Efficient Neural Causal Discovery without Acyclicity Constraints

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

[1]University of Amsterdam, QUVA lab, Science Park 904, Amsterdam, The Netherlands
[2]Qualcomm AI Research*, Science Park 404, Amsterdam, The Netherlands

Abstract

Learning the structure of a causal graphical model using both observational and interventional data is a fundamental problem in many scientific fields. A promising direction is continuous optimization for score-based methods, which efficiently learns the causal graph in a data-driven manner. However, to date, those methods require constrained optimization to enforce acyclicity or lack convergence guarantees. In this paper, we present ENCO, an efficient structure learning method for directed, acyclic causal graphs leveraging observational and interventional data. ENCO formulates the graph search as an optimization of independent edge likelihoods, with the edge orientation being modeled as a separate parameter. Consequently, we can provide convergence guarantees of ENCO under mild conditions without constraining the score function with respect to acyclicity. In experiments, we show that ENCO can efficiently recover graphs with hundreds of nodes, an order of magnitude larger than what was previously possible.

1 INTRODUCTION

Uncovering and understanding causal mechanisms is an important problem not only in machine learning [Bengio et al., 2019, Pearl, 2009, Schölkopf et al., 2021] but also in various scientific disciplines such as computational biology [Friedman et al., 2000, Sachs et al., 2005], epidemiology [Robins et al., 2000, Vandenbroucke et al., 2016], and economics [Hicks et al., 1980, Pearl, 2009]. A common task of interest is *causal structure learning* [Pearl, 2009, Peters et al., 2017] which aims at learning a directed acyclic graph (DAG) in which edges represent causal relations between variables. While observational data alone is in general not sufficient to identify the DAG [Hauser and Bühlmann, 2012, Yang et al., 2018], interventional data can improve identifiability up to finding the exact graph [Eberhardt, 2008, Eberhardt et al., 2005]. With recent advances in gene editing technologies providing large amounts of interventional gene expression data [Dixit et al., 2016, Macosko et al., 2015], there is a need for algorithms that can perform structure learning for graphs with several hundreds of nodes.

Finding the right DAG is challenging as the solution space grows super-exponentially with the number of variables. A promising new direction are continuous-optimization methods [Bengio et al., 2019, Brouillard et al., 2020, Ke et al., 2019, Yu et al., 2019, Zheng et al., 2018, 2020, Zhu et al., 2020] that are more computationally efficient than previous score-based and constraint-based methods [Guo et al., 2020, Peters et al., 2017] by leveraging the expressiveness of neural networks as function approximators. To restrict the search space to acyclic graphs, Zheng et al. [2018] proposed to view the search as a constrained optimization problem using an augmented Lagrangian procedure to solve it. Several follow-up works improved the process [Brouillard et al., 2020, Yu et al., 2019, Zheng et al., 2020]. An alternative approach is the usage of a regularizer in the learning objective that penalizes cyclic graphs and is simpler to optimize [Ke et al., 2019, Zhu et al., 2020]. Nonetheless, methods with such regularizers commonly lack guarantees for converging to the correct causal graph. To date, continuous optimization methods do not scale well to more than 50 variables due to a discrete search space and difficulties in enforcing acyclicity.

In this work, we show that with suitable interventional data, we do not need to limit the search space to DAGs in the first place. Instead, by modeling the orientation of an edge as a separate parameter, the optimization of a likelihood-based objective already converges to the correct, acyclic graph. The parameterization of edge orientations allows us to derive low-variance gradient estimators for the discrete search space and give convergence guarantees under mild conditions. We call this method ENCO, **E**fficient **N**eural **C**ausal

*Qualcomm AI Research in an initiative of Qualcomm Technologies, Inc.

Accepted for the 8th Causal Inference Workshop at UAI 2021.

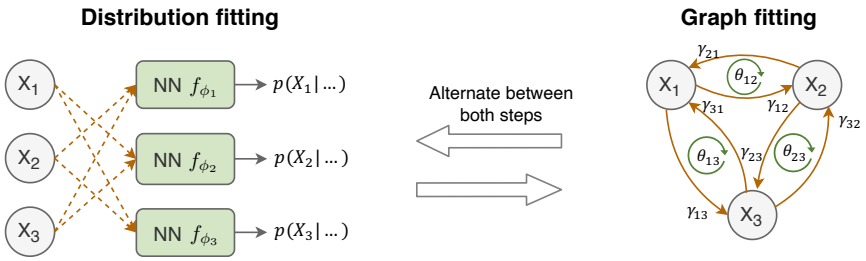

Figure 1: Visualization of the two training stages of ENCO, distribution fitting and graph fitting, on an example graph with 3 variables (X_1, X_2, X_3). The graph on right further shows how the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ correspond to edge probabilities. We learn the parameters by comparing multiple graph samples on their generalization from observational to interventional data.

Disc**o**very. In experiments, we show that ENCO handles various graph settings well, and even recovers graphs with up to 1,000 nodes in less than nine hours of compute using a single GPU (NVIDIA RTX3090) while having less than one mistake on average out of 1 million possible edges.

2 EFFICIENT NEURAL CAUSAL DISCOVERY

2.1 SCOPE AND ASSUMPTIONS

We consider the task of finding a directed acyclic graph $G = (V, E)$ of an unknown causal graphical model (CGM) given observational and interventional samples. We start by assuming that the CGM is causally sufficient, *i.e.*, all common causes of variables are included in the DAG and observable. In the following, however, we also discuss extensions for inferring the causal mechanisms in graphs with latent confounding causal variables. We emphasize that we place no constraints on the domains of the variables: they can be discrete, continuous, or mixed.

The scope with respect to the interventions closely follows Ke et al. [2019]. The interventional data is created via sparse interventions that only affect a single variable, and are retracted before the next intervention is performed. Furthermore, we consider that interventions are perfect, that is, the new variable distribution is independent of the parents. We consider that interventions have been performed for every variable and samples from it including the intervention target are provided. Last, we emphasize that we do not require specific distributions in the interventions.

2.2 LEARNING THE CAUSAL GRAPH

To learn a causal graph from observational and interventional data, ENCO models a probability for every possible directed edge between pairs of variables. The goal is to optimize these probabilities such that all edges in the ground truth graph converge to one, and all others to zero. The optimization exploits the idea of independent causal mechanisms [Pearl, 2009, Peters et al., 2016, Schölkopf et al., 2012]: we search for the graph which generalizes best from observational to interventional data. In the ground-truth causal graph, the conditional distributions for all variables stay invariant under an intervention except the intervened ones. Meanwhile, this does not hold for graphs that model the same distribution but with a flipped, missing or additional edge [Peters et al., 2016]. To implement this optimization, we alternate between two learning stages visualized in Figure 1: distribution fitting and graph fitting.

Distribution fitting trains a neural network f_{ϕ_i} per variable X_i to model its observational, conditional data distribution $p(X_i|...)$. The input to the network are all other variables, $\boldsymbol{X}_{-i}$, but we apply a dropout-like scheme to the input for simulating different sets of parents. Specifically, during training, we randomly set an input variable X_j to zero based on the probability of its corresponding edge $X_j \to X_i$. Similar techniques have been used by previous works [Brouillard et al., 2020, Ivanov et al., 2019, Ke et al., 2019, Li et al., 2020, Yoon et al., 2018]. The training can be summarized as the following optimization problem:

$$\min_{\phi_i} \mathbb{E}_{\boldsymbol{X}} \mathbb{E}_{\boldsymbol{M}} \left[-\log f_{\phi_i}(X_i; \boldsymbol{M}_{-i} \odot \boldsymbol{X}_{-i}) \right] \qquad (1)$$

where $M_j \sim \mathrm{Ber}(p(X_j \to X_i))$. If X_i is a categorical random variable, we apply a softmax as output activation function of f_{ϕ_i}. For continuous cases, a Normalizing Flow [Rezende and Mohamed, 2015] can be used for f_{ϕ_i}.

Graph fitting uses the learned networks and interventional data to score and compare different graphs. For parameterizing the edge probabilities, we use two sets of parameters: $\boldsymbol{\gamma} \in \mathbb{R}^{N \times N}$ represents the existence of edges in a graph, and $\boldsymbol{\theta} \in \mathbb{R}^{N \times N}$ the orientation of the edges. The likelihood of an edge is determined by $p(X_i \to X_j) = \sigma(\gamma_{ij}) \cdot \sigma(\theta_{ij})$, with $\sigma(...)$ being the sigmoid function and $\theta_{ij} = -\theta_{ji}$. The probability of the two orientations always sum to one. The benefit of separating the edge probabilities into two independent parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ is that it gives us more control over the gradient updates. The existence of an (undirected) edge can usually be already learned from observational or

arbitrary interventional data alone, excluding deterministic variables [Pearl, 2009]. In contrast, the orientation can only be reliably detected from data for which an intervention is performed on its adjacent nodes, *i.e.* X_i or X_j for learning θ_{ij}. While other interventions eventually provide information on the edge direction, *e.g.*, intervening on a node X_k which is a child of X_i and a parent of X_j, we do not know the relation of X_k to X_i and X_j at this stage, as we are in the process of learning the structure. Hence, only the interventions on X_i or X_j reliably uncover the orientation θ_{ij}. Despite having just one variable for the orientation, γ_{ij} and γ_{ji} are learned as two independent parameters. This is because on interventional data, an edge can improve the log-likelihood estimate in one direction, but not necessarily the other as well leading to conflicting gradients.

The objective function we use for optimizing the graph parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ is written as:

$$\begin{aligned}\tilde{\mathcal{L}} =& \mathbb{E}_{\hat{I}\sim p_I(I)}\mathbb{E}_{\tilde{p}_{\hat{I}}(\boldsymbol{X})}\mathbb{E}_{p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)}\left[\sum_{i=1}^{N}\mathcal{L}_C(X_i)\right] \\ &+ \lambda_{\text{sparse}}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma(\gamma_{ij})\cdot\sigma(\theta_{ij})\end{aligned} \tag{2}$$

where $p_I(I)$ is the distribution over which variable to intervene on (usually uniform), and $\tilde{p}_{\hat{I}}(\boldsymbol{X})$ the joint distribution of all variables under the intervention $\hat{I}$. In other words, these two distributions represent our interventional data distribution. The distribution over adjacency matrices C under $\boldsymbol{\gamma}, \boldsymbol{\theta}$ is denoted by $p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)$ with $C_{ij} \sim \text{Ber}(\sigma(\gamma_{ij})\sigma(\theta_{ij}))$, and $\mathcal{L}_C(X_i)$ is the negative log-likelihood estimate of the variable X_i conditioned on the parents according to C: $\mathcal{L}_C(X_i) = -\log f_{\phi_i}(X_i; C_{\cdot,i} \odot \boldsymbol{X}_{-i})$. The second term of Equation 2 represents a prior towards sparser graphs, removing redundant edges. It is an ℓ_1-regularizer on the edge probabilities, with the hyperparameter λ_{sparse} as regularization weight. The goal is to optimize $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ such that it minimizes the objective $\tilde{\mathcal{L}}$. For this, we need to determine their gradients through the expectation $\mathbb{E}_{p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)}$ where C is a discrete variable. For this, we apply REINFORCE [Williams, 1992] and obtain a gradient which can be estimated using Monte-Carlo sampling. Specifically, to perform an update step on $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, we sample K graph structures from $p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)$, and use the different likelihood estimates of all variables on a batch of interventional data to determine the gradients of the parameters.

2.3 CONVERGENCE

After training, we obtain a graph prediction by selecting the edges for which $\sigma(\gamma_{ij})$ and $\sigma(\theta_{ij})$ are greater than 0.5. The orientation parameters prevent loops between any two variables, since $\sigma(\theta_{ij})$ can only be greater than 0.5 in one direction. Although the orientation parameters do not guarantee the absence of loops with more variables at any stage of the training, we show that ENCO converges to the correct, acyclic graph under mild conditions. The proof for this convergence contains three steps. First, for every ancestor-descendant pair X_i, X_j, the orientation parameter θ_{ij} converges to $\sigma(\theta_{ij}) = 1$ if X_i and X_j are not conditionally independent on interventional data. Second, every edge $X_i \to X_j$ in the ground truth graph is learned if adding X_i to the any parent set of X_j improves the log-likelihood estimate by at least λ_{sparse}. Finally, all other edges will be removed by the regularizer. We outline a sketch of the proof in the appendix, and show an experimental verification next.

3 EXPERIMENTS

We evaluate ENCO on structure learning on synthetic datasets for systematic comparisons. The experiments focus on graphs with categorical variables. Categorical data is commonly more difficult in structure learning, as regression techniques or assumption on linear noise models cannot be used. Yet, ENCO is also applicable on continuous data.

We compare ENCO to GIES [Hauser and Bühlmann, 2012] and IGSP [Wang et al., 2017, Yang et al., 2018] as greedy score-based approaches, and DCDI [Brouillard et al., 2020] and SDI [Ke et al., 2019] as continuous optimization methods. Pure constraint-based methods do not scale well to the given graph and dataset sizes [Guo et al., 2020, Peters et al., 2017]. We do not compare to methods with observational data only, since those can just recover the graph up to its Markov equivalence class. We perform a separate hyperparameter search for all methods. Since SDI and DCDI use neural networks to fit (observational) distributions as well, we use the same network setup as for ENCO. All methods were executed on the same hardware, namely a 12-core CPU with a single NVIDIA RTX3090 GPU. Our code is publicly available at `https://github.com/phlippe/ENCO`.

3.1 COMMON GRAPH STRUCTURES

We first experiment on synthetic graphs, for which we pick six common graph structures. The graphs `chain` and `full` represent the minimally- and maximally-connected DAGs. `bidiag` is a chain with 2-hop connections, and `jungle` is a tree-like graph. In the `collider` graph, one node has all other nodes as parents. Finally, `random` has a randomly sampled graph structure with a likelihood of 0.3 of two nodes being connected by a direct edge. For each graph structure, we generate 25 graphs with 25 nodes each. The graph generation process follows the setup of Ke et al. [2019]. Following common practice, we use structural hamming distance (SHD) as evaluation metric. SHD counts the number of edges that need to be removed, added, or flipped in order to obtain the ground truth graph.

We report the average performance and standard deviation

Table 1: Comparing structure learning methods in terms of structural hamming distance (SHD) on common graph structures (lower is better), averaged over 25 graphs each. In line with the theoretical guarantees, ENCO can reliably recover five out of the six graph structures without errors.

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)

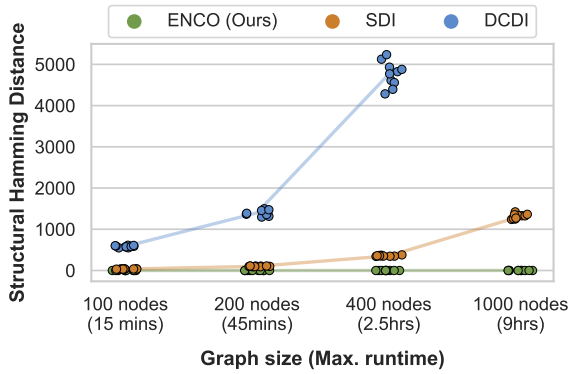

Figure 2: Evaluating SDI, DCDI, and ENCO on large graphs in terms of SHD (lower is better). Dots represent single experiments, lines connect the averages. DCDI ran out of memory for 1000 nodes.

in Table 1. Overall, the continuous optimization methods considerably outperform the greedy search approaches. SDI works reasonably well on sparse graphs, but struggles with nodes that have many parents. The second best is DCDI which performs well on the `collider` graph since its edges can be independently orientated. Although DCDI converges to acyclic graphs, it predicts some incorrectly oriented edges, while being 8 times slower than ENCO on the same hardware. ENCO reliably learns five out of six graph structures without errors, except of rare mistakes on the `full` graph. Therefore, the theoretical guarantees also hold in practice for small graphs.

3.2 SCALABILITY

Next, we test ENCO on graphs with large sets of variables. We create `random` graphs ranging from $N = 100$ to $N = 1,000$ nodes. Every node has on average 8 in- or outgoing edges and a maximum of 10 parents. The challenge of large graphs is that the number of possible edges grows quadractically and the number of DAGs super-exponentially. Hence, efficient methods are needed.

We compare ENCO to the two best performing baselines from Table 1, SDI [Ke et al., 2019] and DCDI [Brouillard et al., 2020]. All methods were given the same setup of neural networks and a maximum runtime which corresponds to 30 epochs for ENCO. We plot the SHD over graph size and runtime in Figure 2. ENCO is capable of recovering the causal graphs perfectly with no errors except for the $1,000$-node graph, for which it misses one out of 1 million edges in 2 out of 10 experiments. SDI and DCDI achieve considerably worse performance. This shows that ENCO can efficiently be applied to $1,000$ variables while maintaining its convergence guarantees. Similar results have also been observed on real-world inspired graphs from the Bayesian Network Repository [Scutari, 2010] including the graphs `diabetes` (413 nodes) and `pigs` (441 nodes).

4 CONCLUSION

In this work, we propose ENCO, an efficient structure learning method leveraging observational and interventional data. ENCO models a graph by independent edge likelihoods with the edge orientation as a separate parameter. As such, its objective is unconstrained with respect to acyclicity while providing convergence guarantees. In experiments, we show that ENCO can be efficiently applied to graphs comprising hundreds of nodes with a very high accuracy.

Aspects that have not been detailed in this extended abstract include the low-variance gradient estimators used for γ and $\boldsymbol{\theta}$. Compared to related work, this estimator has a ten times lower standard deviation, which is crucial for learning large graphs. Further, ENCO can be extended to handle latent confounders which cause unique patterns in the γ-gradients.

Limitations of ENCO include the need for interventional data on all variables. Future work includes investigating the generalization of ENCO to incomplete intervention sets. For instance, despite the absence of interventions, one can still recover undirected edges via γ. A second limitation is that the orientations are missing transitivity: if $X_1 \succ X_2$ and $X_2 \succ X_3$, then $X_1 \succ X_3$ must also be true. Global order distributions such as Plackett-Luce [Luce, 1959, Plackett, 1975] require high variance gradient estimators and struggled with chains in early experiments. That said, a potential direction is to experiment with transitive relations for improving convergence speed.

References

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f8b7aa3a0d349d9562b424160ad18612-Paper.pdf.

A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7): 1853–1866, Dec 2016.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 161–168, Arlington, Virginia, USA, 2008. AUAI Press. ISBN 0974903949.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations among N Variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 178–184, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.*, 53(4), July 2020. ISSN 0360-0300. doi: 10.1145/3397269. URL https://doi.org/10.1145/3397269.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(1):2409–2464, August 2012. ISSN 1532-4435.

John Hicks et al. *Causality in economics*. Australian National University Press, 1980.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxtJh0qYm.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Yang Li, Shoaib Akbar, and Junier Oliva. ACFlow: Flow models for arbitrary conditional likelihoods. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5831–5841. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/li20a.html.

R Duncan Luce. *Individual choice behavior*. John Wiley, Oxford, England, 1959.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346567.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul

2015. PMLR. URL http://proceedings.mlr.press/v37/rezende15.html.

J.M. Robins, M.A. Hernan, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 459–466, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards Causal Representation Learning. *arXiv preprint arXiv:2102.11107*, 2021. URL http://arxiv.org/abs/2102.11107.

Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Jan P Vandenbroucke, Alex Broadbent, and Neil Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International journal of epidemiology*, 45(6):1776–1786, 2016.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/275d7fb2fd45098ad5c3ece2ed4a2824-Paper.pdf.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pages 229–256, 1992.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5541–5550. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/yang18a.html.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/yoon18a.html.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/yu19a.html.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9492–9503, Red Hook, NY, USA, 2018. Curran Associates Inc.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/zheng20a.html.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g2skStPB.

A CONVERGENCE GUARANTEES

In this section we want to give an overview of the conditions under which ENCO is guaranteed to converge to the true, acyclic causal graph given sufficient data and time. We first discuss the assumptions we take for providing the guarantees. Next, we discuss the gradient estimators that are used for optimizing the graph parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ in the graph fitting stage. The gradient estimators are essential to the convergence guarantees as they are unbiased with respect to the objective $\mathcal{L}$ in Equation 2 and can be intuitively explained. Finally, we give a sketch of the proof under which conditions ENCO converges to the correct graph.

A.1 ASSUMPTIONS

Assumption 1 A common assumption in causal structure learning is that the data distribution over all variables $p(\boldsymbol{X})$ is Markovian and faithful with respect to the causal graph we are trying to model. This means that the graph represents the (conditional) independences relations between variables in the data, and (conditional) independence relations in the data reflect the edges in the graph. For ENCO, faithfulness is not strictly required. This is because we work with interventional data. Instead, we rely on the Markov property and assume that for all variables, the parent set $\text{pa}(X_i)$ reflects the inputs to the causal generation mechanism of X_i. This allows us to also handle deterministic variables.

Assumption 2 For this proof, we assume that all variables of the graph are known and observable, and no latent confounders exist. Latent confounders can introduce dependencies between variables which are not reflected by the ground truth graph solely on the observed variables.

Assumption 3 ENCO relies on neural networks to determine the conditional data distributions $p(X_i|...)$. Hence, for providing a guarantee, we assume that in the graph learning step the neural networks have been sufficiently trained such that they accurately model all possible conditional distribution $p(X_i|...)$. In practice, the neural networks might have a slight error. However, as long as enough data, network complexity, and training time is provided, it is fair to assume that the difference between the modeled distribution and the true conditional is smaller than an arbitrary constant ϵ.

A.2 GRADIENT ESTIMATORS

To update $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ based on the objective in Equation 2, we need to determine their gradients through the expectation $\mathbb{E}_{p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)}$ where C is a discrete variable. For this, we apply REINFORCE [Williams, 1992]. We limit our discussion here to the results as those are needed for the convergence proof.

The gradient for the parameter γ_{ij} is:

$$\frac{\partial}{\partial \gamma_{ij}} \tilde{\mathcal{L}} = \sigma'(\gamma_{ij}) \cdot \sigma(\theta_{ij}) \cdot \mathbb{E}_{\boldsymbol{X}, C_{-ij}} \big[\mathcal{L}_{X_i \to X_j}(X_j) - \mathcal{L}_{X_i \not\to X_j}(X_j) + \lambda_{\text{sparse}} \big] \tag{3}$$

where $\mathbb{E}_{\boldsymbol{X}, C_{-ij}}$ summarizes for brevity the three expectations in Equation 2 up to the edge $X_i \to X_j$. This excludes interventions on X_j since we assume the interventions to be perfect, *i.e.* intervened variables are independent of the other variables. Further, $\mathcal{L}_{X_i \to X_j}(X_j)$ denotes the negative log likelihood for X_j under the adjacency matrix C_{-ij} including an edge from X_i to X_j, *i.e.* $C_{ij} = 1$. The gradient in Equation 3 can be intuitively explained: if by the addition of the edge $X_i \to X_j$, the log-likelihood estimate of X_j is improved by more than λ_{sparse}, we increase the corresponding edge parameter γ_{ij}; otherwise, we decrease it.

The orientation parameters $\boldsymbol{\theta}$ are similarly derived as $\boldsymbol{\gamma}$. As mentioned before, we only want to take the gradients for θ_{ij} when the intervention is performed on X_i or X_j. This leads us to:

$$\frac{\partial}{\partial \theta_{ij}} \tilde{\mathcal{L}} = \sigma'(\theta_{ij}) \cdot \Big(p(I_{X_i}) \cdot \sigma(\gamma_{ij}) \cdot T(X_i, X_j) - p(I_{X_j}) \cdot \sigma(\gamma_{ji}) \cdot T(X_j, X_i) \Big) \tag{4}$$

with

$$T(X_k, X_l) = \mathbb{E}_{I_{X_k}, \boldsymbol{X}, C_{-kl}} \left[\mathcal{L}_{X_k \to X_l}(X_l) - \mathcal{L}_{X_k \not\to X_l}(X_l) \right] \tag{5}$$

The probability of taking an intervention on X_i is represented by $p(I_{X_i})$ (usually uniform across variables). When the edge $X_i \to X_j$ improves the log-likelihood of X_j under intervention on X_i, then the gradient increases θ_{ij}. In contrast, when the true edge is $X_j \to X_i$, the correlation between X_i and X_j learned from observational data would yield a worse likelihood estimate on interventional data than without the edge $X_j \to X_i$. This is because $p(X_j|X_i, ...)$ does not stay invariant under intervening on X_i. Lastly, for independent nodes, the expectation of the gradient is zero.

A.3 SKETCH OF PROOF

The proof consists of the following three main steps:

Step 1 We show under which conditions the orientation parameters θ_{ij} will converge to $+\infty$, i.e. $\sigma(\theta_{ij}) \to 1$, if X_i is an ancestor of X_j. Similarly, if X_j is an descendant of X_i, the parameter θ_{ij} will converge to $-\infty$, i.e. $\sigma(\theta_{ij}) \to 0$.

Step 2 Under the assumption that the orientation parameters have converged as in Step 1, we show that for edges in the true graph, γ_{ij} will converge to 1. Note that we need to take additional assumptions/conditions with respect to λ_{sparse} here.

Step 3 Once the parameters γ_{ij} and θ_{ij} have converged for the edges in the ground truth graph, we show that all other edges will be removed by the sparsity regularizer.

The following paragraphs provide more details for each step. Note that causal graphs that do not fulfill all parts of the convergence guarantee can still eventually be recovered. The reason is that the conditions listed in the theorems below ensure that there exists no local minima for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ to converge in. Although local minima exist, the optimization process might converge to the global minimum of the true causal graph.

Theorem A.1. *Consider the edge $X_i \rightarrow X_j$ in the true causal graph. The orientation parameter θ_{ij} converges to $\sigma(\theta_{ij}) = 1$ if the following two conditions are fulfilled:*

(1) for all possible sets of parents of X_j excluding X_i, adding X_i improves the log-likelihood estimate of X_j under the intervention on X_i, or leaves it unchanged:

$$\forall \widehat{pa}(X_j) \subseteq X_{-i,j} : \\ \mathbb{E}_{I_{X_i},\boldsymbol{X}} \left[\log p(X_j|\widehat{pa}(X_j), X_i) - \log p(X_j|\widehat{pa}(X_j))\right] \geq 0 \tag{6}$$

(2) there exists a set of nodes $\widehat{pa}(X_j)$, for which the probability to be sampled as parents of X_j is greater than 0, and the following condition holds:

$$\exists \widehat{pa}(X_j) \subseteq X_{-i,j} : \\ \mathbb{E}_{I_{X_i},\boldsymbol{X}} \left[\log p(X_j|\widehat{pa}(X_j), X_i) - \log p(X_j|\widehat{pa}(X_j))\right] > 0 \tag{7}$$

Proof. Based on the conditions in Equations 6 and 7, we need to show that the gradient of θ_{ij} is negative in expectation, independent of other values of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. Looking at the gradient of θ_{ij} in Equation 4, the conditions correspond to $T(X_i, X_j)$ being smaller or equals to zero. If $T(X_i, X_j)$ is smaller than zero, the gradient of θ_{ij} with respect to the interventions on X_i is negative, *i.e.* increasing θ_{ij}. To guarantee that the whole gradient of θ_{ij} is negative, we also need to show that for interventions on X_j, we obtain $T(X_j, X_i)$ being positive. When intervening on X_j, X_i and X_j become independent as the edge $X_i \rightarrow X_j$ is removed in the intervened graph. Therefore, a distribution $p(X_i|X_j, ...)$ relying on correlations between X_i and X_j from observational data cannot achieve a better estimate than the same distribution when removing X_j. The only situation where X_i and X_j can become conditionally dependent under interventions on X_j is if X_i and X_j share a collider X_k, and X_i is being conditioned on the collider X_k and X_j. However, this requires that θ_{ki} has negative gradients, *i.e.* θ_{ki} increasing, when intervening on X_k. This cannot be the case since under interventions on X_k, X_i and X_k become conditionally independent, and the correlations learned from observational data cannot be transferred to the interventional setting. If X_k and X_i again share a collider, we can apply this argumentation recursively until a node X_n does not share a collider with X_i. The recursion will always come to an end as we have a finite set of nodes, and the causal graph being acyclic. □

The conditions in Theorem A.1 are commonly fulfilled by most causal structures. However, there are some situations where the conditions can fail. One example are structures with three variables X_1, X_2, X_3 where we have the causal edges $X_1 \rightarrow X_2$, $X_1 \rightarrow X_3$, $X_2 \rightarrow X_3$. If knowing X_2 informs the log-likelihood estimate of X_3 more about X_1 than about X_2 itself, an intervention on X_2 could lead to positive gradients and violate the condition in Equation 6. Nonetheless, we did not observe any of these situations in the synthetic and real-world graphs we experimented on. Furthermore, many such graphs can still be learned when $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are initialized with zeros.

Theorem A.2. *Consider a pair of variables X_i, X_j for which X_i is an ancestor of X_j without direct edge in the true causal graph. Then, the orientation parameter of the edge $X_j \rightarrow X_i$ converges to $\sigma(\theta_{ij}) = 1$ if condition 1 of Theorem A.1 holds for the pair of X_i, X_j.*

Proof. To show this theorem, we need to consider two cases for a pair of variables X_i and X_j: X_i and X_j are conditionally independent under a sampled adjacency matrix, or X_i and X_j are not independent. Both cases need to be considered for an intervention on X_i, and an intervention on X_j.

First, we discuss interventions on X_i. If under the sampled adjacency matrix X_j is conditionally independent of X_i, the difference in the log-likelihood estimates $T(X_i, X_j)$ is zero. The variables can be independent if, for example, the parents of X_j are all parents of the true causal graph. If X_j is not conditionally independent of X_i, condition 1 from Theorem A.1 ensures that X_i only has a positive effect on the log-likelihood estimate. Thus, under interventions on X_i, the gradient of θ_{ij} must be smaller or equals to zero in expectation, *i.e.*, increases θ_{ij}.

Next, we consider interventions on X_j. If under the sampled adjacency matrix X_i is conditionally independent of X_j, the difference in the log-likelihood estimates $T(X_j, X_i)$ is zero. The variables can be independent if X_i is conditioned on variables that d-separate X_i and X_j in the true causal graph. For instance, having the children of X_i as parents of X_i creates this scenario. However, for this scenario to take place, one or more orientation parameters of parent-child or ancestor-descendant pairs must be incorrectly converged. In case of a parent-child pair X_i, X_k, Theorem A.1 shows that $\sigma(\theta_{ik})$ will converge to one removing any possibility of a reversed edge to be sampled. In case of an ancestor-descendant pair X_i, X_l, we can apply a recursive argument: as X_l d-separates X_i and X_j, X_l must come before X_j in

the causal order. If for the gradient θ_{il}, we have a similar scenario with X_i being conditionally independent of X_j, the same argument applies. This can be recursively applied until no more variables except direct children of X_i can d-separate X_i and X_j. In that case, $\sigma(\theta_{ik})$ will converge to one, which leads to all other orientation parameters to converge to one as well. If X_i is not conditionally independent of X_j, we can rely back on the argumentation of Theorem 1 when we have an edge $X_i \to X_j$: as in the intervened causal graph, X_i and X_j are independent, any correlation learned from observational data can only lead to a worse log-likelihood estimate. In cases of colliders, we can rely on the recursive argument from before. Thus, under interventions on X_j, the gradient of θ_{ij} must be smaller or equals to zero in expectation, *i.e.*, increases θ_{ij}.

Therefore, we can conclude that $\sigma(\theta_{ij})$ converges to one for any ancestor-descendant pairs X_i, X_j. □

Theorem A.3. *Consider an edge $X_i \to X_j$ in the true causal graph. The parameter γ_{ij} converges to $\sigma(\gamma_{ij}) = 1$ if the following condition holds:*

$$\min_{\hat{pa} \subseteq gpa_i(X_j)} \mathbb{E}_{\hat{I} \sim p_I(I)} \mathbb{E}_{\tilde{p}_{\hat{I}}(\boldsymbol{X})} \big[\log p(X_j | \hat{pa}, X_i) - \log p(X_j | \hat{pa}) \big] > \lambda_{sparse} \tag{8}$$

where $gpa_i(X_j)$ is the set of nodes excluding X_i which, according to the ground truth graph, could have an edge to X_j without introducing a cycle.

Proof. To show this convergence, we assume that the orientation parameters have converged corresponding to Theorem A.1 and A.2. The parameter γ_{ij} converges to $\sigma(\gamma_{ij}) = 1$ if its gradient, $\frac{\partial}{\partial \gamma_{ij}} \tilde{\mathcal{L}}$, is negative independent of other values of $\boldsymbol{\gamma}$ and orientation parameters $\boldsymbol{\theta}$ that are not included in Theorem 1 and 2. The gradient of γ_{ij} includes an expectation over adjacency matrices $p_{\boldsymbol{\gamma},\boldsymbol{\theta}}(C)$. Based on the converged $\boldsymbol{\theta}$-values, we only need to consider sets of nodes as parents for X_j that contain parents, ancestors, or independent nodes according to the ground truth graph. Among those remaining parent sets, we need to ensure that for any such set, the gradient is negative. This is guaranteed by the condition in Equation 8 since the inequality corresponds to $\frac{\partial}{\partial \gamma_{ij}} \tilde{\mathcal{L}} < 0$. If the condition holds for the parent set with the $\hat{pa}$), *i.e.* the maximum gradient, then the gradient of γ_{ij} can be guaranteed to be negative in expectation, independent of the other values of $\boldsymbol{\gamma}$. □

Theorem A.4. *If for all edges $X_i \to X_j$ in the true causal graph, $\sigma(\theta_{ij})$ and $\sigma(\gamma_{ij})$ have converged to one, the likelihood of all other edges, i.e. $\sigma(\theta_{lk}) \cdot \sigma(\theta_{lk})$, will converge to zero.*

Proof. If all edges in the ground truth graph have converged, all other pairs of variables X_l, X_k are (conditionally) independent in the graph. Hence, the difference of the log-likelihood estimate in the gradient of γ_{lk}, *i.e.* $\mathcal{L}_{X_l \to X_k}(X_k) - \mathcal{L}_{X_l \not\to X_k}(X_k)$, is zero in expectation. Thus, the gradient remaining is:

$$\frac{\partial}{\partial \gamma_{lk}} \tilde{\mathcal{L}} = \sigma'(\gamma_{lk}) \cdot \sigma(\theta_{lk}) \cdot \lambda_{\text{sparse}} \tag{9}$$

Since the gradient is positive independent of the values of γ_{lk} and θ_{lk}, γ_{lk} will decrease until it converges to $\sigma(\gamma_{lk}) = 0$. This discussion excludes the situation when X_l is a child and descendant of X_k. However, as discussed in Theorem 1 and 2, the orientation parameters θ_{lk} converge to $\sigma(\theta_{lk}) = 0$ setting those edge likelihoods to zero. Hence, if γ_{lk} is decreasing for all pairs of (conditionally) independent variables X_l, X_k in the ground truth graph, and $\sigma(\theta_{lk})$ converged to zero for children and descendants, the product $\sigma(\gamma_{lk}) \cdot \sigma(\theta_{lk})$ will converge to zero for all edges not existing in the ground truth graph. □

For graphs that fulfill all conditions in the Theorems A.1 to A.4, ENCO is guaranteed to converge given sufficient data and time. The conditions in the theorems ensure that there exist no local minima or saddle points in the loss surface of the objective in Equation 2 with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. For other causal graphs, we might still converge to the correct DAG but cannot guarantee it due to local minima.