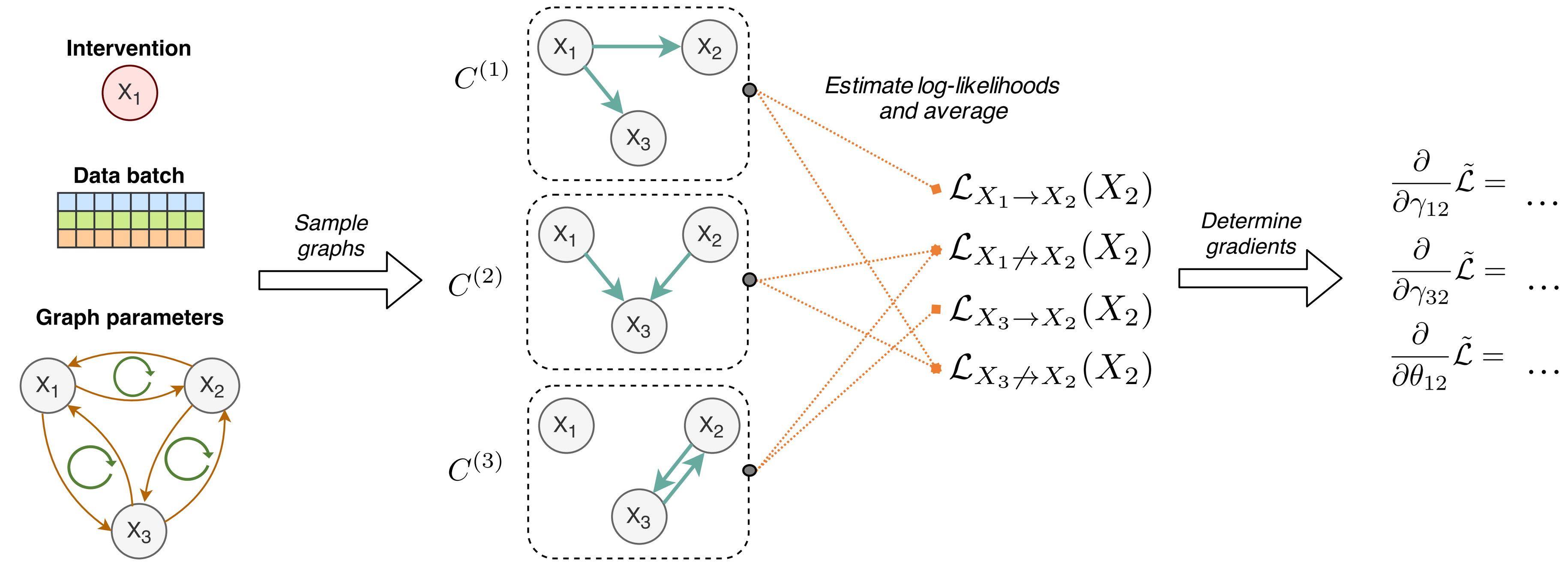
➔ Learn edge and orientation parameters based on fitted distributions

Graph optimization

- Unbiased, low-variance gradient estimator for γ and θ via REINFORCE and Monte-Carlo sampling
- Intuition: sample interventional data and K graphs, and check for each edge whether its existence improved the child's likelihood estimate

$$\frac{\partial}{\partial \gamma_{ij}} \tilde{\mathcal{L}} = \sigma'(\gamma_{ij}) \cdot \sigma(\theta_{ij}) \cdot \mathbb{E}_{\mathbf{X},C_{-ij}} \left[\mathcal{L}_{X_i \rightarrow X_j}(X_j) - \mathcal{L}_{X_i \not\rightarrow X_j}(X_j) + \lambda_{\text{sparse}}\right]$$

- Orientations only updated on interventions



Latent confounders

- Latent confounders between two observable variables without direct causal relation cause a unique pattern in the graph parameters
- An edge between the two variables is disadvantegous on interventional data but beneficial when intervening on any other variable
- Phenomenon can be detected by recording observational and interventional gradients on γ separately, and combine in a score function:

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

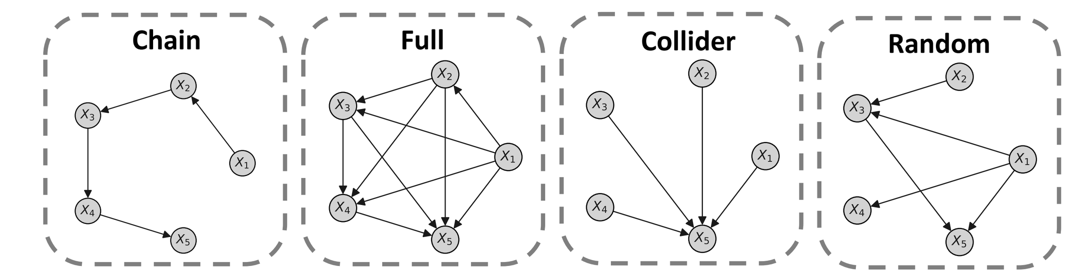
Convergence

- With interventions on all variables, we can guarantee that ENCO converges to the correct, directed acyclic causal graph despite no acyclic constraints
- Central condition: the expected improvement of the likelihood estimate by adding a correct edge must be greater than the sparsity regularizer λ_{sparse}

$$\min_{\hat{\text{pa}} \subseteq \text{gpa}_i(X_j)} \mathbb{E}_{\hat{I}\sim p_I(I)} \mathbb{E}_{\tilde{p}_{\hat{I}}(\mathbf{X})} \left[\log p(X_j|\hat{\text{pa}}, X_i) - \log p(X_j|\hat{\text{pa}})\right] > \lambda_{\text{sparse}}$$

- Tradeoff: low values of λ_{sparse} might require longer training times

Experiments



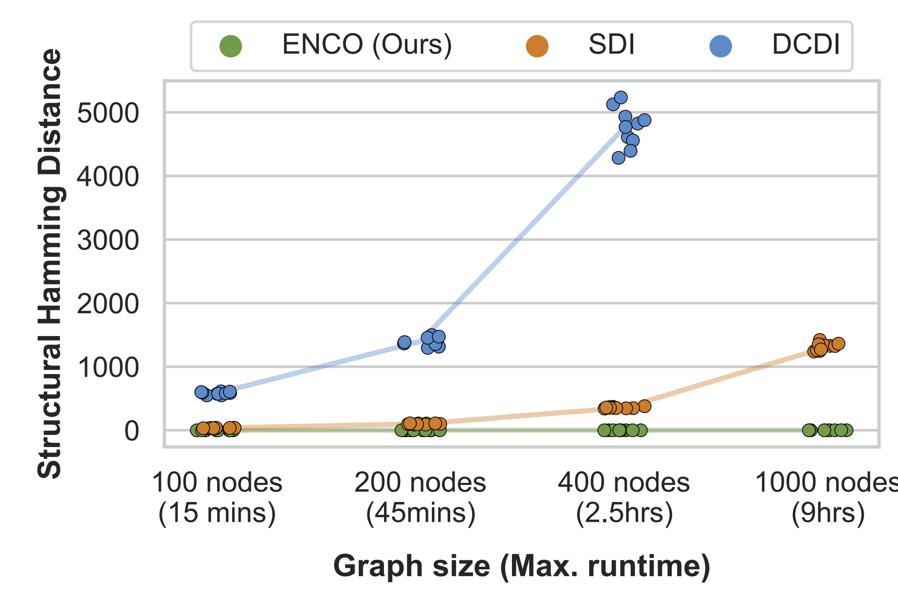
Synthetic graphs

- Synthetically generated graphs, 25 nodes
- 5k observational, 200 interventional samples
- ENCO recovers 4 out of 6 without with less than one mistake on average (SHD)

Graph type	bidiag	chain	collider	full	jungle	random
GIES [Hauser et al., 2012]	33.6 (±7.5)	17.5 (±7.3)	24.0 (±2.9)	216.5 (±15.2)	33.1 (±2.9)	57.5 (±14.2)
IGSP [Wang et al., 2017]	32.7 (±5.1)	14.6 (±2.3)	23.7 (±2.3)	253.8 (±12.6)	35.9 (±5.2)	65.4 (±8.0)
SDI [Ke et al., 2019]	9.0 (±2.6)	3.9 (±2.0)	16.1 (±2.4)	153.9 (±10.3)	6.9 (±2.3)	10.8 (±3.9)
DCDI [Brouillard et al., 2020]	16.9 (±2.0)	10.1 (±1.1)	10.9 (±3.6)	21.0 (±4.8)	17.9 (±4.1)	7.7 (±3.2)
ENCO (ours)	2.2 (±1.4)	1.7 (±1.3)	**1.6** (±1.6)	9.2 (±3.4)	1.7 (±1.3)	4.6 (±1.9)
ENCO-acyclic (ours)	**0.0** (±0.0)	**0.0** (±0.0)	**1.6** (±1.6)	**5.3** (±2.3)	**0.6** (±1.1)	**0.2** (±0.5)

Scalability

- Testing ENCO and strongest baselines on scalability to large, synthetic graphs of 100 to 1000 variables
- With same NN and hardware setting, ENCO is more efficient making only two errors overall in 10 graphs of 1000 variables



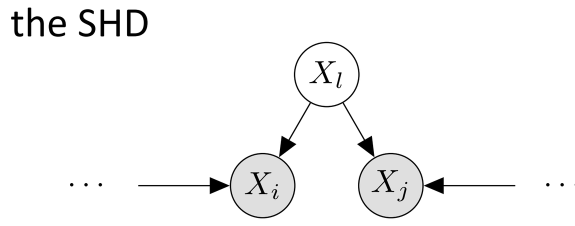
BnLearn Repository

- Real-world inspired causal graphs from the BnLearn Repository [Scutari, 2010]
- ENCO achieves perfect reconstruction for most graphs, including *diabetes* with many deterministic variables

Dataset	asia (8 nodes)	sachs (11 nodes)	child (20 nodes)	alarm (37 nodes)	diabetes (413 nodes)
SDI	4.0	7.0	11.2	24.4	422.4
ENCO (Ours)	**0.0**	**0.0**	**0.0**	**1.0**	**2.0**

Latent confounders

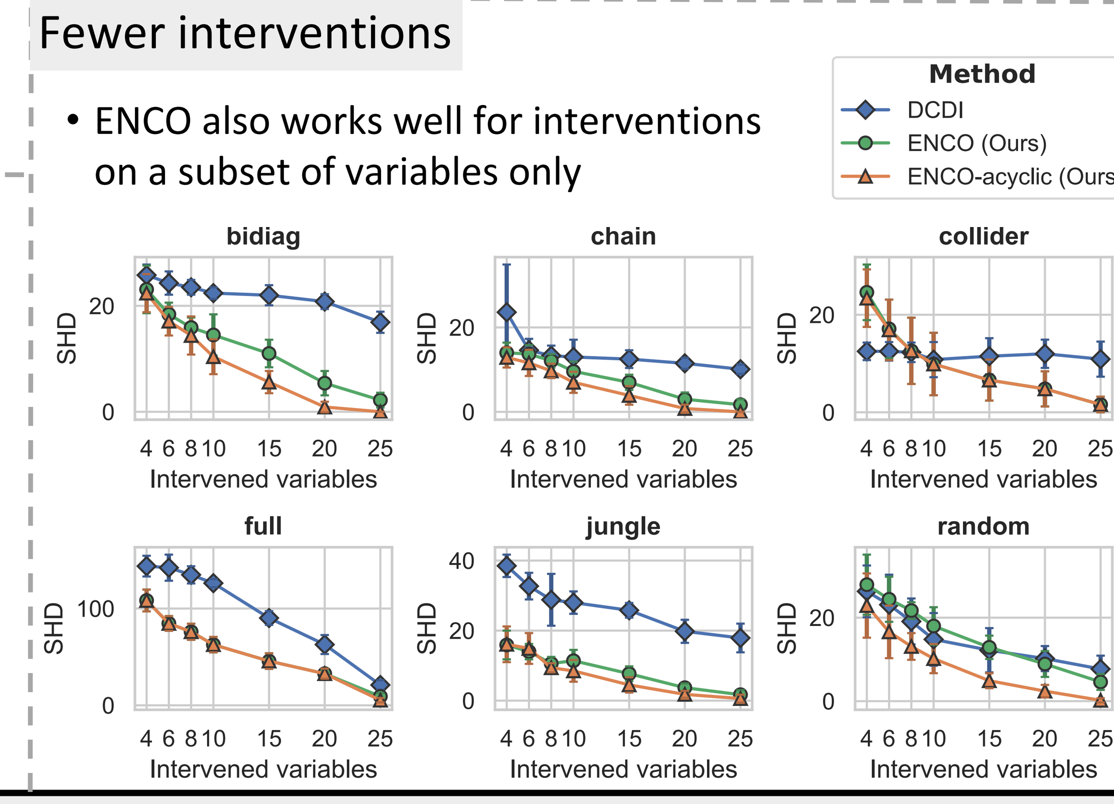
- Synthetic graphs with 25 variables plus 5 latent confounders on arbitrary node pairs except parent-child relations
- ENCO detects most confounders while missed confounders don't effect the SHD



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

Fewer interventions

- ENCO also works well for interventions on a subset of variables only



Takeaways

- Splitting graph parameters into edge existence and orientation for greater control over gradients without acyclicity constraints
- Efficient graph optimization using low-variance gradient estimators by testing generalization to interventions
- Convergence guarantees can be given when interventions on all variables are provided
- ENCO reliably and efficiently recovers causal graphs with up to 1000 variables, including latent confounders

Check out our paper and code for details!