GRADIENT-BASED NEURAL DAG LEARNING WITH INTERVENTIONS

Philippe Brouillard[1], **Alexandre Drouin**[2], **Sébastien Lachapelle**[1],
Alexandre Lacoste[2], **Simon Lacoste-Julien** [1,3]
[1]Mila, Université de Montréal; [2]Element AI; [3]Canada CIFAR AI Chair

ABSTRACT

Decision making based on statistical association alone can be a dangerous endeavor due to non-causal associations. Ideally, one would rely on causal relationships that enable reasoning about the effect of interventions. Several methods have been proposed to discover such relationships from observational and interventional data. Among them, GraN-DAG, a method that relies on the constrained optimization of neural networks, was shown to produce state-of-the-art results among algorithms relying purely on observational data. However, it is limited to observational data and cannot make use of interventions. In this work, we extend GraN-DAG to support interventional data and show that this improves its ability to infer causal structures.

1 INTRODUCTION

Causal inference from observation is a fundamental problem in science with applications in fields such as genomics, economics, and policy making (Koller & Friedman, 2009). The goal is to uncover causal relationships among observed variables. Knowledge of such relationships is crucial to decision making, since it allows reasoning about the effect of interventions. That is, answering questions such as "What would be the effect of acting to change the value of this variable?"

Several prior works have focused on learning graphical representations of causal relationships from observational data alone (Bühlmann et al., 2014; Shimizu, 2014; Hauser & Bühlmann, 2012; Spirtes et al., 2000). Others have shown that observing the effect of some interventions could improve the identification of causal relationships (Eberhardt, 2008; Eberhardt et al., 2005; Yang et al., 2018). Many of these methods are based on enumerating and scoring candidate graphs, resulting in a time-consuming combinatorial search. Zheng et al. (2018) recently showed that this search could be replaced by a continuous optimization problem. Building on this idea, Lachapelle et al. (2019) proposed GraN-DAG, a method that relies on neural networks to identify causal graphs. While this method shows state-of-the-art performance for this task, it is unable to make use of information about interventions.

In this work, we extend the GraN-DAG method to support interventions with known targets. We start by proposing a slightly modified loss function adapted to the interventional setting. We then evaluate this method on real and simulated data sets. Our findings indicate that considering interventions does improve the accuracy of GraN-DAG and that it compares favorably to state-of-the-art methods leveraging interventional data.

2 BACKGROUND

Causal models: A causal graphical model (CGM) is defined by a distribution P_X over a random vector $X = (X_1, \dots, X_d)$ and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $i \in \mathcal{V}$ is associated to a corresponding random variable X_i and each edge $(i, j) \in \mathcal{E}$ indicates a causal influence of X_i on X_j. We use the notation X_S with $S \subseteq \mathcal{V}$ to refer to the random vector $(X_i)_{i \in S}$ and x_S to refer to a specific value in the support of X_S. The distribution P_X of a CGM is Markovian to its graph $\mathcal{G}$, which means that the density $p(x_1, \dots, x_d)$ can be factorized as $\prod_{j=1}^{d} p_j(x_j | x_{\pi_j^{\mathcal{G}}})$ where $\pi_j^{\mathcal{G}}$ is the set of parents of j in graph $\mathcal{G}$.

Interventions: We consider *stochastic interventions* (Korb et al., 2004) to model the effect of intervening on a set of variables $I \subseteq \mathcal{V}$. In such interventions, the conditional density $p_j(x_j|x_{\pi_j^\mathcal{G}})$ is replaced by a new marginal $\tilde{p}_j(x_j)$ in the joint density for all $j \in I$. Formally, given the *intervention target* I, the *interventional joint density* is defined as

$$p(x_1, ..., x_d|do(X_I)) \triangleq \prod_{j \notin I} p_j(x_j|x_{\pi_j^\mathcal{G}}) \prod_{j \in I} \tilde{p}_j(x_j)$$

This formulation encapsulates the idea that each *mechanism* giving rise to a variable given its parents can be manipulated by an external agent without affecting the other mechanisms (Peters et al., 2017).

Causal learning: The causal structure learning problem consists of inferring the underlying graph $\mathcal{G}$ given samples from P_X. There are mainly two types of methods: score-based and constraint-based (Heinze-Deml et al., 2018). In the score-based setting, a score $\mathcal{S}(\mathcal{D}, \mathcal{G})$ is used to evaluate how well the data $\mathcal{D}$ can be fitted using a graph $\mathcal{G}$. The optimization is performed over the set of all DAGs in order to find the graph with the highest score, $\hat{\mathcal{G}} \triangleq \arg\max_{\mathcal{G} \in \text{DAG}} \mathcal{S}(\mathcal{D}, \mathcal{G})$. Since the size of this set is super-exponential in the number of nodes (Chickering, 2003), many score-based methods rely on greedy heuristic search algorithms. Recently, Zheng et al. (2018) showed that this combinatorial search could be replaced by a constrained continuous optimization problem, opening the door to differentiable algorithms.

GraN-DAG: Lachapelle et al. (2019) introduced *Gradient-Based Neural DAG Learning* (GraN-DAG), a score-based method that extends the continuous constrained optimization framework of Zheng et al. (2018) to allow for nonlinear relationships. Each conditional distribution $p_j(x_j|x_{-j};\phi_j)$, where x_{-j} denotes all variables except x_j, is learned by a neural network parametrized by ϕ_j. It receives X_{-j} as input and outputs a parameter θ_j of a parametric distribution for variable X_j. To make sure $\prod_{j=1}^d p_j(x_j|x_{-j};\phi_j)$ is a valid joint density function, GraN-DAG is optimized under an acyclicity constraint adapted from Zheng et al. (2018). The key idea is to construct a *weighted adjacency matrix* A_ϕ which depends on $\phi \triangleq \{\phi_1, \dots, \phi_d\}$, i.e. all the weights of all neural networks. Intuitively, $(A_\phi)_{ij} \geq 0$ quantifies the strength of the edge $i \to j$. Moreover, $(A_\phi)_{ij} = 0$ implies the edge $i \to j$ is absent (see Lachapelle et al. (2019) for more details). The parameters of the model are learned by approximately solving the following constrained problem:

$$\max_\phi \mathbb{E}_{X \sim P_X} \sum_{j=1}^d \log p_j(X_j|X_{-j};\phi_j) \quad \text{s.t.} \quad \operatorname{Tr} e^{A_\phi} = d \tag{1}$$

Since the objective of (1) is a valid log-likelihood function whenever the constraint is satisfied, its solution corresponds to the maximum likelihood estimator. In practice, this objective is optimized by an augmented Lagrangian approach as in Zheng et al. (2018), where each sub-problem is approximately solved by stochastic gradient descent. See the original paper for details regarding thresholding and graph pruning.

3 GraN-DAG with interventions

In its original formulation, GraN-DAG can learn the underlying causal graph given the right assumptions and may thus be used to answer queries about interventions. However, it does not support interventional data as input. Interventional data are beneficial for identifiability since they reduce the size of the class of graphs that could have generated P_X (Yang et al., 2018; Hauser & Bühlmann, 2012). Here, we extend GraN-DAG to support interventional data with known targets.

Interventional setting: The learner receives a dataset of n observations of the form $\{(X^{(1)}, I^{(1)}), ..., (X^{(n)}, I^{(n)})\}$ where $I^{(i)}$ is the interventional target associated to observation $X^{(i)}$. Recall that an interventional target I is a subset of $\mathcal{V}$, i.e. the nodes targeted by the intervention. The data generation process is assumed to be the following:

$$\begin{aligned} I^{(i)} &\sim P(I) \text{ i.i.d. } \forall i \\ X^{(i)}|I^{(i)} &\sim P(X|I = I^{(i)}) \triangleq p(x_1, ..., x_d|do(X_{I^{(i)}})) \ \forall i \end{aligned} \tag{2}$$

where $P(I)$ is a distribution over a collection of interventional targets, denoted by $\mathcal{I}$.

Optimization problem: Conceptually, it is useful to think of the CGM we are learning as a family of models of the form $\{\prod_{j\notin I} p_j(x_j|x_{\pi_j^{\mathcal{G}}};\phi_j)\prod_{j\in I}\tilde{p}_j(x_j;\omega_j^I)|I\in\mathcal{I}\}$ where we introduced the parameter $\omega^I \triangleq \{\omega_j^I\}_{j\in I}$ for each $I\in\mathcal{I}$ to model the distribution of the variables on which we are intervening. This interpretation together with the acyclicity constraint of GraN-DAG suggests the following maximum log-likelihood program:

$$\max_{\phi,\{\omega^I\}_{I\in\mathcal{I}}} \mathbb{E}_{(X,I)\sim P(X,I)}\left[\sum_{j\notin I}\log p_j(X_j|X_{-j};\phi_j)+\sum_{j\in I}\log p_j(X_j;\omega_j^I)\right] \quad \text{s.t.} \quad \operatorname{Tr} e^{A_\phi}=d \tag{3}$$

In principle, the parameters ω^I could be learned, but this would not contribute to learning the ground truth graph. Since (3) can be trivially decomposed in a sum of a $\max$ over ϕ and a $\max$ over $\{\omega^I\}_{I\in\mathcal{I}}$, it suffices to solve

$$\max_{\phi} \mathbb{E}_{(X,I)\sim P(X,I)}\sum_{j\notin I}\log p(X_j|X_{-j};\phi_j) \quad \text{s.t.} \quad \operatorname{Tr} e^{A_\phi}=d. \tag{4}$$

Again, we solve it using augmented Lagrangian together with stochastic gradient descent. Of note, other score-based methods (Hauser & Bühlmann, 2012) use a similar form of objective to deal with interventions with known targets. Note that we use similar thresholding and pruning strategies as suggested by Lachapelle et al. (2019) (see Appendix A.3).

3.1 Experiments

We compare our method to the *greedy interventional equivalence search* (GIES) method (Hauser & Bühlmann, 2012) and a modified version of the *causal additive model* (CAM) method (Bühlmann et al., 2014). GIES is an extension of GES (Chickering, 2003) that was designed for interventions with known targets. GIES assumes a linear model with Gaussian noise and greedily searches the space of *interventional equivalence classes* by maximizing the Bayesian information criterion. CAM assumes an *additive noise model* (ANM) where the nonlinear functions are additive. We use a modified version of CAM (*CAM**) where its maximum likelihood objective has been adapted for the interventional case. GraN-DAG assumes a Gaussian ANM model.

For each task, the performance of each method is assessed by two distances on the retrieved graph compared to the ground truth graph: i) the *structural Hamming distance* (SHD) and ii) the *structural interventional distance* (SID). The SHD is simply the number of edges that differ between the two DAGs (either reversed, missing or superfluous). While the SHD is a purely structural measurement, the SID is especially interesting for causal learning, since it assesses how two DAGs differ with respect to their causal inference statements (Peters & Bühlmann, 2015).

3.1.1 Synthetic data sets

We used 3 types of data sets with different mechanisms: i) linear functions with Gaussian noise, ii) Gaussian ANM, and iii) neural network. For each type of data set, graphs vary in term of nodes ($d = 20$ or 50) and expected number of edges per node ($e = 1$ or 4). For each setting, we sampled 10 DAGs following the *Erdős-Rényi* scheme and then data were generated with the corresponding mechanisms (See the Appendix A.1). A total of $n = 1000$ examples were sampled per graph, equally divided between the $d + 1$ intervention targets. Of these $d + 1$ targets, one is observational (no interventions) and for the d others, interventions were done on each node, one at a time. The target nodes were replaced by samples from a $\mathcal{N}(0, 1)$ without changing the other mechanisms.

We report the SHD and SID for all methods on the nonlinear Gaussian ANM data sets (Table 1), the linear data sets (Table 2) and the neural network data sets (Table 3). For each data set a hyperparameter search was performed on 50 hyperparameter combinations (see Appendix A.3). For each metric, we report the mean performance over the 10 data sets and its standard deviation. *GraN-DAG* denotes the version that supports interventional data and *GraN-DAG no interv* denotes the GraN-DAG method without interventions, trained on the same sample size ($n = 1000$), but on data sampled from the observational distribution of the same causal model. This way, we can evaluate the advantage of considering interventions (for a more systematic assessment, see Appendix A.2).

Table 1: Results for the nonlinear Gaussian ANM data sets with graphs of 20 and 50 nodes with expected connectivity (e) of 1 and 4. For both metrics, lower is better.

	20 nodes, $e = 1$		20 nodes, $e = 4$		50 nodes, $e = 1$		50 nodes, $e = 4$	
Method	SHD	SID	SHD	SID	SHD	SID	SHD	SID
GraN-DAG	2.4 ± 3.3	**2.1 ± 4.4**	**33.0 ± 13.8**	**126.2 ± 37.9**	**5.7 ± 2.9**	**33.5 ± 30.3**	**98.9 ± 17.1**	**1029.4 ± 159.5**
GraN-DAG no interv	**0.6 ± 1.3**	**2.2 ± 6.3**	**43.2 ± 14.4**	**159.7 ± 40.8**	**5.7 ± 3.0**	**35.3 ± 19.8**	**105.9 ± 16.0**	**1062.8 ± 176.2**
GIES	12.1 ± 6.1	19.4 ± 19.1	64.8 ± 10.0	275.3 ± 47.4	49.0 ± 12.6	154.8 ± 79.5	168.2 ± 24.8	1863.8 ± 163.8
CAM*	2.6 ± 2.3	**5.5 ± 4.9**	**41.0 ± 20.3**	**132.7 ± 40.1**	**6.1 ± 3.1**	**44.1 ± 31.3**	**95.7 ± 20.3**	**1089.6 ± 219.3**

In Table 1, we observe that *GraN-DAG*, *GraN-DAG no interv* and *CAM** tend to yield the best results. Since *GraN-DAG no interv* makes the right assumptions and that the graph is identifiable (Peters et al., 2014), purely observational data are sufficient to identify the graph. Nevertheless, with a finite sample, observing interventions seems to lead to slightly better performance on denser graphs. *GIES* has poor performance since it cannot adequately model the nonlinear functions.

Table 2: Results for the linear data sets with graphs of 20 and 50 nodes and expected connectivity (e) of 1 and 4. For both metrics, lower is better.

	20 nodes, $e = 1$		20 nodes, $e = 4$		50 nodes, $e = 1$		50 nodes, $e = 4$	
Method	SHD	SID	SHD	SID	SHD	SID	SHD	SID
GraN-DAG	**4.8 ± 3.6**	16.3 ± 13.8	57.9 ± 13.4	262.0 ± 28.6	20.6 ± 7.7	106.9 ± 70.1	**154.7 ± 22.8**	1604.5 ± 209.1
GraN-DAG no interv	11.3 ± 4.5	39.3 ± 16.8	79.6 ± 8.5	323.4 ± 24.0	32.2 ± 7.0	169.3 ± 70.1	179.8 ± 26.3	1815.8 ± 117.8
GIES	**3.3 ± 2.1**	**3.8 ± 12.0**	**23.3 ± 18.7**	**74.9 ± 60.5**	17.3 ± 3.8	**3.7 ± 6.4**	248.5 ± 78.7	**1130.3 ± 444.8**
CAM*	**3.4 ± 3.4**	**8.7 ± 14.5**	62.2 ± 24.6	181.1 ± 60.7	**5.2 ± 5.1**	21.2 ± 24.8	**149.1 ± 29.9**	1660.3 ± 288.0

In Table 2, we observe that *GIES* has the best performance for several conditions. This was expected, since this method specifically makes the linear Gaussian assumption. For a few settings, *CAM** and *GraN-DAG* have the best performance. Of note, *GraN-DAG* clearly has a better performance than *GraN-DAG no interv*. This can be explained by the fact that, unlike the ANM data set, the graph is not identifiable from observational data alone.

Table 3: Results for the neural network data sets with graphs of 20 and 50 nodes and expected connectivity (e) of 1 and 4. For both metrics, lower is better.

	20 nodes, $e = 1$		20 nodes, $e = 4$		50 nodes, $e = 1$		50 nodes, $e = 4$	
Method	SHD	SID	SHD	SID	SHD	SID	SHD	SID
GraN-DAG	**4.5 ± 3.6**	**20.2 ± 14.7**	**43.7 ± 5.1**	**222.7 ± 32.6**	**13.8 ± 4.5**	77.5 ± 25.2	**90.2 ± 14.0**	**1337.4 ± 183.9**
GraN-DAG no interv	**6.0 ± 3.0**	28.8 ± 21.1	54.8 ± 8.9	**253.8 ± 30.0**	**16.4 ± 6.1**	78.4 ± 40.7	109.2 ± 34.7	**1507.3 ± 190.2**
GIES	10.1 ± 4.4	**10.3 ± 11.3**	63.6 ± 13.2	**217.1 ± 37.3**	43.4 ± 9.0	**41.1 ± 24.1**	182.1 ± 51.6	**1380.7 ± 269.3**
CAM*	8.3 ± 4.6	23.7 ± 16.9	92.5 ± 34.0	**203.9 ± 57.7**	19.3 ± 10.6	109.2 ± 90.5	135.9 ± 35.9	1616.5 ± 200.3

We also explored the performance on a data set with functions that cannot be modeled by the different methods. In fact, neural networks with random initialization yield complex functions, often with heteroscedastic noise. In Table 3 we observe that, while *GraN-DAG no interv* is on par with *GraN-DAG* in some cases, *GraN-DAG* has the overall best performance.

3.1.2 Real-world data sets

As our real-world task, we used the cytometry data set of Sachs et al. (2005) which is commonly used in the causal literature. The measurements are the level of expression of phosphoproteins and phospholipids in human cells recorded under different experimental conditions, where reagents were used to activate or inhibit the measured proteins.

Since in some of these experimental conditions the perturbations were not directly done on a measured protein, we use only 5846 measurements of the 7466 measurements as in Wang et al. (2017). Of the 5846 measurements, 1755 measurements are considered observationals, while the other 4091 measurements are from five different interventions (with the following proteins as targets: Akt, PKC, PIP2, Mek, PIP3). The graph reconstructed by Sachs et al. (2005) is used as the ground truth DAG. It contains 11 nodes and 17 edges.

In Table 4, we present the SHD and SID for *GraN-DAG*, *GIES* and *CAM**. To have a clearer comparison, we also present the true positive, false negative, false positive, reversed edges, and the F_1 score. Overall, *GIES* shows the worst performance. This is likely because the linear assumption does not hold in this data set. *GraN-DAG* has the lowest SHD, but also the highest SID. This high SID can partially be explained by the relatively high number of reversed edges. *CAM** is clearly superior in term of the F_1 score, followed by *GraN-DAG* and *GIES*. One possible explanation might be that *CAM** has the right inductive bias for the mechanisms present in this data set. As part of future work, we intend to evaluate *GraN-DAG* on other real-world tasks such as the one from Dixit et al. (2016).

Table 4: Results for the cytometry data sets

Method	SHD	SID	tp	fn	fp	rev	F_1 score
GraN-DAG	33	35	7	3	23	7	0.35
GIES	45	34	10	0	41	7	0.33
CAM*	35	20	12	1	30	4	0.51

3.2 Discussion

We proposed an extension of GraN-DAG to the interventional setting. Although other methods (like *CAM**) yield better results in some conditions, GraN-DAG is often on par with the other methods on a variety of data sets. Also, GraN-DAG generally outperforms its purely observational counterpart, showing the beneficial effect of the interventions. Since the introduction of the continuous constraint from Zheng et al. (2018), several recent works (Yu et al., 2019; Zheng et al., 2019; Ng et al., 2019; Kalainathan et al., 2018) proposed causal discovery methods using neural networks with observational data alone. To the best of our knowledge, this is the first score-based method that frames the problem of learning a causal model from observational and interventional data as a continuous constrained optimization problem. Our results are promising and we intend to make a more extensive comparison to other methods with additional real-world data sets. This first step opens several new opportunities of future work. Since the identifiability problem is less severe in the presence of interventional data, it could be interesting to explore more expressive functions than Gaussian ANM. Moreover, in real-world applications we often do not know which variables have been targeted during interventions. This setting, referred to as *unknown interventions*, was addressed by Eaton & Murphy (2007). Recently, Ke et al. (2019) proposed a method based on neural networks for this problem. As part of future work, we intend to explore an extension of GraN-DAG to this challenging setting.

References

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 12 2014.

David Maxwell Chickering. Optimal structure identification with greedy search. In *Journal of Machine Learning Research*, volume 3, pp. 507–554, 2003. doi: 10.1162/153244303321897717.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866, 12 2016.

Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *Journal of Machine Learning Research*, volume 2, pp. 107–114, 2007.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, UAI 2008*, pp. 161–168, 6 2008.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments suficient and in the worst case necessary to identify all causal relations among N variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*, pp. 178–184, 7 2005.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv preprint arXiv:1803.04929*, 2018.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning Neural Causal Models from Unknown Interventions. 2019.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

Kevin B. Korb, Lucas R. Hope, Ann E. Nicholson, and Karl Axnick. Varieties of causal intervention. In *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, volume 3157, pp. 322–331. Springer, Berlin, Heidelberg, 2004.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-Based Neural DAG Learning. 2019.

Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, and Zhitang Chen. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 3 2015.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

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

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 4 2005.

Shohei Shimizu. Lingam: Non-Gaussian Methods for Estimating Causal Structures. *Behaviormetrika*, 41(1):65–98, 1 2014.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pp. 5823–5832, 2017.

Karren D. Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *35th International Conference on Machine Learning, ICML 2018*, volume 12, pp. 8823–8839. International Machine Learning Society (IMLS), 2018. ISBN 9781510867963.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. *arXiv preprint arXiv:1904.10098*, 2019.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pp. 9472–9483, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Learning sparse nonparametric dags. *arXiv preprint arXiv:1909.13189*, 2019.

A Appendix

A.1 Synthetic data sets generation

For each type of data set, we first sample a DAG following the *Erdős-Rényi* scheme and then we sample the parameters of the different mechanisms as stated below. For the observational case, we then sample $n/(d+1)$ examples (n examples in the case of *GraN-DAG no interv*). For each intervention, one node is uniformly chosen and replaced by a $\mathcal{N}(0,1)$ and then, $n/(d+1)$ examples are sampled (if n is not divisible by $d+1$, some intervention setting may have one extra sample in order to have a total of n samples). For all data sets, the source nodes are Gaussian with zero mean and variance sampled from $\mathcal{U}[1,2]$. The noise variables N_j are mutually independent and sampled from $\mathcal{N}(0,\sigma_j^2)\ \forall j$.

- The *nonlinear Gaussian ANM* data sets are generated following $X_j := f_j(X_{\pi_j^{\mathcal{G}}}) + N_j$ $\forall j$ where the functions f_j are independently sampled from a Gaussian process with a unit bandwidth RBF kernel and $\sigma_j^2 \sim \mathcal{U}[0.4, 0.8]$.
- The *linear* data sets are generated following $X_j|X_{\pi_j^{\mathcal{G}}} \sim w_j^T X_{\pi_j^{\mathcal{G}}} + 0.2 \cdot N_j\ \forall j$ where $\sigma_j^2 \sim \mathcal{U}[1,2]$ and w_j is a vector of $|\pi_j^{\mathcal{G}}|$ coefficients each sampled uniformly from $[-1,-0.25] \cup [0.25, 1]$.
- The *neural network* data sets are generated following $X_j := f_j(X_{\pi_j^{\mathcal{G}}}, N_j)\ \forall j$ where the functions f_j are fully connected neural networks with one hidden layer of 20 units and $\tanh$ as nonlinearities. The weights of each neural network are randomly initialized from $\mathcal{N}(0,1)$.

A.2 Impact of the number of interventions

We explored to what extent the number of interventions has an impact on the graph recovery. Using the linear data sets with graphs of 50 nodes and edge connectivity of 4, we tested the performance of GraN-DAG with interventions on data sets with $\{0, 5, 10, \ldots, 50\}$ interventions. The interventional targets are singleton, i.e. only one variable is intervened upon at a time. In Figure 1 and 2, we present the mean SHD and SID over 10 data sets for each number of interventions and their standard deviation. As expected, the number of interventions has a beneficial impact on the performance of GraN-DAG with intervention: more interventions lead to better SHD and SID.

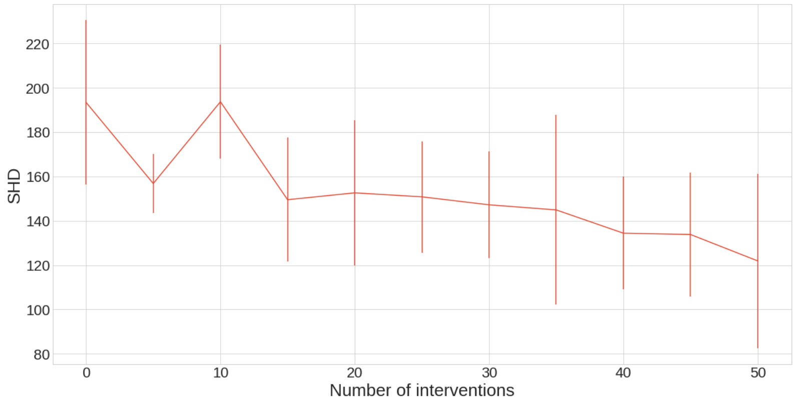

Figure 1: Effect of the number of interventions on SHD

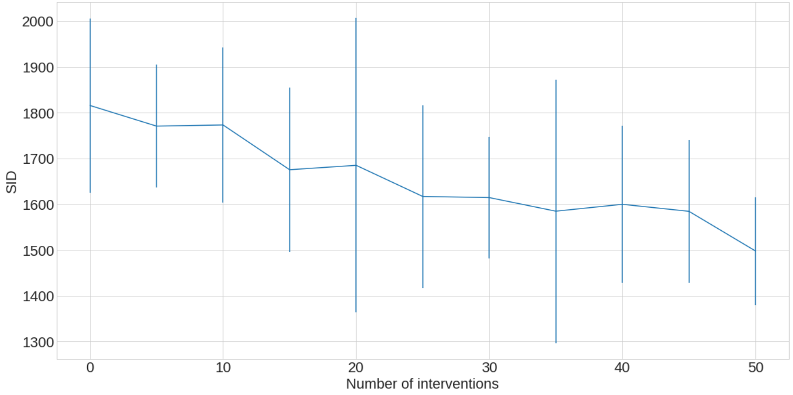

Figure 2: Effect of the number of interventions on SID

A.3 Hyperparameter search

For all methods, we performed a hyperparameter search over 50 hyperparameter combinations using a random search following the sampling scheme of Table 5. For each data set, the models were trained on 80% examples and evaluated on the 20% remaining examples (never on the ground truth DAG). For GIES, since the performances were always worse, we kept the results with the default value for the regularizer coefficient. Unless otherwise stated, GraN-DAG, with or without intervention, had the same hyperparameters as in Lachapelle et al. (2019): RMSprop was used as the optimizer, the NN's activation functions were leaky-ReLU and minibatches of size 64 were used. For *Jac-thresh* $= 1$, the Jacobian matrix trick is used for thresholding. Preliminary neighborhood selection (PNS) and pruning were also used.

Table 5: Hyperparameter search spaces for each algorithm

	Hyperparameter space
GraN-DAG (with/without interventions)	$\log_{10}$(learning rate) $\sim U[-2,-3]$ (first subproblem) $\log_{10}$(learning rate) $\sim U[-3,-4]$ (other subproblems) # hidden units $\sim U\{4,8,16,32,64\}$ # hidden layers $\sim U\{1,2,3\}$ jac-thresh $\sim U\{0,1\}$ PNS threshold $\sim U[0.5,0.75,1,2]$ $\log_{10}$(edge clamping threshold) $\sim U\{-3,-4,-5\}$ $\log_{10}$(pruning cutoff) $\sim U\{-7,-6,-5,-4,-3,-2,-1\}$ $\log_{10}$(constraint convergence tolerance) $\sim U\{-6,-8,-10\}$
CAM*	$\log_{10}$(pruning cutoff) $\sim U[-7,0]$
GIES	$\log_{10}$(regularizer coefficient) $\sim U[-4,4]$