<Society logo(s) and publication title will appear here.>

DAGMA-DCE: Interpretable, Non-Parametric Differentiable Causal Discovery

Daniel Waxman [1], Graduate Student Member, IEEE, Kurt Butler [1], Graduate Student Member, IEEE, and Petar M. Djurić [1], Life Fellow, IEEE

[1]Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA

Corresponding author: Petar M. Djurić (petar.djuric@stonybrook.edu)

ABSTRACT We introduce DAGMA-DCE, an interpretable and model-agnostic scheme for differentiable causal discovery. Current non- or over-parametric methods in differentiable causal discovery use opaque proxies of "independence" to justify the inclusion or exclusion of a causal relationship. We show theoretically and empirically that these proxies may be arbitrarily different than the actual causal strength. Juxtaposed to existing differentiable causal discovery algorithms, DAGMA-DCE uses an interpretable measure of causal strength to define weighted adjacency matrices. In a number of simulated datasets, we show our method achieves state-of-the-art level performance. We additionally show that DAGMA-DCE allows for principled thresholding and sparsity penalties by domain-experts. The code for our method is available open-source at https://github.com/DanWaxman/DAGMA-DCE, and can easily be adapted to arbitrary differentiable models.

INDEX TERMS Causal discovery, interpretable machine learning, differential causal effect

I. INTRODUCTION

The discovery of causal relationships is the subject of many scientific questions, in medicine, economics, the social sciences, and more, and is therefore a question of utmost importance in science and engineering. Given an observational signal, one may therefore be interested in learning the causal structure underlying the data. Under Pearl's formalism of causality [1], the causal structure is captured by a *causal graph*, which is a directed acyclic graph (DAG) with edges denoting direct causation of one variable to another.

A core principle of causal inference is the notion of *intervention*, where one can interact with the system that produces signals. This is in contrast to causal inference from *observational* data, where we can only observe the outputs of a system, as is typically the case in signal processing or machine learning. Unfortunately, causal discovery using observational data is generally ill-posed and requires strong assumptions about the system we observe [2, pp. 135]. Different methods may use different assumptions, where the validity of the assumptions may be extremely domain- or problem-specific. Even when assumptions are made to ensure

(at least partial) identifiability of the causal graph, causal discovery can be extremely expensive, in terms of both the computation and the amount of required data.

One major class of methods, referred to as *constraint-based* methods, is based on the observation that a causal graph encodes a set of conditional independencies (CIs). By assuming that all CIs are encoded in the graph, an assumption known as *causal faithfulness*, constraint-based methods utilize statistical CI tests to determine which causal graphs are consistent with the data.

While some of the earliest successful causal discovery algorithms are constrained-based, such as the PC Algorithm and Fast Causal Inference [3], they can be very costly. Namely, they are both expensive computationally, scaling exponentially with the number of variables in the worst case, and they are expensive in terms of data. Additionally, CI tests tend not to be very powerful, and constrained-based causal discovery methods require conditional data for many different values and combinations of random variables. Moreover, causal faithfulness is not a testable assumption in general, though some weaker versions are [4].

An alternative approach to causal discovery using observational data is to use *score-based* methods. These methods attempt to find a causal model $\mathcal{M}$ and its associated causal graph $\mathcal{G}$, which maximizes a score function $\mathcal{S}(\mathcal{M}, \mathcal{D})$ for observational data $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ [2, pp. 148]. For example, one can search for the causal model which maximizes the likelihood or minimizes the mean square error (MSE). In practice, one typically optimizes a version of these objectives which is penalized for model complexity, such as the Bayesian information criterion. For certain classes of causal models, this approach is simple and effective — indeed, early attempts at score-based causal discovery were made contemporaneously to early constraint-based approaches [5]. Despite their simplicity, score-based methods suffer from the super-exponential search space of DAGs with d variables, as well as identifiability issues [2, pp. 149].

Fundamentally troubling to score-based methods is the difficulty of optimization in discrete spaces, which tend to be combinatorially more difficult than continuous objectives. To address this, NOTEARS (short for *Non-combinatorial **O**ptimization via **T**race **E**xponential and **A**ugmented lag**R**angian for Structure learning*) [6] proposes a novel algebraic constraint on adjacency matrices, such that the constraint is satisfied if and only if the adjacency matrix $\mathbf{A}$ represents a DAG. NOTEARS then performs constrained continuous optimization on a penalized MSE objective to learn a causal graph without explicitly exploring the discrete space of all DAGs.

NOTEARS was originally defined for linear models, with the adjacency matrix entries being coefficients of the corresponding linear functions. NOTEARS+ [7] extends NOTEARS to nonlinear functions and achieves state-of-the-art performance on simulated data. In lieu of linear coefficients, NOTEARS+ uses opaque proxies to define $\mathbf{A}$, providing advantages in optimization but at the cost of interpretability and generality of the results.

Several methods have provided alternative acyclicity constraints which are faster to compute or are otherwise preferable in optimization. One striking example is DAGMA (short for ***D**AGs via **M**-matrices with **A**cyclicity*) [8], which in practice can optimize up to an order of magnitude faster than NOTEARS.

Our contribution is to provide a novel learning objective that results in obtaining interpretable weighted causal graphs. Our method has several important strengths. It is model-agnostic in nature, with the ability to use any twice-differentiable model trainable with gradient descent. Furthermore, the sparsity penalty used in the scoring function $\mathcal{S}(\cdot, \cdot)$ is also model-agnostic. Lastly, we can interpret the causal graphs with the differentiable causal effect (DCE) metric [9]. We base our optimization on the method of DAGMA, leading us to name our method DAGMA-DCE.

The rest of this paper is structured as follows: in Section II, we review score-based causal discovery and the current state-of-the-art in differentiable causal discovery. In Section III, we show that the adjacency matrix returned by NOTEARS and DAGMA is not interpretable, first by a theoretical argument and then by a toy example. We propose a new graph definition and optimization problem in Section IV, forming DAGMA-DCE, and discuss its interpretability in Section V. Finally, we run experiments on synthetic data in Section VI showing that DAGMA-DCE performs competitively with other methods and returns legitimately different adjacency matrices than DAGMA. Concluding remarks are provided in Section VII.

II. BACKGROUND

We begin by reviewing some concepts of causal inference, namely structural causal models (SCMs), which formalize causality. We then introduce differentiable causal discovery methods, which utilize continuous optimization for score-based causal discovery.

A. Structural Causal Models and Causal Discovery

SCMs provide a formalism for causality, where causal relationships are described by a DAG and accompanying functional assignments. To make this more clear, let $\mathbf{x} = [x_1, \ldots, x_d]$ be d-dimensional random vector, and $\mathcal{G}$ be a DAG whose vertices are the components of $\mathbf{x}$. Then, according to $\mathcal{G}$, an SCM describes each component x_j as a function of its parents $\mathrm{Pa}(x_j)$ and independent noise z_j, $f_j(\mathrm{Pa}(x_j), z_j)$, i.e.,

$$x_j := f_j\left(\mathrm{Pa}(x_j), z_j\right), \quad j = 1, \ldots, d. \tag{1}$$

We use the notation $:=$ to reinforce the somewhat philosophical point that these are *assignments*, and not merely equivalence [2]. We will frequently refer to the SCM of a system as simply the "causal model" $\mathcal{M}$.

The goal of structure learning is then to identify the SCM of a system, given data $\mathcal{D}$. If one cannot interact with the system, then we must make assumptions on the form of each f_j and the distribution of the noise z_j to uniquely determine the SCM [2, pp. 139]. These assumptions can be parametric, for example, assuming each f_j is linear and each z_j is non-Gaussian [10], or they can be non-parametric, for example, assuming the f_j are strictly nonlinear with the z_j being additive and Gaussian [11, Corollary 31]. Once we make such a restriction to our model, we call it *identifiable*.

After establishing identifiability, there are several families of techniques for learning the SCM from observational data. Here, we focus on score-based methods, though the interested reader is pointed to [12] for a recent, comprehensive review. In score-based causal discovery, we find a causal model $\mathcal{M}$ which maximizes a score function $\mathcal{S}(\mathcal{M}, \mathcal{D})$. In practice, we parameterize models by some vector $\boldsymbol{\theta}$ and then find the model $\mathcal{M}_{\boldsymbol{\theta}}$ which optimizes $\mathcal{S}(\cdot, \cdot)$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{S}(\mathcal{M}_{\boldsymbol{\theta}}, \mathcal{D}) \text{ s.t. } \mathcal{G}_{\boldsymbol{\theta}} \text{ is a DAG}. \tag{2}$$

The main drawback of naïve score-based methods is that they have to search over the space of DAGs with d nodes, which grows extremely quickly. For $d = 5$, this number

is already 29,281; for $d = 14$, it is already greater than 10^{36} [13, A003024]. Therefore, for even small graphs, one must resort to greedy or approximate techniques [2, pp. 150].

B. Differentiable Causal Discovery

Differentiable causal discovery methods are score-based methods that recast Eq. (2) as a constrained continuous optimization problem; as before, let $\mathcal{S}(\cdot, \cdot)$ be a score function that reflects how well a model $\mathcal{M}_{\boldsymbol{\theta}}$ explains the data $\mathcal{D}$, and $\mathbf{A}_{\boldsymbol{\theta}} \in \mathbb{R}_{\geq 0}^{d \times d}$ be a weighted adjacency matrix derived from $\mathcal{M}_{\boldsymbol{\theta}}$. Then, using a function $h(\cdot)$, which is 0 if and only if $\mathbf{A}_{\boldsymbol{\theta}}$ represents a DAG, the optimization problem Eq. (2) becomes

$$\begin{aligned} \max_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \quad & \mathcal{S}(\mathcal{M}_{\boldsymbol{\theta}}, \mathcal{D}), \\ \text{s.t.} \quad & h(\mathbf{A}_{\boldsymbol{\theta}}) = 0. \end{aligned} \quad (3)$$

The earliest differentiable causal discovery algorithm with a DAG constraint is NOTEARS [6], which uses a trace of a nonlinear function on the adjacency matrix $\mathbf{A}_{\boldsymbol{\theta}}$, i.e.,

$$h_{\text{NOTEARS}}(\mathbf{A}_{\boldsymbol{\theta}}) = \text{tr}\left(\exp(\mathbf{A}_{\boldsymbol{\theta}} \odot \mathbf{A}_{\boldsymbol{\theta}})\right) - d,$$

where $\odot$ is the Hadamard (elementwise) product. The motivation for this constraint is that powers of the adjacency matrix $\mathbf{A}_{\boldsymbol{\theta}}^k$ reflect how many closed walks of length k are present in the corresponding graph $\mathcal{G}_{\boldsymbol{\theta}}$. Since DAGs have no closed walks of any length, $h(\mathbf{A}_{\boldsymbol{\theta}})$ is 0 precisely when $\mathbf{A}_{\boldsymbol{\theta}}$ is a DAG.

Quickly, several simplifications on the same constraint were derived. For example, one need not calculate the entire matrix exponential, with a polynomial sufficing instead [14].

The alternative constraint that we use is DAGMA [8]; with $s > 0$, the DAGMA constraint function is

$$h_{\text{DAGMA}}(\mathbf{A}_{\boldsymbol{\theta}}) = -\log\det\left(s\mathbb{1}_d - \mathbf{A}_{\boldsymbol{\theta}} \odot \mathbf{A}_{\boldsymbol{\theta}}\right) + d\log s, \quad (4)$$

where $\mathbb{1}_d$ is the $d \times d$ identity matrix. This constraint is motivated by the fact that $\mathbf{A}_{\boldsymbol{\theta}}$ describes a DAG if and only if it is nilpotent. While the constraint given by Eq. (4) is, in general, a relaxation of nilpotency, the authors of DAGMA characterize the set of matrices for which zeroness of Eq. (4) is equivalent to acyclicity, called $\mathbf{M}$-matrices. They then show that the set of $\mathbf{M}$-matrices includes all DAGs, and that the gradient of $h_{\text{DAGMA}}(\cdot)$ is sufficiently well-behaved, to enable constrained optimization within $\mathbf{M}$-matrices.

It remains to specify the model $\mathcal{M}_{\boldsymbol{\theta}}$, the score function $\mathcal{S}$, and how to derive $\mathcal{G}_{\boldsymbol{\theta}}$ from $\mathcal{M}_{\boldsymbol{\theta}}$. The approach introduced in NOTEARS+ and used in DAGMA[1] is to define

$$\mathbf{A}_{\boldsymbol{\theta} ij} \triangleq \|\partial_i f_j\|_{L^2} = \sqrt{\int_{\mathbb{R}^d} (\partial_i f_j(\mathbf{x}))^2 \, d\mathbf{x}}, \quad (5)$$

where f_j is the function $\mathbf{x} \mapsto x_j$ according to $\mathcal{M}_{\boldsymbol{\theta}}$. To guarantee Eq. (5) is well-defined, it is required that $\|\partial_i f_j\|_{L^2}$ is finite. Accordingly, the NOTEARS+ authors restrict $\mathcal{M}_{\boldsymbol{\theta}}$ to a space of functions called the Sobolev space $H^1(\mathbb{R}^d)$.

However, the NOTEARS+ authors argue that for causal discovery, the zeroness of $\mathbf{A}_{\boldsymbol{\theta} ij}$ is what matters, and so a condition equivalent to $\mathbf{A}_{\boldsymbol{\theta} ij} = 0$ is derived for several classes of models. For example, in a multilayer perception (MLP) network, the L^2 norm of the first layer in the neural network is chosen. The score function is defined to be a sparsity-penalized MSE[2],

$$\mathcal{S}(\mathcal{M}_{\boldsymbol{\theta}}, \mathcal{D}) = -(\text{MSE}(\mathcal{M}_{\boldsymbol{\theta}}, \mathcal{D}) + \lambda_1 g(\mathcal{M}_{\boldsymbol{\theta}})), \quad (6)$$

where λ_1 is a hyperparameter and $g(\mathcal{M}_{\boldsymbol{\theta}})$ defines a sparsity penalty. For MLPs, NOTEARS+ and DAGMA use the ℓ_1 norm of the first layer MLP weights.

In practice, it is very difficult to exactly follow the constraint in Eq. (3). For this reason, as well as to avoid false positives, an additional thresholding step is performed, where $\mathbf{A}_{ij}$ is set to 0 if it is less than some predetermined hyperparameter C.

III. NON-INTERPRETABILITY OF DAGMA

A key motivation for DAGMA-DCE is that existing methods utilize a non-interpretable weighted adjacency matrix $\mathbf{A}_{\boldsymbol{\theta}}$. In this section, we argue theoretically that the adjacency matrix of DAGMA is not interpretable and subsequently show empirically that DAGMA fails to return intuitively-interpretable results on linear SCMs.

For the purposes of this section, we restrict ourselves to models where f_j is modeled with an MLP network, though our arguments can be adjusted for other cases. The theoretical section can safely be skipped, but shows explicitly that adjacency matrices derived from the first layer of a neural network may be quite different than those defined by derivatives.

A. Theoretical Argument

For the purposes of this section, let $f_j \colon \mathbb{R}^d \to \mathbb{R}$ be an MLP with weight matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(M)}$ and activation function $\sigma(\cdot)$. Then the $(M-1)$-hidden layer MLP f_j can be expressed as

$$f_j(\mathbf{x}) = \mathbf{A}^{(M)} \sigma\left(\mathbf{A}^{(M-1)} \sigma\left(\cdots \sigma\left(\mathbf{A}^{(1)} \mathbf{x}\right)\right)\right). \quad (7)$$

Despite the exact analytic result that

$$\|\partial_i f_j\|_{L^2} = 0 \iff \|\mathbf{A}_{i\cdot}^{(1)}\|_{L^2} = 0, \quad (8)$$

differentiable causal discovery methods rely on thresholding small, nonzero values. The following lemma shows that when using sigmoid activation functions, as is done in NOTEARS+ and DAGMA, the two norms in Eq. (8) may be arbitrarily different.

Lemma 1. *Let $\sigma(\cdot)$ denote the sigmoid activation function. Then for any $\delta, \epsilon > 0$, there exists an MLP f_j with weight matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(M)}$ such that $\|\mathbf{A}_{i\cdot}^{(1)}\|_{L^2} < \epsilon$ but $\|\partial_i f_j\|_{L^2} > \delta$.*

---

[1]In general, DAGMA refers to both linear and nonlinear models using Eq. (4). To avoid confusion, DAGMA exclusively refers to the nonlinear case here.

[2]DAGMA uses the log-likelihood instead for benchmarks, but we use the MSE to simplify comparisons, and avoid assumptions on the data-generating process.

Proof:

As (7) is a repeated composition of smooth functions, it suffices to show the result for the 1-hidden layer MLP

$$g(\mathbf{x}) = \mathbf{A}^{(2)} \sigma\left(\mathbf{A}^{(1)}\mathbf{x}\right).$$

If $\mathbf{A}^{(2)}$ is a $K \times H$ matrix, then

$$[g(x)]_k = \sum_h^H \mathbf{A}_{kh}^{(2)} \sigma\left(\sum_{j=1}^d \mathbf{A}_{hj}^{(1)} x_j\right).$$

For notational clarity, let

$$z_h = \sigma\left(\sum_{j=1}^d \mathbf{A}_{hj}^{(1)} x_j\right).$$

Additionally, note that $\sigma'(\cdot) > 0$. Then so long as $\mathbf{A}_{hi}^{(1)}$ is nonzero for some h, we have by chain rule that

$$\left|\frac{\partial z_h}{\partial x_i}\right| > \epsilon'$$

for some $\epsilon' > 0$. Then by directly calculating the norm,

$$\|\partial_i f_j\|_{L^2}^2 = \int \sum_{k=1}^K \left(\frac{\partial g_k}{\partial x_i}\right)^2 dx_i$$

$$= \int \sum_{k=1}^K \left(\sum_{h=1}^H \mathbf{A}_{kh}^{(2)} \frac{\partial z_h}{\partial x_i}\right)^2 dx_i$$

$$\geq \int \sum_{k=1}^K \left(\sum_{h=1}^H \mathbf{A}_{kh}^{(2)} \epsilon'\right)^2 dx_i.$$

we conclude that it is unbounded. Therefore, we can choose $\mathbf{A}^{(2)}$ such that the function norm is arbitrarily large, and in particular, larger than δ. ∎

While this result may not be particularly "generic," in the sense that the resulting function f_j may be unlikely to be learned in practice, it does show that $\partial_i f_j$ can be arbitrarily different than any nonzero $\mathbf{A}^{(1)}$. This situation only gets worse for the more popular rectified linear unit (ReLU) activation function, where $\|\partial_i f_j\|_{L^2}$ is arbitrary for a *fixed* f_j. This is formalized in the following lemma:

Lemma 2. *Let $\sigma(\cdot)$ denote the ReLU activation function. Then for any $s > 0$ and $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(M)}$, there exist matrices $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(M)}$ such that*

$$f_j(\mathbf{x}) = \mathbf{A}^{(M)} \sigma\left(\mathbf{A}^{(M-1)} \sigma\left(\cdots \sigma\left(\mathbf{A}^{(1)}\mathbf{x}\right)\right)\right)$$

$$= \mathbf{B}^{(M)} \sigma\left(\mathbf{B}^{(M-1)} \sigma\left(\cdots \sigma\left(\mathbf{B}^{(1)}\mathbf{x}\right)\right)\right),$$

and $\|\mathbf{B}_{i\cdot}^{(1)}\|_{L^2} = s$.

Proof:

Note that for any real z, the ReLU function is invariant to positive rescaling: $\sigma(z) = \sigma(az)/a$ for any $a > 0$. This implies, more generally, that $\mathbf{S}^{-1}\mathbf{A}^{(2)}\sigma(\mathbf{S}\mathbf{A}^{(1)}\mathbf{x}) = \mathbf{A}^{(2)}\sigma(\mathbf{A}^{(1)}\mathbf{x})$, where $\mathbf{S}$ is a diagonal matrix [15]. The lemma then immediately follows by setting $\mathbf{B}^{(1)} = \mathbf{S}\mathbf{A}^{(1)}$ and $\mathbf{B}^{(2)} = \mathbf{S}^{-1}\mathbf{A}^{(2)}$, with the appropriate choice of $\mathbf{S}$. ∎

B. Empirical Result with a Linear SCM

In the lemmas above, we have shown that the adjacency matrix of NOTEARS+ and DAGMA may be arbitrarily different than the L^2 function norm. Here, we illustrate this fact in practice by applying DAGMA to a linear structural equation model (SEM), and showing that the returned adjacency matrix does not correspond to the intuitive notion of causal strength.

For this purpose, we simulated an Erdös-Rényi (ER) graph with 10 random variables and an expectation of 20 edges, creating a DAG as in [6]. Given a DAG, a SEM assigns functional relationships between a variable and its parents. In this case, a linear SEM is assigned by randomly sampling linear coefficients from $\mathcal{U}\left((-2.0, -0.5) \cup (0.5, 2.0)\right)$.

The use of a linear SEM allows for an obvious ground-truth weighted graph to compare to, namely the linear coefficients. We then ran DAGMA with each f_j being modeled as an MLP with one hidden layer of 10 units and compared the resulting weighted adjacency matrix to the ground truth in Fig. 1a. For the purposes of this demonstration, we tested with $\lambda_1 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and chose the value which minimized the Frobenius norm $\|\mathbf{A}_{\text{est}} - \mathbf{A}_{\text{true}}\|_F$ to give a favorable comparison. This resulted in using $\lambda_1 = 10^{-2}$, with all other hyperparameters set to their default values.

Two things are visually apparent: first, DAGMA learns the graph topology quite well, including one erroneous edge in an otherwise perfect prediction. Second, however, is that the weights recovered by DAGMA do not obviously correspond to the interpretable, linear coefficients, i.e., it cannot easily be interpreted as a "strength of causation." Even worse, it is not easy to predict when DAGMA is over- or under-estimating the strength of an edge. While we defer a definition of DAGMA-DCE and discussion of Fig. 1b to the subsequent sections, a view of Fig. 1b shows a much closer resemblance to linear coefficients. It is important to note that this does not indicate "errors" in DAGMA, but instead that the magnitude of the weighted graph *is not defined* to be the true "causal strength" — their very definition has lost meaning compared to the linear case.

IV. PROPOSED SOLUTION

In this section, we provide a definition of the learning objective for DAGMA-DCE, and subsequently discuss practical details of implementing such a scheme. In particular, we discuss how DAGMA-DCE is model-agnostic in both formulation and implementation, and we also explore simple methods for accelerating and pre-training models.

A. Definition of the Optimization Problem

An initial solution would be to estimate Eq. (5) directly; however, we argue $H^1(\mathbb{R}^d)$ is the wrong space to define models over. Not only are many interesting functions outside of $H^1(\mathbb{R}^d)$, but also simple functions, for example linear functions, do not have a finite Sobolev norm over all of $\mathbb{R}^d$. Additionally, from a data-centric perspective, there is little

<Society logo(s) and publication title will appear here.>
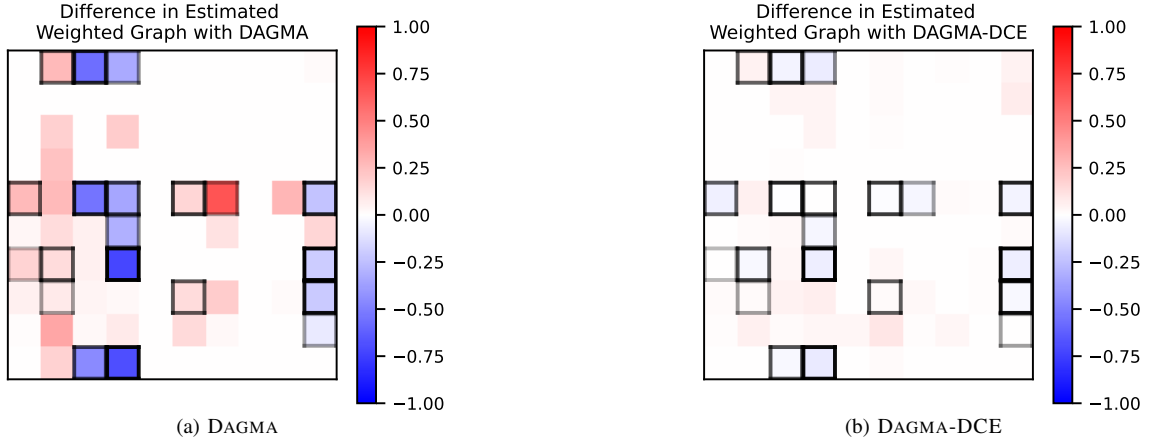

(a) DAGMA

(b) DAGMA-DCE

FIGURE 1: The difference between the magnitude of the true derivatives in a linear causal model to the magnitude of the weighted graph in DAGMA for a random 10×10 Erdös-Rényi directed graph with 20 expected edges. Gray boxes surrounding each cell denote the magnitude of the ground-truth linear coefficient.

reason to care about the behavior of the model where there is a vanishingly small probability mass.

Instead, we choose to define the adjacency matrix as the L^2 norm of the derivative $\partial_i f_j$ with respect to the probability distribution of $\mathbf{X}$, which we denote $\mathbb{P}^{\mathbf{X}}$. While the theory of Sobolev spaces plays no further role in our work, to adapt the language of [7], models are defined over the *weighted Sobolev space* $H^1(\mathbb{R}^d, \mathbb{P}^{\mathbf{X}})$ [16]. This is simply to say that the models' function and first (weak) derivatives are square-integrable with respect to $\mathbb{P}^{\mathbf{X}}$. We then offer the alternative definition and Monte Carlo approximation

$$\mathbf{A}_{ij} \triangleq \|\partial_i f_j\|_{L^2(\mathbb{P}^{\mathbf{x}})} = \sqrt{\int_{\mathbb{R}^d} (\partial_i f_j(\mathbf{x}))^2 \, p_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}}$$

$$\approx \sqrt{\frac{1}{N} \sum_{n=1}^N (\partial_i f_j(\mathbf{x}_n))^2}, \qquad (9)$$

where $p_{\mathbf{X}}(\mathbf{x})$ is the density of $\mathbb{P}^{\mathbf{X}}$ with respect to the Lebesgue measure, and $\mathbf{x}_n \sim p_{\mathbf{X}}(\mathbf{x})$. By noting that the entries of $\mathbf{A}$ are the root-mean-square entries of the Jacobian matrix, it is then straightforward to calculate $\mathbf{A}_{ij}$ for any differentiable model, with the above being a Monte Carlo approximation over empirical Jacobian matrices.

This definition of $\mathbf{A}$ also allows for a principled sparsity penalty — indeed, if $\partial_i f_j$ is square-integrable with respect to $\mathbb{P}^{\mathbf{X}}$, then Jensen's inequality implies that it is also integrable, which allows us to define and approximate

$$g(\mathcal{M}_{\boldsymbol{\theta}}) = \|\partial_i f_j\|_{L^1(\mathbb{P}^{\mathbf{x}})} \approx \frac{1}{N} \sum_{n=1}^N |\partial_i f_j(\mathbf{x}_n)|. \qquad (10)$$

Combining Eqs. (3), (4), (6), (9) and (10), we arrive at the optimization problem

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \quad & \mathrm{MSE}(\mathcal{M}_{\boldsymbol{\theta}}, \mathcal{D}) + \lambda_1 \|\partial_i f_j\|_{L^1(\mathbb{P}^{\mathbf{x}})} \\ \mathrm{s.t.} \quad & -\log \det \left(s \mathbb{1}_d - \mathbf{A}_{\boldsymbol{\theta}} \odot \mathbf{A}_{\boldsymbol{\theta}} \right) + d \log s = 0. \end{aligned} \qquad (11)$$

Similar to DAGMA, we use a central path optimization technique to solve the constrained optimization problem.

Our approach has the same convergence properties as NOTEARS and DAGMA, using similar assumptions. Namely, we assume causal sufficiency, i.e., a lack of latent confounders, and identifiability, which ensures that the optima of the score are the true causal graph. There are many options for ensuring identifiability, for example, thrice-differentiable strictly-nonlinear additive noise models suffice [11]. Additionally, while it is not necessary for the optimization problem to be well-defined, the interpretability of DAGMA-DCE relies on once-differentiability of the underlying f_j. Namely, this excludes discrete or mixed-type data, since how a model interpolates functions between discrete labels is arbitrary.

B. DAGMA-DCE *is Model Agnostic*

The derivatives in Eq. (9) may be available analytically for many models. For the simplest case, take $\mathcal{M}_{\boldsymbol{\theta}}$ to be linear, i.e.,

$$x_j := \sum_{i=1}^d \beta_{ij} x_i.$$

Then the linear coefficients β_{ij} are the derivatives used in Eq. (9). Note that Eq. (11) then becomes the problem of linear DAGMA, meaning DAGMA-DCE is a legitimate generalization.

Simple (semi-)analytic expressions are also available for additive models:

$$x_j := \sum_{k=1}^K \beta_{jk} g_k(\mathbf{x}), \qquad (12)$$

in which case

$$\partial_i f_j = \sum_{k=1}^K \beta_{jk} \frac{\partial g_k}{\partial x_i}.$$

Analytic expressions of coefficients are even available for several classes of non-parametric models. For example, by the representer theorem [17, pp. 132], a Gaussian process (GP) posterior can be written in the form of Eq. (12), where $g(\mathbf{x})$ depends on the kernel $\kappa(\cdot, \cdot)$ and $\beta_{j1}, \ldots, \beta_{jn}$ are easily calculated:

$$
\begin{aligned}
f_j &\sim \mathcal{GP}(\mathbf{m}, \kappa); \\
x_j &\coloneqq f_j(\mathbf{x}),
\end{aligned}
\quad \Longleftrightarrow \quad x_j \coloneqq \sum_{n=1}^{N} \beta_n \kappa(\mathbf{x}, \mathbf{x}_n). \quad (13)
$$

Closed-form expressions for several common kernels, including the automatic relevance detection squared exponential (ARD-SE) and Matérn-3/2 kernels, are derived in [9].

Furthermore, even when analytic expressions become cumbersome to derive by hand, the derivative can often be obtained via automatic differentiation (AD) tools. Among many other differentiable models, a notable case is with MLPs. It is also the case for extensions of NOTEARS+-like objectives to time series, for example the one-dimensional convolutional neural networks used in NTS-NOTEARS [18].

C. Practical Considerations

Several general comments are made in Section 4.1 of [8] which also hold for DAGMA-DCE, particularly concerning the optimization details. For example, the hyperparameter s can be changed freely, even between optimization iterations. This can affect performance, as s controls the volume of the feasible region. The authors of [8] recommend starting with $s = 1$ and slowly reducing its value. Additionally, different optimizers can be used to solve sub-problems along the central path; DAGMA uses Adam [19], which we also use. Finally, the use of central path methods also creates additional hyperparameters, including central path coefficients and weight initializations, which are discussed.

While backward passes over the Jacobian is indeed expensive and a drawback of the DAGMA-DCE, the Monte Carlo approximation can be effectively parallelized with respect to N, particularly on GPUs. Therefore, while scaling with d is inherently difficult, scaling with N is quite inexpensive.

In practice, we additionally see benefits from pre-training for DAGMA-DCE, namely by training a single iteration of DAGMA. Intuitively, while DAGMA and DAGMA-DCE are fundamentally different optimization problems, their underlying task is the same. Therefore, a single iteration of DAGMA serves as a good starting point, and saves costly backward passes over the Jacobian.

Finally, DAGMA-DCE has several hyperparameters, including the strength of L_1 regularization λ_1 and thresholds. In the next section, we discuss how the formulation of DAGMA-DCE makes the threshold hyperparameter interpretable. However, the issue of setting λ_1 remains. One option, explored in [7] and [18], is to optimize λ_1 with respect to some metric using cross-validation if labeled data/causal graph pairs exist. How to choose λ_1 when these labeled pairs are not available is an interesting task for future research.

V. INTERPRETABILITY OF DAGMA-DCE

The interpretability of the DAGMA-DCE approach stems from the observation that the partial derivatives, $\partial_i f_j$, can be interpreted through the lens of causal strength. We begin this section with general notes on causal strengths, followed by the differential causal effect, which forms the basis of interpretability for DAGMA-DCE. We revisit the empirical example of Section III-B and show that DAGMA-DCE recovers the intuitive result. Finally, we discuss how thresholding in DAGMA-DCE is a fundamentally interpretable operation, allowing the incorporation of prior knowledge from experts and decision-makers.

A. Measuring Causal Strength

The quantification of causal strength is a highly nontrivial task, with several competing definitions [20]. Despite this, we find it helpful to ground our reasoning in the linear case, where an unambiguous notion of causal strength is available. In particular, the linear model coefficients quantify exactly how a change in one variable will induce changes in its children. Consider a linear SCM with coefficients given by a matrix $[\mathbf{B}]_{ij} = \beta_{ij}$ and additive noise z_j:

$$
x_j \coloneqq \sum_{i=1}^{d} \beta_{ij} x_i + z_j, \qquad j = 1, \ldots, d.
$$

Then the entry β_{ij} represents the direct influence of x_i to x_j. Note that $\mathbf{B}$ can be viewed as a causal graph, provided that it is acyclic, or equivalently, nilpotent.

Occasionally, these direct coefficients summarize all causal information that x_i holds about x_j. Indeed, if $\mathbf{B}$ is viewed as a graph, this is the case if the only path x_i and x_j is the direct one, in which case $\mathbf{B}_{ij}$ is the total causal strength. Otherwise, in general, we must sum across all paths in the graph, and multiply the coefficients along each path. This can be expressed compactly by a geometric sum,

$$
\mathbf{T} = \sum_{\ell=1}^{d} \mathbf{B}^{\ell}.
$$

Causal sufficiency of the system then tells us that the entries $\mathbf{T}_{ij}$ of the matrix $\mathbf{T}$ are sufficient to describe how any change to x_i will affect x_j, and thus, $\mathbf{T}$ can be useful to ask counterfactual questions or to understand the impact of various causes to a single effect.

B. Differential Causal Effect and DAGMA-DCE

The DCE provides a nonlinear version of the linear model causal strength by considering the local behavior of the model [9]. The Jacobian matrix, $\mathcal{J}_f$, replaces the role of $\mathbf{B}$, describing how changes to a parent affect the child node when there are no other paths in the system. Similarly, the geometric sum of the Jacobian also defines the total causal effect from source to effect, for small changes.

Relating to DAGMA-DCE, the adjacency matrix defined in (9) can be interpreted as the root-mean-squared DCE — as a result, $\mathbf{A}_{ij} \neq 0$ if and only if the DCE, $\partial_i f_j$, is not

<Society logo(s) and publication title will appear here.>

identically zero. The magnitude of $\mathbf{A}_{ij}$ quantifies how strong the interaction is, in terms of the energy of the DCE functions.

C. Empirical Results with a Linear SCM

We quickly revisit the experiment of Section III-B, where DAGMA was fit to a linear SCM, and its weighted adjacency matrix compared to the ground truth. The setup is nearly identical, with the same parameterization of each function f_j. The main difference is that we do not tune the sparsity penalty hyperparameter, instead we simply set $\lambda_1 = 0$. Like DAGMA, the DAGMA-DCE method obtains an accurate estimate of the graph topology, recovering the underlying graph perfectly. Unlike DAGMA, we additionally recover the "intuitive" notion of causal strength.

D. Thresholding in DAGMA-DCE

An ad-hoc component of NOTEARS+-like methodologies is the thresholding step. Indeed, across different datasets, the thresholding hyperparameter can have large consequences to accuracy, as shown in [7]. However, many conceivable causal inference tasks will not have any validation set to calibrate the thresholding parameter with, meaning the user will have to make a heuristic estimate.

On the contrary, DAGMA-DCE's weighted adjacency matrix entries are interpretable, so the choice of a threshold carries direct meaning to the modeler or expert. This decision can even be made on an edge-by-edge basis, or across (a possibly rescaled version of) $\mathbf{A}$.

For example, one may decide to normalize the entry A_{ij} of by the marginal variance of x_j for the purposes of thresholding, reminiscent of ANOVA [21]. An alternative scheme could be normalizing by the sum of the jth column; while subadditivity of norms means this does not actually scale by the total derivative ∂f_j, it can be viewed as scaling by the total causal effect. Finally, since the entries of $\mathbf{A}$ are now unconstrained, one could attempt thresholding methods developed for Granger causality, for example via clustering [22] or denoising [23]. How this thresholding step should be interpreted and executed forms an interesting question for future work.

VI. EXPERIMENTS

We perform several experiments similar to those conducted by the authors of DAGMA [8], providing comparisons to DAGMA and NOTEARS+. We begin with a brief discussion of metrics used to evaluate results, followed by experimental details, results, and discussion.

A. Metrics

Evaluating results in causal discovery is complex in the sense that several different metrics exist with different goals and calculations. The two most basic metrics we report are the structural Hamming distance (SHD), which calculates the number of incorrect edges a graph contains, and the F_1 score, which is the geometric mean of precision and recall.

Additionally, we include the structural intervention distance (SID) [24], which is arguably more well-suited to causal inference. The SID is motivated by the idea that, when calculating causal effects via "adjustment sets," some errors are more severe than others. For example, including spurious relations does not affect causal effects, while reversing the order of an edge does. The SID calculates the number of edges $i \to j$ such that the adjustment set corresponding to the calculation of the causal effect of x_i on x_j is not valid.

Finally, we report the time elapsed in minutes in each of our synthetic benchmarks.

B. Experimental Details

We replicate a common synthetic dataset in the causal discovery literature, generating ER-4 DAGs according to the procedure of [7], with functional relationships assigned in one of two ways. The first way is using additive GPs:

$$f_j(\mathbf{x}) = \sum_{i \in \mathrm{Pa}(x_j)} g_{ij}(x_i),$$

where each g_{ij} is sampled from an RBF GP with lengthscale 1. The other is with a random MLP network with hidden size 100 and a sigmoid activation. Crucially, all weights are samples from $\mathcal{U}\left((-2.0, -0.5) \cup (0.5, 2.0)\right)$.

As noted in Section IV-C, the hyperparameters of each method may generally be tuned with cross-validation. Nonetheless, to facilitate a fair comparison between methods, we use a set of "defaults" as in [7], [8], [18]. For DAGMA and NOTEARS+, we use the hyperparameters noted in Appendix C.2.2 of [8]. However, since the sparsity penalty of DAGMA-DCE is fundamentally different than that of other methods, it is difficult to find a "default" setting for a fair comparison. We chose $\lambda_1 = 3.5 \times 10^{-2}$, which provided a sparsity penalty of similar magnitude in the MLP experiments. This is imperfect, but allows for the initial score functions of DAGMA and DAGMA-DCE to be similar without unfairly optimizing λ_1. Similarly, we used 0.25 as thresholds for comparison, which provides a similar true positive rate in the MLP experiments. Additionally, due to the added cost of DAGMA-DCE, we set the maximum number of iterations to 10% of the corresponding values in DAGMA.

All experiments were run on a NVIDIA Titan RTX GPU. The authors' implementations of NOTEARS+[3] and DAGMA[4] were used, with slight modifications to allow GPU acceleration. All methods (including DAGMA-DCE) used the same MLP architecture used in the experiments of [7], [8].

C. Results & Discussion

Our results show that DAGMA-DCE achieves the state-of-the-art on common synthetic datasets: in additive GPs (Fig. 2a), the median SID is universally lower than DAGMA or NOTEARS+. Similarly, the F_1 score is higher for all but $D = 50$ nodes, and the SHD is significantly lower for $d \leq 30$

---

[3]https://github.com/xunzheng/notears
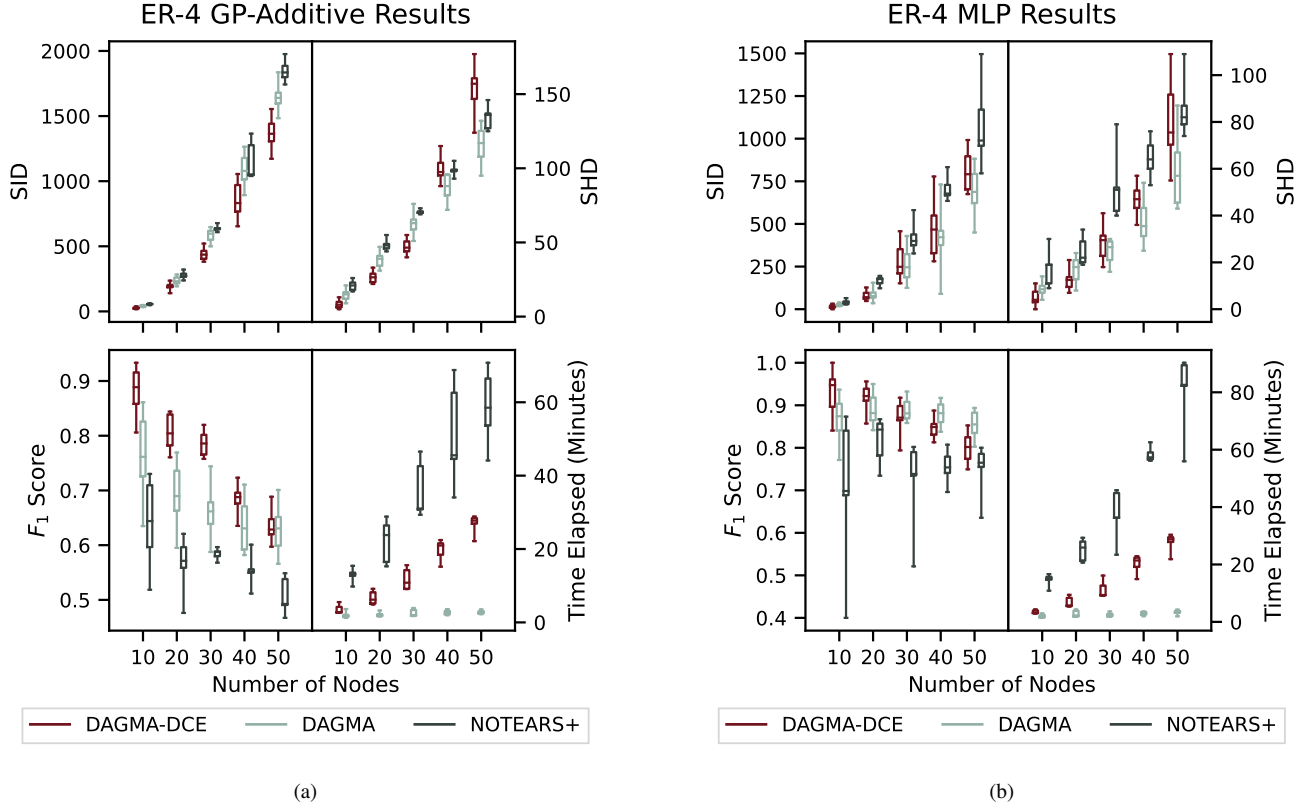[4]https://github.com/kevinsbello/dagma

(a)

(b)

FIGURE 2: Resulting SID (top left), SHD (top right), F_1 Score (bottom left), and time elapsed (bottom right) for random data generated from (a) the ER-4 GP-additive model and (b) the ER-4 MLP model, as detailed in Section B. Boxes show the median and quartiles across $T = 10$ trials for DAGMA and DAGMA-DCE, and $T = 5$ trials for NOTEARS, with whiskers showing the minimum and maximum values.

nodes. While DAGMA-DCE is indeed more computationally expensive than DAGMA, it is still substantially faster than NOTEARS across all graph sizes.

In the MLP experiments (Fig. 2b), our results are somewhat less impressive, but still competitive. This is despite what we argue is a generating process that gives a distinct advantage to DAGMA and NOTEARS+: by recalling that these methods perform thresholding on the norm of layer weights, sampling weights from $\mathcal{U}\left((-2.0, -0.5) \cup (0.5, 2.0)\right)$ builds-in identifiability in a perhaps unrealistic manner.

Finally, without asserting what the "correct" magnitude of weighted adjacency matrices is, it is interesting to note that DAGMA and DAGMA-DCE return adjacency matrices in which the relative magnitudes differ considerably. We measure this with Kendall's τ-b and the Spearman correlation coefficient, which measure rank correlation. These both take on values in $[-1, 1]$, where a value of ± 1 indicates complete agreement or disagreement, and a value of 0 indicates that orderings are unrelated. The resulting values for our synthetic experiments can be found in Table 1.

TABLE 1: Rank correlations between adjacency matrix entries of DAGMA and DAGMA-DCE across all trials for synthetic datasets, given as the mean plus-minus one standard deviation.

Dataset	Kendall's τ-b	Spearman's ρ
Additive GPs	0.40 ± 0.09	0.53 ± 0.11
MLP	0.55 ± 0.06	0.74 ± 0.07

VII. CONCLUSION

We reformulated the objectives of NOTEARS+ and DAGMA in order to extract a principled, interpretable weighted adjacency matrix. With this comes principled, interpretable notions of an ℓ_1 sparsity penalty, and thresholding the resulting graph.

A drawback of this method is its computational complexity, requiring a backward pass over the Jacobian; however, empirical results indicate it is still much faster than some other competing methods that use an augmented Lagrangian scheme for optimization. Furthermore, we were able to achieve state-of-the-art performance on synthetic datasets with small-to-medium-sized graphs. Interesting directions for future work include applications to real datasets and a formal workflow for determining sparsity and thresholding hyperparameters.

<Society logo(s) and publication title will appear here.>

REFERENCES

[1] Judea Pearl, *Causality*, Cambridge University Press, 2009.

[2] J. Peters, D. Janzing, and B. Scholkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, Adaptive Computation and Machine Learning Series. MIT Press, 2017.

[3] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman, *Causation, Prediction, and Search*, MIT Press, 2000.

[4] Jiji Zhang and Peter Spirtes, "Detection of unfaithfulness and robust causal inference," *Minds and Machines*, vol. 18, pp. 239–271, 2008.

[5] David Maxwell Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research*, vol. 3, pp. 507–554, 2002.

[6] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing, "Learning sparse nonparametric DAGs," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3414–3425.

[8] Kevin Bello, Bryon Aragam, and Pradeep Ravikumar, "DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8226–8239, 2022.

[9] Kurt Butler, Guanchao Feng, and Petar M Djurić, "A differential measure of the strength of causation," *IEEE Signal Processing Letters*, vol. 29, pp. 2208–2212, 2022.

[10] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan, "A linear non-Gaussian acyclic model for causal discovery.," *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[11] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf, "Causal discovery with continuous additive noise models," *Journal of Machine Learning Research*, vol. 15, pp. 2009–2053, 2014.

[12] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden, "D'ya like DAGS? A survey on structure learning and causal discovery," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–36, 2022.

[13] OEIS Foundation Inc., "The On-Line Encyclopedia of Integer Sequences," 2023, Published electronically at http://oeis.org.

[14] Dennis Wei, Tian Gao, and Yue Yu, "DAGs with No Fears: a closer look at continuous optimization for learning Bayesian networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3895–3906, 2020.

[15] Joachim Bona-Pellissier, François Bachoc, and François Malgouyres, "Parameter identifiability of a deep feedforward ReLU neural network," *arXiv preprint arXiv:2112.12982*, 2021.

[16] Alois Kufner and Bohumír Opic, "How to define reasonably weighted Sobolev spaces," *Commentationes Mathematicae Universitatis Carolinae*, vol. 025, no. 3, pp. 537–554, 1984.

[17] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning series. MIT Press, 2005.

[18] Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart, "NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 1942–1964.

[19] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[20] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf, "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[21] Elliott Sober, "Apportioning causal responsibility," *The Journal of Philosophy*, vol. 85, no. 6, pp. 303–318, 1988.

[22] Nattaporn Plub-in and Jitkomut Songsiri, "Estimation of Granger causality of state-space models using a clustering with Gaussian mixture model," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3853–3858.

[23] Khashaye Bayati, Karthikeyan Umapathy, and Soosan Beheshti, "Granger causality pattern learning equipped with noise invalidation soft thresholding," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 2028–2032.

[24] Jonas Peters and Peter Bühlmann, "Structural intervention distance for evaluating causal graphs," *Neural Computation*, vol. 27, no. 3, pp. 771–799, 2015.