

## CAUSALITY IN MACHINE LEARNING

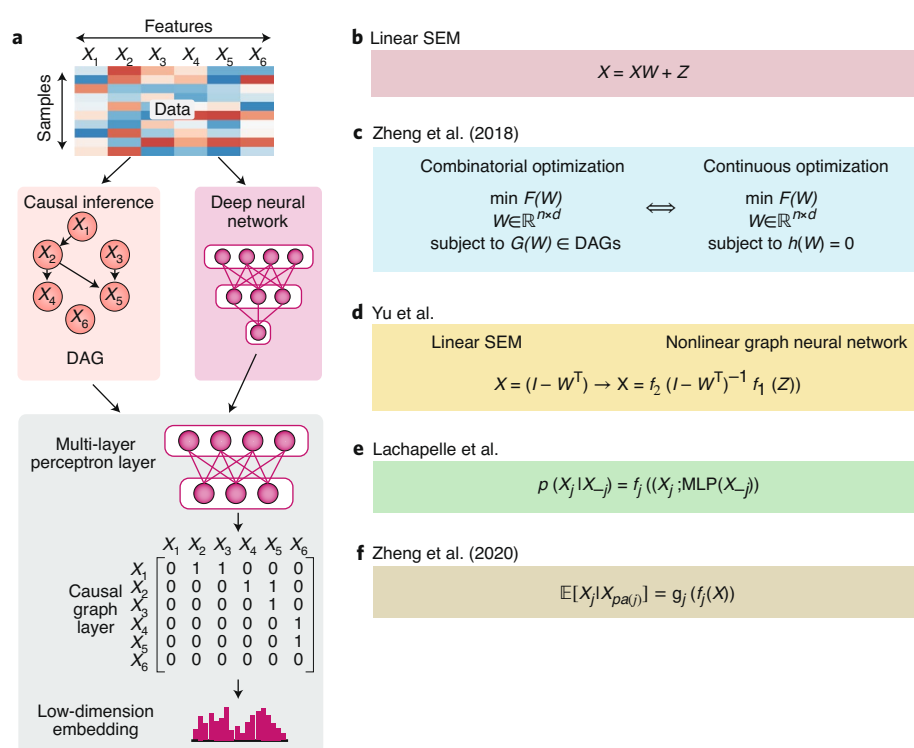
## When causal inference meets deep learning

Bayesian networks can capture causal relations, but learning such a network from data is NP-hard. Recent work has made it possible to approximate this problem as a continuous optimization task that can be solved efficiently with well-established numerical techniques.

Yunan Luo, Jian Peng and Jianzhu Ma

Learning causal relations, rather than correlations, is a fundamental problem in both statistical machine learning and computer sciences. For instance, when studying a complex system or environment, it is natural to ask ‘what would happen to event  $X$  if we change factor  $Y$ ?’ Systematically answering such questions requires deciphering the causal relationships among different components of the system, usually from existing experience or observations (data). One of the most popular causal models, known as Bayesian networks (BN)<sup>1</sup>, encodes the conditional independencies between variables using directed acyclic graphs (DAGs). BN models have been useful for addressing important challenges in several machine learning directions, such as interpretable learning<sup>2</sup> and fairness-aware learning<sup>3</sup>. Exact inference of the DAG structure of BN is computationally intractable, due to the combinatorial explosion in the search space. Most widely used algorithms are approximating a solution by using constraint<sup>4</sup>- or score-based<sup>5</sup> heuristics. A constraint-based method can use background knowledge to infer the direction of causality between variables using statistical tests while a score-based method directs its search through the space of all possible DAGs using a score-function as a heuristic. While these methods can circumvent the enormous search space, they are still computationally costly due to their combinatorial nature and the inefficient search strategies. Building on a recent framework, which established an elegant mathematical connection between the classical combinatorial optimization of searching the space of possible DAG solutions and the continuous optimization of machine learning<sup>6</sup>, Lachapelle et al.<sup>7</sup> describe how the statistical problem can be turned into a pure neural network learning one while Zheng et al.<sup>8</sup> transform it into a general, non-parametric problem that requires no modelling assumptions.

The key innovation of the framework in ref.<sup>6</sup> was to transform the discrete DAG



**Fig. 1 | Causal inference with deep learning.** **a**, Causal inference has been using DAG to describe the dependencies between variables. Deep learning is able to model nonlinear, higher-order dependencies in the data. Leveraging both the effectiveness of deep learning and the interpretability of causal inference is a promising direction. **b**, The linear SEM is a classical model to characterize conditional dependencies between variables, where  $X$  is the data matrix,  $W$  is the adjacency matrix of the DAG and  $Z$  is additive noises. **c**, Zheng et al. (2018)<sup>6</sup> converted the conventional combinatorial problem into a continuous optimization problem. **d**, Yu et al.<sup>10</sup> generalized SEM to a variational autoencoder to model the nonlinear relationships between random variables. **e**, Lachapelle et al.<sup>7</sup> inferred the conditional dependencies from the weights of the neural network. **f**, Zheng et al. (2020)<sup>8</sup> proposed a nonparametric model that can be applied to a variety of parametric and semiparametric models, including generalized linear SEMs, additive models, and index models as special cases.

constraints/space into a smooth function  $h$ , which quantifies the ‘DAG-ness’ of the adjacency matrix of a graph. The original combinatorial acyclic constraints are thus replaced by a continuous equality constraint. The authors parameterized the causal inference using the commonly used linear structural equation model (SEM)<sup>9</sup>. Given a sample-by-feature matrix  $X$ , the

method tries to find a graph that encodes in its adjacency matrix  $W$  the relations between the variables in a self-regression model (Fig. 1b). The authors then further constrained the adjacency matrix to ensure that the graph does not contain cycles (Fig. 1c). This constraint changes the combinatorial nature of the original problem and transforms the problem into


a continuous optimization problem. More importantly, it is now differentiable, so the whole objective function can be optimized by using gradient descent, which is much more efficient compared to previous heuristic algorithms.

Further work in recent years has generalized the linear part of the framework in ref. <sup>6</sup> to a nonlinear model using various deep neural networks<sup>10</sup>. Lachapelle et al. proposed an alternative solution for causal inference with deep learning. Instead of directly inferencing the DAG structure of BN, they chose to learn it from the weights of a neural network while still considering the constraint defined in ref. <sup>6</sup> (Fig. 1e). For each random variable  $X_i$ , the method trains a neural network using other variables to predict this variable. Analysing the weights of the network reveals whether there is a relationship between a variable and  $X_i$ ; if for all paths leading from a variable to  $X_i$ , at least one weight is 0,  $X_i$  does not depend on that variable. The adjacency matrix was optimized together with the acyclic constraint by maximizing the log-likelihood using the augmented Lagrangian method.

Zheng et al.<sup>8</sup> further extended the DAG learning into a nonparametric problem, substantially generalizing the above approaches (Fig. 1f). They proposed a generic formulation of the DAG learning problem as  $E[X_j|X_{\text{parent}(j)}] = g_j(f_j(X))$ , where  $f_j(u_1, u_2, \dots, u_d)$  only depends on variables in the set  $\text{parent}(j)$  while  $g_j$  are typically additive noises. The formulation does not involve the weighted adjacency matrix  $W$ , so the acyclic constraint proposed in ref. <sup>6</sup> cannot be directly applied. The authors addressed this problem by noting that  $f_j$  does not depend on  $X_i$  if and only if its partial derivatives with respect to  $X_i$  are equal to zero. This observation led to a new definition of an adjacency matrix such that the DAG constraints still apply. Since this problem is infinite-dimensional, the authors

approximate it with a finite-dimensional optimization problem with tractable function space.

The hardness of causal inference mainly stems from the intractable combinatorial search space. With the recently developed methods, new types of global solutions are now available that can be scaled to larger sizes of problems where combinatorial algorithms are not. For example, recently it has been shown that these new methods can be scaled to infer causal transcriptome networks with more than 10,000 genes<sup>11</sup>, which cannot be done by combinatorial algorithms in a reasonable timeframe. In addition, it can be expected that prior knowledge would further enhance the utility of this paradigm for causal inference. For instance, it is easy to incorporate domain knowledge such as 'event  $X$  never leads to event  $Y$ ' or 'gene  $A$  enhances the expression of gene  $B$  in most of the tissue types'. From a deep learning perspective, it has been frequently observed that it is hard to extract any explicit structures of the data from the deep neural networks to facilitate the interpretation. Such models, although accurate, provide no meaningful insights about how their decisions are made. Recent studies<sup>12,13</sup> indicated that a certain degree of interpretability could be achieved by properly designing the architecture of the neural network based on domain knowledge. Lachapelle et al. and Zheng et al.<sup>8</sup> directly inferred part of the neural network architecture from data to encode the causality among random variables, which provides model interpretation from another perspective<sup>14</sup>. Currently, most of the neural architecture search (NAS) algorithms optimize the neuron connections with respect to prediction accuracy. The models discussed here provide us with an unprecedented opportunity to develop a new generation of NAS frameworks that could achieve both model accuracy and transparency simultaneously. That is, part

of the architecture could be optimized as a regularization as normal NAS frameworks to improve the model accuracy and another part of the architecture could be optimized with respect to various types of interpretations, such as causality or feature modularity. 

Yunan Luo<sup>1</sup>, Jian Peng<sup>1</sup> and Jianzhu Ma<sup>1,2,3</sup> 

<sup>1</sup>Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA. <sup>2</sup>Department of Computer Science, Purdue University, West Lafayette, IN, USA. <sup>3</sup>Department of Biochemistry, Purdue University, West Lafayette, IN, USA.

 e-mail: majianzhu@purdue.edu

Published online: 12 August 2020  
<https://doi.org/10.1038/s42256-020-0218-x>

## References

1. Heckerman, D., Geiger, D. & Chickering, D. *Mach. Learn.* **20**, 197–243 (1995).
2. Kim, C. & Bastani, O. Preprint at <https://arxiv.org/abs/1901.08576> (2019).
3. Madras, D., Creager, E., Pitassi, T. & Zemel, R. in *Proc. Conf. Fairness, Accountability, and Transparency* 349–358 (ACM, 2019).
4. Meek, C. in *Proc. Eleventh Conf. Uncertainty in Artificial Intelligence* 403–410 (ACM, 1995).
5. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
6. Zheng, X., Aragam, B., Ravikumar, P. K. & Xing, E. P. in *Advances in Neural Information Processing Systems* Vol. 31 (eds. Bengio, S. et al.) 9472–9483 (Curran Associates, 2018).
7. Lachapelle, S., Brouillard, P., Deleu, T. & Lacoste-Julien, S. in *Proc. Eighth Int. Conf. Learning Representations (ICLR, 2020)*.
8. Zheng, X., Dan, C., Aragam, B., Ravikumar, P. & Xing, E. P. in *Proc. Twenty Third Int. Conf. Artificial Intelligence and Statistics* Vol. 108 3414–3425 (PMLR, 2020).
9. Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. *J. Mach. Learn. Res.* **7**, 2003–2030 (2006).
10. Yu, Y., Chen, J., Gao, T. & Yu, M. in *Proc. 36th Int. Conf. Machine Learning* Vol. 97 7154–7163 (PMLR, 2019).
11. Lee, H.-C., Danieletto, M., Miotto, R., Cherg, S. T. & Dudley, J. T. in *Pacific Symp. Biocomputing* Vol. 25 391–402 (PSB, 2020).
12. Ma, J. et al. *Nat. Methods* **15**, 290–298 (2018).
13. Lin, C., Jain, S., Kim, H. & Bar-Joseph, Z. *Nucleic Acids Res.* **45**, e156 (2017).
14. Lipton, Z. C. Preprint at <https://arxiv.org/abs/1606.03490> (2016).

## Competing interests

The authors declare no competing interests.