Variational Causal Networks: Approximate Bayesian Inference over Causal Structures

YASHAS ANNADANI, ETH Zurich and Mila, Université de Montréal
JONAS ROTHFUSS, ETH Zurich
ALEXANDRE LACOSTE, ElementAI/ ServiceNow
NINO SCHERRER, ETH Zurich
ANIRUDH GOYAL, Mila, Université de Montréal
YOSHUA BENGIO, Mila, Université de Montréal
STEFAN BAUER, Max Planck Institute for Intelligent Systems

Learning the causal structure that underlies data is a crucial step towards robust real-world decision making. The majority of existing work in causal inference focuses on determining a single *directed acyclic graph (DAG)* or a Markov Equivalence Class (MEC) thereof. However, a crucial aspect to acting intelligently upon the knowledge about causal structure which has been inferred from finite data demands reasoning about its uncertainty. For instance, planning *interventions* to find out more about the causal mechanisms that govern our data requires quantifying *epistemic uncertainty* over DAGs. While Bayesian causal inference allows us to do so, the posterior over DAGs becomes intractable even for a small number of variables. Aiming to overcome this issue, we propose a form of variational inference over the graphs of Structural Causal Models (SCMs). To this end, we introduce a parametric variational family modeled by an autoregressive distribution over the space of discrete DAGs. Its number of parameters do not grow exponentially with the number of variables and can be tractably learned by maximizing an Evidence Lower Bound (ELBO). In our experiments, we demonstrate that the proposed variational posterior is able to provide a good approximation of the true posterior.

Additional Key Words and Phrases: Causal Inference, Bayesian Structure Learning, Variational Inference

ACM Reference Format:
Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer. 2021. Variational Causal Networks: Approximate Bayesian Inference over Causal Structures. In *BCIRWIS 2021: Bayesian causal inference for real world interactive systems - (KDD 2021 Workshop), August 2021 14–15, 2018, Virtual.* ACM, New York, NY, USA, 13 pages.

1 INTRODUCTION

Moving from learning correlation and association in data to causation is a critical step towards increased robustness, interpretability and real-world decision-making [23, 30]. While the majority of work on causal inference [5–7, 33] deals with getting a single underlying causal structure without a probabilistic treatment, quantifying the *epistemic uncertainty* in case of non-identifiability is crucial and is not possible in these approaches. In this work, we take a Bayesian approach to causal structure learning. Given *only finite observational data*, Bayesian inference allows us to quantify the uncertainty in the causal structure of the data generating process and thus reason about how much information we gain when performing specific interventions on the graph. However, due to the super-exponential

Manuscript submitted to ACM

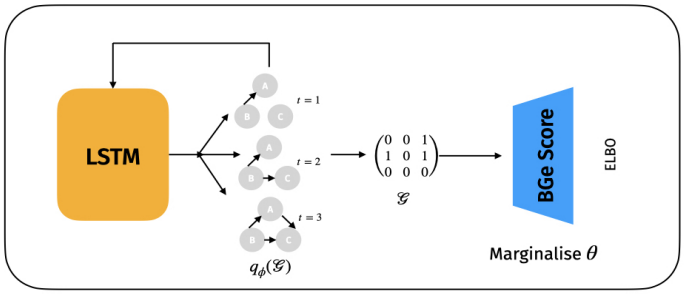

Fig. 1. Schematic diagram of the proposed approach. We perform autoregressive sampling of the $\{0, 1\}$ adjacency matrix with Bernoulli parameters from the model and then compute the ELBO by marginalising the parameters of SCM using the BGe score [10, 19].

growth of the space of Directed Acyclical Graphs (DAGs) in the number of variables, Bayesian inference on discrete causal structures is highly intractable. Sampling from structural posteriors using MCMC techniques involve lot of heuristics and usually show slow mixing and convergence [2, 9, 21]. Aiming to overcome these challanges, we propose to form a variational approximation of the posterior over DAGs which can be learned using gradient descent techniques. To this end, we contribute a family of distributions over adjacency matrices and the necessary tools to ensure that the modeled adjacency matrices correspond to acyclic graphs. In our experiments, we evaluate our proposed Variational Causal Networks (VCN) approach on linear Gaussian SCM's with additive noise (App.A).

2 VARIATIONAL CAUSAL NETWORKS

In variational inference [4], we aim to approximate the Bayesian posterior over graph structures $\mathcal{G}$ given data $\mathcal{D}$ $p(\mathcal{G}|\mathcal{D})$ by a variational distribution $q(\mathcal{G})$ that has a tractable density. In particular, we use a density $q_\phi(\mathcal{G})$ that is parameterized by ϕ.

To learn the parameters of our variational distribution, we minimize the KL-Divergence between $q_\phi(\mathcal{G})$ and the true posterior $p(\mathcal{G})$ which is equivalent to maximizing the *Evidence Lower Bound (ELBO)*, given by

$$\log p(\mathcal{D}) \geq \mathcal{L}(\phi; \mathcal{D}) = \mathop{\mathbb{E}}_{q_\phi(\mathcal{G})} \left[\log p(\mathcal{D}|\mathcal{G})\right] - \mathrm{D}_{\mathrm{KL}} \left(q_\phi(\mathcal{G})||p(\mathcal{G})\right) \tag{1}$$

To maximize the ELBO, we estimate the gradients $\nabla_{(\phi)} \mathcal{L}(\phi; \mathcal{D})$ to employ a form of gradient ascent on the objective.

A key challenge to variational inference over SCMs concerns the choice of the variational distribution $q_\phi(\mathcal{G})$. For instance, if we choose $q_\phi(\mathcal{G})$ naively as categorical distribution over the space of DAGs, the dimensionality of ϕ grows super-exonentially in the number of variables d (i.e. $O(2^{d^2})$), rendering this choice infeasible.

Instead, we represent graphs by their adjacency matrix $\mathbf{A} \in \{0, 1\}^{d \times d}$ and model q_ϕ as a discrete distribution over such $\{0, 1\}$-matrices. While Ke et al. [16] represent each entry of the adjacency matrix as independent Bernoulli variable, such approach is insufficient to capture any dependencies which ensure that the graph is a DAG. In addition, a factorisable distribution can only capture unimodal distributions while the posterior over causal structures in the non-identifiable case could have exponential number of modes. In order to address these issues, we model this discrete distribution as an autoregressive distribution over entries of the adjacency matrix using an LSTM [15].

Let $q_\phi(\mathcal{G}) = q_\phi(\mathbf{A}_\mathcal{G})$ be defined in an autoregressive manner as follows:

$$q_\phi(\mathbf{A}_\mathcal{G}) = \prod_{i=1}^{d(d-1)} q_\phi(a_{\mathcal{G}_i}|a_{\mathcal{G}_{1:i-1}}) \quad \text{s.t.} \quad \phi(a_{\mathcal{G}_i}|a_{\mathcal{G}_{1:i-1}}) := \mathrm{Ber}\left(a_{\mathcal{G}_i}; f_\phi\left(a_{\mathcal{G}_{1:i-1}}\right)\right), \quad \mathbf{A}_\mathcal{G} = \{a_{\mathcal{G}_i}\}_{i=1}^{d(d-1)}$$

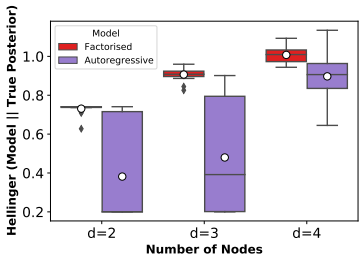

Fig. 2. Hellinger distance of full posterior of the model with the true posterior.

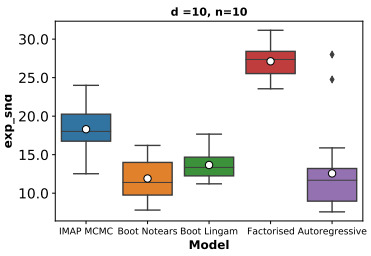

Fig. 3. $\mathbb{E}$ [SHD] for $d = 10$ node ER random graphs (lower is better).

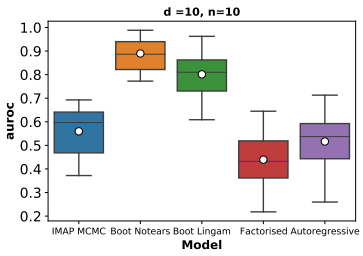

Fig. 4. AUROC for $d = 10$ node ER random graphs (higher is better).

where f is a function which predicts the Bernoulli parameter based on the previous realisations of the entries of the adjacency matrix. We model this function using an LSTM [15]. In causal discovery, it is usually a requirement that the search space of all graphs is a DAG. However, in the parameterisation described above, the discrete distribution is defined over all possible graphs, including the one which contains cycles. Hence, to enforce acyclicity and encode additional assumptions about the data generating process such as sparsity, we use the prior $p(\mathcal{G}) \propto \exp\left(-\lambda_t g(\mathbf{A}_\mathcal{G}) - \lambda_s \left\|\mathbf{A}_\mathcal{G}\right\|_1\right)$ where $g(\mathbf{A}_\mathcal{G})$ is the DAG constraint given by the matrix exponential [33], i.e $g(\mathbf{A}_\mathcal{G}) = \text{tr}\left[e^{\mathbf{A}_\mathcal{G}}\right] - d$, and λ_t and λ_s are tunable hyperparameters. Typically, $\lambda_t \to \infty$ if graphs with cycles should have zero probability. In order to marginalise the parameters of the linear Gaussian SCM, we use standard parameter priors given by the *BGe score* [10, 19]. We note that since we take a Monte Carlo estimate of the KL term in the ELBO over a discrete space, we need an unbiased estimator of the gradients. Therefore, we use a score-function estimator [31] with exponential moving average baseline for obtaining gradients.

3 EXPERIMENTS

We evaluate our method on synthetic dataset. For generating synthetic data, we sample a DAG at random from an Erdos-Renyi (ER) model with expected number of edges equal to d. Each reported result is over 20 different random graphs. For low dimensional variables ($d \leq 4$), we can enumerate the true posterior for all graphs. Hence, in this setting we compute the Hellinger distance between the variational posterior learned from data and the true posterior. For higher dimensional variables, we evaluate on Expected Structural Hamming Distance $\mathbb{E}$[SHD] and Area Under Reciever Operating Curve (AUROC) [26] (See App. D.1). Detailed experimental settings and baselines are described in App. D.

4 DISCUSSION AND CONCLUSION

As can be seen from Figures 5,7 and 8, the proposed approach obtains a better performance than the factorised distribution, mainly due to the non-identifiability of graphs and the fact that the true posterior is multimodal. It gives competitive results against other baselines in higher dimensions. We would like to note that there is no perfect metric which can quantify how well the approximation is in higher dimensions and is in general hard. The evaluated metrics, while being imperfect, nevertheless shed light on some of the downstream tasks the uncertainty estimates might be useful in[1]. One of the main limitations of this approach is the scalability of this method to larger dimensions. The difficulty in scaling comes from the fact that the score function estimator has high variance in higher dimensions. Nevertheless, we believe that the proposed approach is promising and the uncertainty estimates obtained in this framework can be useful for selecting the most informative interventions, performing budgeted interventional experiments as well as representation learning of high-dimensional signals like natural images [3].

[1] A model which has perfect approximation of the true posterior might still not give highest possible scores on these metrics.

REFERENCES

[1] Raj Agrawal, Chandler Squires, Karren Yang, Karthik Shanmugam, and Caroline Uhler. 2019. ABCD-strategy: Budgeted experimental design for targeted causal structure discovery. *arXiv preprint arXiv:1902.10347* (2019).

[2] Raj Agrawal, Caroline Uhler, and Tamara Broderick. 2018. Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In *International Conference on Machine Learning*. PMLR, 89–98.

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

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.* 112, 518 (Apr 2017), 859–877. https://doi.org/10.1080/01621459.2017.1285773

[5] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. 2020. Differentiable Causal Discovery from Interventional Data. *arXiv preprint arXiv:2007.01754* (2020).

[6] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* (2014).

[7] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research* (2002).

[8] Byron Ellis and Wing Hung Wong. 2008. Learning causal Bayesian network structures from experimental data. *J. Amer. Statist. Assoc.* (2008).

[9] Nir Friedman and Daphne Koller. 2003. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning* 50, 1 (2003), 95–125.

[10] Dan Geiger, David Heckerman, et al. 2002. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics* (2002).

[11] Marco Grzegorczyk and Dirk Husmeier. 2008. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning* (2008).

[12] Alain Hauser and Peter Bühlmann. 2012. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of Machine Learning Research* (2012). http://jmlr.org/papers/v13/hauser12a.html

[13] David Heckerman, Christopher Meek, and Gregory Cooper. 1999. A Bayesian approach to causal discovery. *Computation, causation, and discovery* (1999).

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

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[16] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. 2019. Learning Neural Causal Models from Unknown Interventions. *arXiv preprint arXiv:1910.01075* (2019).

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Jack Kuipers and Giusi Moffa. 2017. Partition MCMC for inference on acyclic digraphs. *J. Amer. Statist. Assoc.* (2017).

[19] Jack Kuipers, Giusi Moffa, David Heckerman, et al. 2014. Addendum on the scoring of Gaussian directed acyclic graphical models. *Annals of Statistics* 42, 4 (2014), 1689–1691.

[20] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. 2019. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226* (2019).

[21] David Madigan, Jeremy York, and Denis Allard. 1995. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* (1995).

[22] Teppo Niinimäki, Pekka Parviainen, and Mikko Koivisto. 2016. Structure discovery in Bayesian networks by sampling partial orders. *The Journal of Machine Learning Research* (2016).

[23] Judea Pearl. 2009. *Causality*. Cambridge university press.

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

[25] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference*. The MIT Press.

[26] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. 2010. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS one* (2010).

[27] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* (2017).

[28] Thomas Schaffter, Daniel Marbach, and Dario Floreano. 2011. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* (2011).

[29] Shohei Shimizu. 2014. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika* (2014).

[30] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[31] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* (1992).

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

[33] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*. 9472–9483.

A PROBLEM SETTING

Causal modeling. Consider a set of d random variables $\mathbf{X} := \{X_1, \ldots, X_d\}$. A Structural Causal Model (SCM) [25] over $\mathbf{X}$ is defined as set of structural assignments

$$X_i := f_i(\mathbf{x}_{\pi_{\mathcal{G}}(i)}, \epsilon_i)\ , i = 1, ..., d \tag{2}$$

corresponding to a Directed Acyclic Graph (DAG) $\mathcal{G}$ with vertices $\mathbf{X}$. In here, $\pi_{\mathcal{G}}(i)$ are the parents of X_i in $\mathcal{G}$, ϵ_i are independent noise variables with probability density P_{ϵ_i} and the f_i's are (potentially non-linear) functions. The SCM entails a joint probability distribution $P_{\mathbf{x}}$ over the random variables. In this work, we assume that all the endogenous variables are observed, that is, there are no hidden confounders.

A popular instantiation of this generic framework are linear SCMs with additive noise, given by:

$$x_i := \boldsymbol{\theta}_i^T \mathbf{x}_{\pi_{\mathcal{G}}(i)} + \epsilon_i\ \boldsymbol{\theta}_i \in \mathbb{R}^{|\pi_{\mathcal{G}}(i)|} \tag{3}$$

wherein $\boldsymbol{\theta}_i$ are parameters (edge weights) of the linear functions f_i. Alternatively, this can be written as

$$x_i := \boldsymbol{\theta}_i^T \left(\mathbf{X} \circ \mathbf{A}_{\mathcal{G}_i}\right) + \epsilon_i\ \boldsymbol{\theta}_i\ ,\ \mathbf{A}_{\mathcal{G}_i} \in \mathbb{R}^d\ , \tag{4}$$

where $\circ$ corresponds to elementwise product, $\mathbf{A}_{\mathcal{G}_i}$ is the i^{th} row of $\mathbf{A}_{\mathcal{G}} \in \mathbb{R}^{d \times d}$, the $(0, 1)$-adjacency matrix of $\mathcal{G}$. We restrict our further exposition to continuous random variables. However, the framework presented in the remainder of the paper applies to discrete random variables as well. We focus on linear SCMs with Gaussian variables, as they are non-identifiable in this setting and hence obtaining uncertainty estimates is useful. Non-linear additive noise SCM's with Gaussian variables are identifiable, and hence a single point estimate suffices.

NOTEARS - Causal inference as continuous optimization problem. In practice, we often have data while the causal structure that underlies the data generating process is unknown to us. Causal discovery is concerned with recovering this causal structure, e.g. in the form of an SCM, from observational data or interventional data (or both). Assuming a linear SCM, this coincides with estimating the graph $\mathcal{G}$, i.e. its adjacency matrix $\mathbf{A}$, and its corresponding edge weights $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_d\}$, given data $\mathcal{D} := \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathcal{X}^n$.

In score-based causal discovery, we use a score function $F : \{0, 1\}^{d \times d} \times \mathcal{X}^n \mapsto \mathbb{R}$ to find the graph that best corresponds to the data as follows:

$$\underset{\mathbf{A} \in \{0,1\}^{d \times d}}{\arg\min}\ F(\mathbf{A}, \mathcal{D}) \quad \text{s.t. } \mathcal{G}(\mathbf{A}) \in \text{DAG}(d) \tag{5}$$

where $\mathbf{A}$ is the $\{0, 1\}$ adjacency matrix of graph $\mathcal{G}$.

However, since the search space of DAGs of this combinatorial optimization problem grows in the order of $O(2^{d^2})$, solving (5) becomes infeasible even for a relatively small number of variables d.

Addressing this issue, Zheng et al. [33] propose an alternative formulation that converts the combinatorial problem into the following continuous program:

$$\underset{\boldsymbol{W} \in \mathbb{R}^{d \times d}}{\arg\min}\ F(\boldsymbol{W}, \mathcal{D}) \quad \text{s.t. } \text{tr}\left(e^{\boldsymbol{W}}\right) - d = 0\ . \tag{6}$$

that can be solved with standard tools of constrained optimization such as the Lagrange method. In here, we have converted the adjacency matrix $\mathbf{A} \in \{0,1\}^{d\times d}$ into a weighted adjacency matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ such that $a_{i,j} = 1 \Leftrightarrow w_{i,j} \neq 0$.

Bayesian inference over DAGs. An important aspect towards acting intelligently upon the knowledge about the causal structure which has been inferred from finite data demands reasoning about its uncertainty. For instance, in scientific inquiry, planning *interventional experiments* to find out more about the causal mechanisms that govern a data-generating system of interest requires quantifying *epistemic uncertainty* over DAGs. Bayesian inference gives us a principled framework for the treatment of such uncertainty (Figure **??**).

In the Bayesian framework, we encode our prior structural knowledge in form of a *prior* $p(\mathcal{G})$ over DAGs. By combining the prior with the *likelihood* $p(\mathcal{D}|\mathcal{G})$ through Bayes' rule, we obtain a *posterior distribution* over DAGs:

$$p(\mathcal{G}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{G})p(\mathcal{G})}{p(\mathcal{D})} \tag{7}$$

where $p(\mathcal{D}) = \sum_{\mathcal{G}} \int_{\theta} p(\mathbf{x}|\mathcal{G},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{G})p(\mathcal{G})d\boldsymbol{\theta}$ is the *model evidence*. A key issue with (7) is the intractable evidence term in the denominator. Even if the integral over $\boldsymbol{\theta}$ can be solved analytically such as in the case of linear structural equations with Gaussian prior and likelihood with standard parameter priors [10], the sum over possible DAGs grows in the order of $O(2^{d^2})$, making its computation infeasible even for a small number of variables d.

B DETAILED METHODOLOGY

In this section, we present our variational inference (VI) framework for approximating Bayesian posteriors over causal structures (Equation 7). First, we derive the evidence lower bound and sketch out how to perform VI on DAGs. Unlike the variational inference in a latent variable model [4], the evidence lower bound is still intractable in the case of causal models due to the superexponential number of graph configurations. Therefore, we introduce a novel parametric variational family which makes the intractable ELBO into a tractable one while still having the flexibility to model complex distributions over DAGs. Such a variational family can be modelled using autoregressive models like LSTM [15]. This particular instantiation of approximate Bayesian inference for SCMs with the presented variational family is called Variational Causal Networks (VCN).

B.1 Variational Inference for Causal Structures

In variational inference [4], we aim to approximate the Bayesian posterior $p(\mathcal{G}|\mathcal{D})$ by a variational distribution $q_\phi(\mathcal{G})$, parameterized by ϕ, that has a tractable density. To learn the parameters of our variational distribution, we minimize the KL-Divergence between $q_\phi(\mathcal{G})$ and the true posterior $p(\mathcal{G}|\mathcal{D})$ which is equivalent to maximising the *Evidence Lower Bound (ELBO)*, given by the following proposition.

Proposition 1. ***(ELBO)*** *Let $q_\phi(\mathcal{G})$ be the variational posterior over causal structures. Then the evidence lower bound (ELBO) is given by:*

$$\log p(\mathcal{D}) \geq \mathcal{L}(\phi;\mathcal{D}) = \mathop{\mathbb{E}}_{q_\phi(\mathcal{G})}\left[\log p(\mathcal{D}|\mathcal{G})\right] - \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\mathcal{G})||p(\mathcal{G})\right) \tag{8}$$

Proof.

$$\begin{aligned}
&\arg\min_{\phi} \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\mathcal{G})||p(\mathcal{G}|\mathcal{D})\right) \\
&\overset{Eq.7}{=} \arg\min_{\phi} \mathop{\mathbb{E}}_{q_\phi(\mathcal{G})}\left[\log p(\mathcal{D}|\mathcal{G}) + \log p(\mathcal{D}) + \log q_\phi(\mathcal{G}) - \log p(\mathcal{G})\right] \\
&= \arg\min_{\phi} \mathop{\mathbb{E}}_{q_\phi(\mathcal{G})}\left[-\log p(\mathcal{D}|\mathcal{G})\right] + \log p(\mathcal{D}) + \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\mathcal{G})||p(\mathcal{G})\right)
\end{aligned}$$

which gives us a lower bound on the marginal log-likelihood of the data.

$$\log p(\mathcal{D}) \geq \mathop{\mathbb{E}}_{q_\phi(\mathcal{G})}\left[\log p(\mathcal{D}|\mathcal{G})\right] - \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\mathcal{G})||p(\mathcal{G})\right)$$

□

To maximise the ELBO, we estimate the gradients $\nabla_{(\phi)}\mathcal{L}(\phi;\mathcal{D})$ to employ a form of gradient ascent on the objective.

B.2 A Variational Family for Causal Structures

A key challenge to variational inference over DAGs concerns the choice of the variational distribution $q_\phi(\mathcal{G})$ over discrete DAGs $\mathcal{G}$. For instance, if we choose $q_\phi(\mathcal{G})$ naively as categorical distribution over the space of DAGs, the dimensionality of ϕ grows super-exponentially in the number of variables d (i.e. $O(2^{d^2})$), rendering this choice impractical. Hence, we need to find a way of representing $q_\phi(\mathcal{G})$ that is not combinatorial in its nature while ensuring a rich family of distributions over graphs. In addition, to be able to compute Monte Carlo gradient estimates of the ELBO in (1), $q_\phi(\mathcal{G})$ needs to have a tractable probability density.

Aiming to fulfil these requirements, we represent graphs by their adjacency matrix $\mathbf{A} \in \{0,1\}^{d\times d}$ and model q_ϕ as a discrete distribution over such $\{0,1\}$-matrices. While Ke et al. [16] represent each entry of the adjacency matrix as independent Bernoulli variable, such approach is insufficient to capture any dependencies between the entries of the adjacency matrix which are necessary to ensure that $q(\mathbf{A})$ only assigns positive probabilities to adjacency matrices corresponding to DAGs. For instance, to avoid cycles of length two, $(\mathbf{A})_{i,j}$ needs to have a strong negative dependency $(\mathbf{A})_{j,i}$, i.e., if there is a directed edge $i \to j$ there must be no edge $j \to i$, and reverse. In addition, a factorisable distribution can only capture unimodal distributions while the posterior over causal structures in the non-identifiable case could have exponential number of modes. Hence, the variational distribution needs to be parameterised such that multiple modes could be captured. In order to address these issues, we model this discrete distribution as an autoregressive distribution over entries of the adjacency matrix using an LSTM [15].

Let $q_\phi(\mathcal{G}) = q_\phi(\mathbf{A}_\mathcal{G})$ be defined in an autoregressive manner as follows:

$$\begin{gathered}
q_\phi(\mathbf{A}_\mathcal{G}) = \prod_{i=1}^{d(d-1)} q_\phi(a_{\mathcal{G}_i}|a_{\mathcal{G}_{1:i-1}}) \\
s.t.\ q_\phi(a_{\mathcal{G}_i}|a_{\mathcal{G}_{1:i-1}}) := \text{Bernoulli}\left(a_{\mathcal{G}_i}; f_\phi\left(a_{\mathcal{G}_{1:i-1}}\right)\right) \\
\mathbf{A}_\mathcal{G} = \{a_{\mathcal{G}_i}\}_{i=1}^{d(d-1)}
\end{gathered}$$

where only the non-diagonal elements of the adjacency matrix are modelled and f_ϕ is a function which predicts the Bernoulli parameter based on the previous realisations of the entries of the adjacency matrix, thus making the distribution autoregressive. We model this function using an LSTM [15].

Note that using an autoregressive distribution on the entries of the adjacency matrix helps to implicitly keep a distribution over super-exponential number of graphs, and also be able to sample from that distribution.

Prior over graph structures. Choosing the appropriate prior $p(\mathcal{G})$ over graph structures is important to learn good approximation of the posterior. In causal discovery, it is usually a requirement that the search space of all graphs is a DAG. However, in the parameterisation described above, the discrete distribution is defined over all possible graphs, including the one which contains cycles. Having support over graphs with cycles and using maximum likelihood estimation results in high probability mass corresponding to a fully connected graph in the approximate posterior. Therefore, appropriate DAG regularizers are required. We employ the result of NOTEARS [33] to define a Gibbs distribution as the prior which helps us to limit the support of graphs to just DAGs. We can also encode additional assumptions about the data generating process such as sparsity. The Gibbs distribution in this case could be defined as:

$$p(\mathcal{G}) = \frac{\exp\left(-\lambda_t g(\mathbf{A}_{\mathcal{G}}) - \lambda_s \left\|\mathbf{A}_{\mathcal{G}}\right\|_1\right)}{Z_d} \tag{9}$$

where $g(\mathbf{A}_{\mathcal{G}})$ is the DAG constraint given by the matrix exponential [33], i.e $g(\mathbf{A}_{\mathcal{G}}) = \text{tr}\left[e^{\mathbf{A}_{\mathcal{G}}}\right] - d$. Matrix binomial can also be used to define g, as given in [32]. The second term in the Gibbs distribution corresponds to the sparsity constraint. λ_t and λ_s are tunable hyperparameters. Typically, $\lambda_t \to \infty$ if graphs with cycles should have zero probability mass under the posterior. However, any large value should still give a reasonable prior with very low probabilities for cyclic graphs. Z_d is the partition function of the Gibbs distribution which in practice cannot be computed easily for data with dimensionality $d \geq 5$. However, we note that computing this is not required for the optimisation of the ELBO (Eq. 1) as it turns out to be a constant.

Marginal Likelihood. We use standard parameter priors which ensure marginal likelihood $\log p(\mathcal{D}|\mathcal{G})$ is "score-equivalent", i.e. all the DAGs which are in the same MEC have the same marginal likelihood. This can be ensured with a Gaussian-Wishart prior over the parameters $\boldsymbol{\theta}$ and the marginalisation can be done in closed form [10] (called as the *BGe score*). Using this parameter prior to calculate marginal likelihood in the ELBO makes the ELBO score-equivalent.

B.3 Estimating Gradients

Since the KL regulariser is defined on a discrete distribution with many parameters, obtaining a Monte Carlo estimate of this term (and hence the corresponding ELBO) requires unbiased gradients of ϕ. Therefore, we use the score-function gradient estimator [31] with exponential moving average baseline for obtaining the gradients. As the score function estimator does not require the estimand function to be differentiable, we can perform closed form marginalisation of the parameters with the standard parameter priors like that of Gaussian-Wishart.

The detailed algorithm is given in Algorithm 1.

C RELATED WORK

Causal Discovery: Broadly, causal structure learning approaches can be mainly categorised to two different paradigms with regards to kind of data used: methods involving learning structure from purely observational data and methods involving learning structure from both observational and interventional data. Both these approaches strongly rely on the identifiability of SCM. The approaches which learn structures from purely observational data usually assume additive gaussian noise model where the noise is mutually independent. While these approaches have the advantage that interventional data is not needed which might be unavailable or impossible to obtain in some cases, they have

Algorithm 1 VCN Algorithm

$\phi \leftarrow$ initialise parameters
repeat
  Sample L graphs $\{\mathcal{G}^{(i)}\}_{i=1}^{L}$ autoregressively from $q_\phi(\mathbf{A}_\mathcal{G})$ using an LSTM.
  Calculate the per-sample marginal log-likelihood $\log p(\mathcal{D}|\mathcal{G}^{(i)})$ by marginalising $\boldsymbol{\theta}$ to obtain the BGe Score [].
  $g \leftarrow \frac{1}{L}\sum_{i=1}^{L} \nabla_\phi \left[\log p(\mathcal{D}|\mathcal{G}^{(i)}) - \left[\log q_\phi(\mathcal{G}^{(i)}) - \log p(\mathcal{G}^{(i)})\right]\right]$
  $\phi \leftarrow$ Update parameters using gradients g.
until $\phi \leftarrow$ converge.

to make rather restrictive assumptions about the SCM. Lachapelle et al. [20], Yu et al. [32], Zheng et al. [33] employ this assumption and use a score-based objective with global search over DAG space to learn the structure. Lachapelle et al. [20] extend the linear model of [33] to non-linear case using neural networks. Apart from these approaches, a few of the other ones use local search heuristics and maximize a score based on Bayesian Information Criterion (BIC). GES [7, 27] greedily searches over the space of CPDAG's. LiNGAM [29] uses an Independent Component Analysis (ICA) based approach by making the assumption that either the variables or one of the noise variables are non-Gaussian. PC [30] algorithm uses a constraint based approach for causal discovery.

Different from the above approaches, a few approaches use a combination of observational and interventional data to learn the causal structure. ICP [24] assumes a linear SCM and resorts to conditional independence testing to test the hypothesis of invariance, a concept wherein the plausible causal predictors of a given variable are stable across different interventions other than intervention on the variable of interest. [14] extend ICP to non-linear SCMs. These methods do not infer the complete graph but instead just the causal parents of a particular variable. The difficulty of these approaches rest in the fact that conditional independence tests are hard to perform. [12] derive a graph theoretic criterion to characterize the Markov Equivalence Class under the presence of interventional data called Interventional Markov Equivalence Class and generelize the GES [7] to the case of interventional data. More recently, [3] and [16] use a meta-learning objective to learn the causal structure using interventional data.

Bayesian Approach to Structure Learning: Existing approaches for Bayesian learning of causal graphs are mainly concerned with efficient sampling of graph structures from the posterior distribution. A popular choice for such an approach is the Markov Chain Monte Carlo (MCMC) with suitably chosen energy functions and heuristics [8, 13, 18, 21, 22]. While Niinimäki et al. [22] is concerned with causal discovery by sampling graph structures from partial orders of edge orientation, Madigan et al. [21] proposes an approach for learning probabilistic graphical models with discrete variables. Grzegorczyk and Husmeier [11], Kuipers and Moffa [18] and Ellis and Wong [8] propose improved MCMC techniques for sampling graph structures by restricting graph space to certain topology and sampling from a restricted space. Heckerman et al. [13] use a BIC based scoring function to learn a bayesian network by sampling from Metropolis-Hastings under different energy configurations. An estimate of MAP is obtained using Maximum Likelihood to get the final graph structure. Agrawal et al. [1] use DAG bootstrapping to estimate the posterior over graphs and use this for budgeted experimental design for causal discovery. While all these techniques are intended for a Bayesian approach to structure learning, they face the problem of efficiently approximating the posterior over graph structures and hence resort to either heuristics or restricted graph structures. However, we do not make any simplifying assumptions about the graph structures and we can efficiently approximate the posterior while being able to sample exactly from them.

D DETAILED EXPERIMENTS

To validate the modelling choice presented in the main section, we perform experiments mainly focusing on the following aspects: (1) since we approximate the posterior, we evaluate how close is the approximation to the true posterior in lower dimensional settings (≤ 4) where enumeration of true posterior is possible. (2) We outline the difficulty involved in evaluating the Bayesian Causal inference models in higher dimensions where enumeration is not possible and suggest possible metrics which alleviate the problem and quantify how well the true posterior is approximated. We further evaluate our model on these metrics.

Key Findings. We summarise the experimental results as follows: (1) The proposed autoregressive approach of VCN performs better than a factorised distribution [16] in all the considered settings. This is due to the ability of an autoregressive distribution to model multi-modal posteriors as compared to the unimodal factorised distribution. (2) In higher dimensions ($d \geq 10$), the proposed approach performs better or comparable to competitive baselines on various metrics. (3) VCN achieves a good approximation of the posterior with a smaller runtime, thus making it a suitable approach for Bayesian causal inference. (4) VCN achieves favourable results on the real gene-expression dataset of *Dream4* in-silico network challenge [28].

Experimental Settings. We evaluate our method on both synthetic and real datasets. For generating synthetic data, we follow the procedure of NOTEARS [33]. We sample a DAG at random from an Erdos-Renyi (ER) model with expected number of edges equal to d. Each reported result is over 20 different random graphs. The models were trained with a learning rate of $1e-2$ using the Adam optimiser [17] for 30k epochs. We consider the settings when we have $n = 10, 100$ samples. For taking the Monte Carlo estimate of the ELBO, we take $L = 1000$ samples. λ_s is fixed to 0.01 and λ_t is annealed from 10 to 1000 with an exponential annealing schedule.

Baselines. Parameterising the posterior with a **factorised distribution** [16] is our main baseline as it involves differentiable causal learning based techniques similar to our approach. We also compare with DAG Bootstrap [1] where the posterior is estimated by bootstrapping the data with any causal discovery algorithm and then forming an empirical estimate of the posterior based on frequency count. We use LiNGAM [29] and NOTEARS [33] as underlying causal discovery algorithms for DAG Bootstrap. They are denoted by **Boot Lingam** and **Boot Notears** respectively. In addition, we compare with the MCMC approach of Minimal **IMAP MCMC** [2]. We use the default hyperparameters and settings which is suggested in the respective approaches. For DAG Bootstrap, we perform 1000 bootstrap sets.

D.1 Evaluation Metrics

For low dimensional variables ($d \leq 4$), we can enumerate the true posterior for all graphs. Therefore, in this setting we compute the distance between the variational posterior learned from data and the true posterior.

For higher dimensional variables, where enumerating the true posterior is not possible, evaluating Bayesian causal inference algorithms is not straightforward and is in general hard. Directly comparing the likelihoods do not necessarily ensure that the posterior learned is good, as graphs with more edges always have higher likelihoods in the case of structure learning. While there is no perfect metric to evaluate uncertainty quantification, we adapt metrics for evaluating point estimates with sampling based distributional generalisations. Therefore, we employ the following metrics:

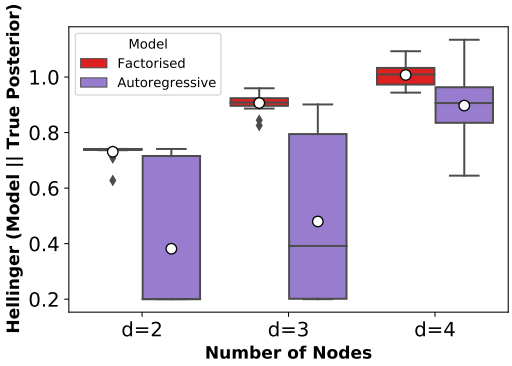

Fig. 5. Hellinger distance of the full posterior of the approximation with the true posterior.

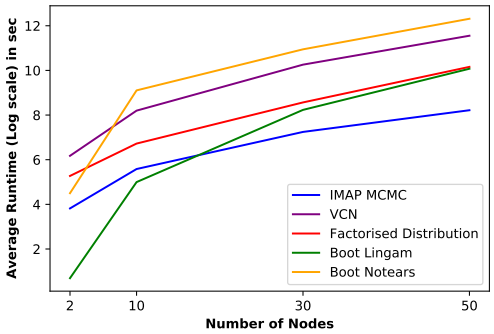

Fig. 6. Average runtime (in seconds, log scale) of different Bayesian causal inference approaches.

Expected Structural Hamming Distance ($\mathbb{E}[SHD]$*).* Given that the true posterior has modes over all the graphs inside the MEC corresponding to the ground truth graph, we can sample the graphs from the model and then compute the Structural Hamming Distance (SHD) between these samples and the ground truth. We can then compute the empirical mean of these SHDs as a metric. As all the graphs inside the MEC have the same number of edges, this metric indicates how well the approximated posterior is close to the ground truth on average. If $\mathcal{G}_{\text{GT}}$ is the ground truth data generating graph, then $\mathbb{E}[\text{SHD}]$ is given by

$$\mathop{\mathbb{E}}_{q_\phi(\mathcal{G})}[\text{SHD}] \approx \frac{1}{T}\sum_{i=1}^{T}\left[\text{SHD}(\mathcal{G}^{(i)}, \mathcal{G}_{\text{GT}})\right] \;\; \mathcal{G}^{(i)} \sim q_\phi(\mathcal{G}) \tag{10}$$

Area Under Reciever Operating Curve (AUROC). If we care for just feature probabilities of quantities like the presence of an edge, we can compute edge beliefs and compute the Receiver Operating Curve (ROC). The area under this curve can be treated as a metric for evaluating the Bayesian model. However, it does not necessarily give a full picture of the multiple modes of the posterior and whether the sampled graph is far away /close to the MEC of the true graph. Nevertheless, it can still be informative of the edge beliefs of the learned approximation. Details of computing this is give in [26].

D.2 Results

D.2.1 Estimation of True Posterior. As indicated before, when the number of variables in the graph is small, we can enumerate the true posterior and compute the divergence/distance between the approximation and the true posterior. Figure 5 presents the Hellinger distance of the approximation up to four nodes. It can be seen that the autoregressive distribution of VCN reconstructs the true posterior much better than the factorised distribution, mainly due to the non-identifiability of graphs and the fact that the true posterior is multimodal.

D.2.2 Evaluation in higher dimensions. Figure 7 reports the $\mathbb{E}[\text{SHD}]$ of the approximation for different number of nodes. It can be seen that the autoregressive distribution of VCNs gives better performance than the factorised distribution in all the settings. In addition, the autoregressive distribution gives better results as compared to all the competitive baselines when there are 10 samples, a regime common in biological applications. When the number of samples are increased to 100, the proposed approach compares favourably to the baselines which do not involve differentiable

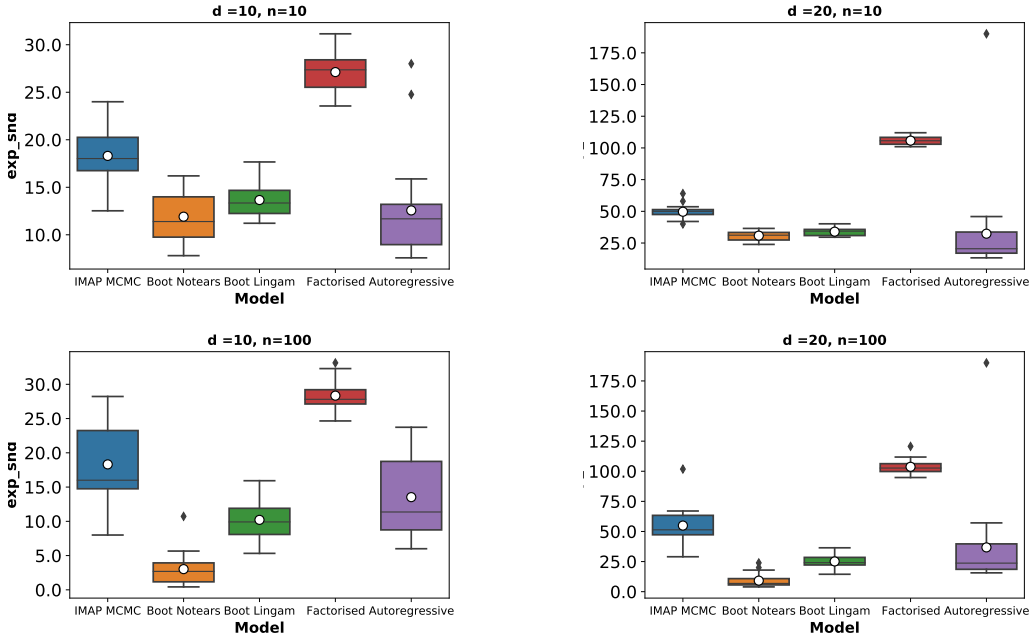

Fig. 7. $\mathbb{E}$ [SHD] for $d = 10$ and $d = 20$ node ER random graphs (lower is better). Results obtained using 20 different random graphs.

learning based techniques, while being much better than the factorised distribution. Figure 8 presents the AUROC of these approaches. The autoregressive distribution of VCN performs better than the learning based method of factorised distribution, while comparing favourably to IMAP MCMC. We found that Boot Notears usually performs well on this metric. However, DAG Bootstrap in general has some limitations. Though DAG Bootstrap can capture multiple modes, its support is limited to the DAGs estimated using the bootstrap procedure and hence does not necessarily have full support. In addition, it can be significantly slower than VCN (Figure 6). Therefore, the proposed approach has a good tradeoff of runtime versus performance against the evaluated metrics and is strictly favourable in terms of purely learning based approaches.

D.2.3 Real Dataset. We evaluate our method in terms of AUROC on edge feature probabilities against all the baselines on the *Dream4* in-silico network challenge on gene regulation. As the typical metric for this challenge is the AUROC, we restrict our focus to this metric on this dataset. In particular, we examine the multifactorial dataset consisting of ten nodes and ten observations. Table 1 reports the AUROC of all the methods including the baselines. As the table suggests, our method achieves favourable performance while still being faster.

Table 1. Results on the Dream4 gene expression dataset in terms of AUROC.

	AUROC
IMAP MCMC	0.438 ± 0.02
Boot Notears	0.330
Boot Lingam	**0.689**
Factorised	0.489 ± 0.01
Autoregressive (VCN)	0.519 ± 0.03

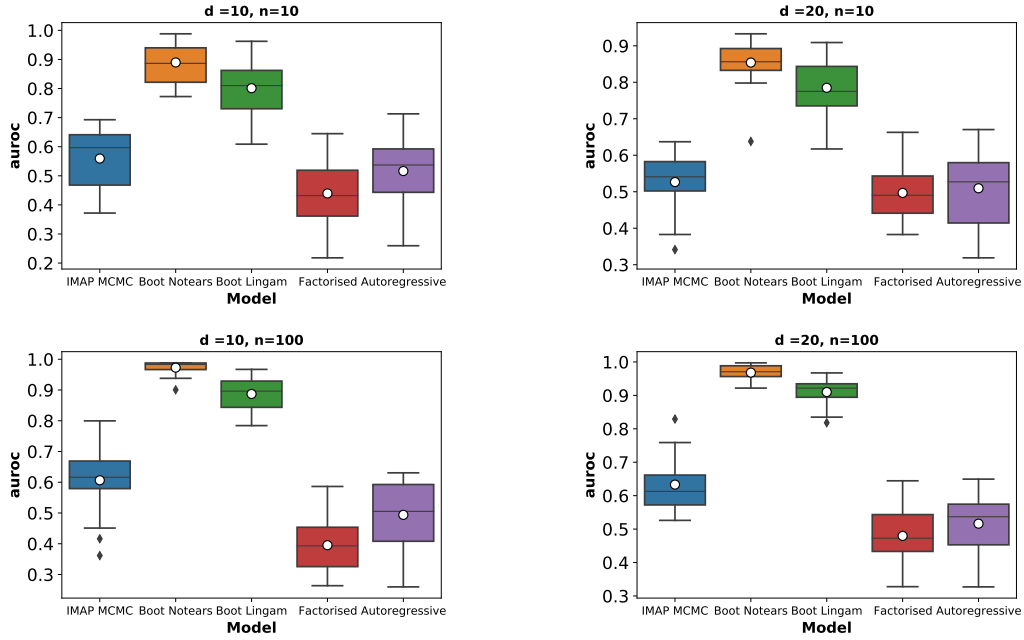

Fig. 8. AUROC for $d = 10$ and $d = 20$ node ER random graphs (higher is better). Results obtained using 20 different random graphs.