Sample Complexity Bounds for Score-Matching: Causal Discovery and Generative Modeling

Zhenyu Zhu[†], **Francesco Locatello**[‡*], **Volkan Cevher**[†*]
† École Polytechnique Fédérale de Lausanne ‡ Institute of Science and Technology Austria
†{zhenyu.zhu, volkan.cevher}@epfl.ch ‡francesco.locatello@ista.ac.at

Abstract

This paper provides statistical sample complexity bounds for score-matching and its applications in causal discovery. We demonstrate that accurate estimation of the score function is achievable by training a standard deep ReLU neural network using stochastic gradient descent. We establish bounds on the error rate of recovering causal relationships using the score-matching-based causal discovery method of Rolland et al. [2022], assuming a sufficiently good estimation of the score function. Finally, we analyze the upper bound of score-matching estimation within the score-based generative modeling, which has been applied for causal discovery but is also of independent interest within the domain of generative models.

1 Introduction

Score matching Hyvärinen [2005], an alternative to the maximum likelihood principle for unnormalized probability density models with intractable partition functions, has recently emerged as a new state-of-the-art approach that leverages machine learning for scalable and accurate causal discovery from observational data Rolland et al. [2022]. However, the theoretical analysis and guarantees in the finite sample regime are underexplored for causal discovery even beyond score-matching approaches.

Contributions: In this work, we give the first sample complexity error bounds for score-matching using deep ReLU neural networks. With this, we obtain the first upper bound on the error rate of the method proposed by Rolland et al. [2022] to learn the topological ordering of a causal model from observational data. Thanks to the wide applicability of score-matching in machine learning, we also discuss applications to the setting of score-based generative modeling. Our main contributions are:

1. We provide the analysis of sample complexity bound for the problem of score function estimation in causal discovery for non-linear additive Gaussian noise models which has a convergence rate of log n/n with respect to the number of data. Importantly, our results require only mild additional assumptions, namely that the non-linear relationships among the causal variables are bounded and that the score function is Lipschitz. To the best of our knowledge, this is the first work to provide sampling complexity bounds for this problem.
2. We provide the first analysis of the state-of-the-art topological ordering-based causal discovery method SCORE [Rolland et al., 2022] and provide a correctness guarantee for the obtained topological order. Our results demonstrate that the algorithm's error rate converges linearly with respect to the number of training data. Additionally, we establish a connection between the algorithm's error rate and the average second derivative (curvature) of the non-linear relationships among the causal variables, discussing the impact of the causal model's inherent characteristics on the algorithm's error rate in identification.

[*]Share the senior authorship

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3. We present sample complexity bounds for the score function estimation problem in the standard score-based generative modeling method, ScoreSDE [Song et al., 2021]. In contrast to previous results [Chen et al., 2023a], our bounds do not rely on the assumption of low-dimensional input data, and we extend the applicability of the model from a specific encoder-decoder network architecture to a general deep ReLU neural network.

High-level motivation and background: Causal discovery and causal inference refer to the process of inferring causation from data and reasoning about the effect of interventions. They are highly relevant in fields such as economics [Varian, 2016], biology [Sachs et al., 2005], and healthcare [Sanchez et al., 2022]. In particular, some causal discovery methods aim to recover the causal structure of a problem solely based on observational data.

The causal structure is typically represented as a directed acyclic graph (DAG), where each node is associated with a random variable, and each edge represents a causal mechanism between two variables. Learning such a model from data is known to be NP-hard [Chickering, 1996]. Traditional approaches involve testing for conditional independence between variables or optimizing goodness-of-fit measures to search the space of possible DAGs. However, these greedy combinatorial optimization methods are computationally expensive and difficult to extend to high-dimensional settings.

An alternative approach is to reframe the combinatorial search problem as a topological ordering task [Teyssier and Koller, 2012, Solus et al., 2021, Wang et al., 2021, Rolland et al., 2022, Montagna et al., 2023b,a, Sanchez et al., 2023], where nodes are ordered from leaf to root. This can significantly speed up the search process in the DAG space. Once a topological ordering is found, a feature selection algorithm can be used to prune potential causal relations between variables, resulting in a DAG.

Recently, Rolland et al. [2022] proposed the SCORE algorithm, which utilizes the Jacobian of the score function to perform topological ordering. By identifying which elements of the Jacobian matrix of the score function remain constant across all data points, leaf nodes can be iteratively identified and removed. This approach provides a systematic way to obtain the topological ordering and infer the causal relations within the entire model. This method has achieved state-of-the-art results on multiple tasks Rolland et al. [2022] and has been extended to improve scalability Montagna et al. [2023b] also using diffusion models Sanchez et al. [2023] and to non-Gaussian noise Montagna et al. [2023a]. Interestingly, these approaches separate the concerns of statistical estimation of the score function from the causal assumption used to infer the graph (e.g., non-linear mechanisms and additive Gaussian noise). This opens an opportunity to study the convergence properties of these algorithms in the finite data regime, which is generally under-explored in the causal discovery literature. In fact, if we had error bounds on the score estimate without additional complications from causal considerations, we could study their downstream effect when the score is used for causal discovery.

Unfortunately, this is far from trivial as the theoretical research on score matching lags behind its empirical success and progress would have far-reaching implications. Even beyond causal discovery, error bounds on the estimation of the score functions would be useful for score-based generative modeling (SGM) [Song and Ermon, 2019, Song et al., 2021]. These have achieved state-of-the-art performance in various tasks, including image generation [Dhariwal and Nichol, 2021] and audio synthesis [Kong et al., 2021]. There has been significant research investigating whether accurate score estimation implies that score-based generative modeling provably converges to the true data distribution in realistic settings [Chen et al., 2023b, Lee et al., 2022, 2023]. However, the error bound of score function estimation in the context of score-based generative modeling remains an unresolved issue due to the non-convex training dynamics of neural network optimization.

Notations: We use the shorthand $[n] := \{1, 2, \ldots, n\}$ for a positive integer n. We denote by $a(n) \lesssim b(n)$: there exists a positive constant c independent of n such that $a(n) \leqslant cb(n)$. The Gaussian distribution is $\mathcal{N}(\mu, \sigma^2)$ with the μ mean and the σ^2 variance. We follow the standard Bachmann–Landau notation in complexity theory e.g., $\mathcal{O}$, o, Ω, and Θ for order notation. Due to space constraints, a detailed notation is deferred to Appendix A.

2 Preliminaries

As this paper concerns topics in score matching estimation, diffusion models, neural network theory, and causal discovery, we first introduce the background and problem setting of our work.

2.1 Score matching

For a probability density function $p(\boldsymbol{x})$, we call the score function the gradient of the log density with respect to the data $\boldsymbol{x}$. To estimate the score function $\nabla \log p(\boldsymbol{x})$, we can minimize the ℓ_2 loss over the function space $\mathcal{S}$.

$$\min_{\boldsymbol{s}\in\mathcal{S}} \mathbb{E}_p[\|\boldsymbol{s}(\boldsymbol{x}) - \nabla \log p(\boldsymbol{x})\|^2], \quad \hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}\in\mathcal{S}} \mathbb{E}_p[\|\boldsymbol{s}(\boldsymbol{x}) - \nabla \log p(\boldsymbol{x})\|^2].$$

The corresponding objective function to be minimized is the expected squared error between the true score function and the neural network:

$$J_{\text{ESM}}(\boldsymbol{s}, p(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x})}\left[\frac{1}{2}\left\|\boldsymbol{s}(\boldsymbol{x}) - \frac{\partial \log p(\boldsymbol{x})}{\partial \boldsymbol{x}}\right\|^2\right], \tag{1}$$

We refer to this formulation as explicit score matching (ESM).

Denoising score matching (DSM) is proposed by Vincent [2011] to convert the inference of the score function in ESM into the inference of the random noise and avoid the computing of the second derivative. For the sampled data $\boldsymbol{x}$, $\hat{\boldsymbol{x}}$ is obtained by adding unit Gaussian noise to $\boldsymbol{x}$. i.e. $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. We can derive the conditional probability distribution and its score function:

$$p(\hat{\boldsymbol{x}}|\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp(-\frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2}{2\sigma^2}), \qquad \frac{\partial \log p(\hat{\boldsymbol{x}}|\boldsymbol{x})}{\partial \hat{\boldsymbol{x}}} = \frac{\boldsymbol{x} - \hat{\boldsymbol{x}}}{\sigma^2}.$$

Then the DSM is defined by:

$$J_{\text{DSM}}(\boldsymbol{s}, p(\boldsymbol{x}, \hat{\boldsymbol{x}})) = \mathbb{E}_{p(\boldsymbol{x},\hat{\boldsymbol{x}})}\left[\frac{1}{2}\left\|\boldsymbol{s}\big(\hat{\boldsymbol{x}} - \frac{\partial \log p(\hat{\boldsymbol{x}}|\boldsymbol{x})}{\partial \hat{\boldsymbol{x}}}\big)\right\|^2\right] = \mathbb{E}_{p(\boldsymbol{x},\hat{\boldsymbol{x}})}\left[\frac{1}{2}\left\|\boldsymbol{s}(\hat{\boldsymbol{x}}) - \frac{\boldsymbol{x} - \hat{\boldsymbol{x}}}{\sigma^2}\right\|^2\right]. \tag{2}$$

Vincent [2011] have proven that minimizing DSM is equivalent to minimizing ESM and does not depend on the particular form of $p(\hat{\boldsymbol{x}}|\boldsymbol{x})$ or $p(\boldsymbol{x})$.

2.2 Neural network and function space

In this work, we consider a standard depth-L width-m fully connected ReLU neural network. Formally, we define a DNN with the output $\boldsymbol{s}_l(\boldsymbol{x})$ in each layer

$$\boldsymbol{s}_l(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} & l = 0, \\ \phi(\langle \boldsymbol{W}_l, \boldsymbol{s}_{l-1}(\boldsymbol{x})\rangle) & 1 \le l \le L-1, \\ \langle \boldsymbol{W}_L, \boldsymbol{s}_{L-1}(\boldsymbol{x})\rangle & l = L, \end{cases} \tag{3}$$

where the input is $\boldsymbol{x} \in \mathbb{R}^d$, the output is $\boldsymbol{s}_L(\boldsymbol{x}) \in \mathbb{R}^d$, the weights of the neural networks are $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$, $\boldsymbol{W}_l \in \mathbb{R}^{m\times m}$, $l = 2, \ldots, L-1$ and $\boldsymbol{W}_L \in \mathbb{R}^{d\times m}$. The neural network parameters formulate the tuple of weight matrices $\boldsymbol{W} := \{\boldsymbol{W}_i\}_{i=1}^L \in \{\mathbb{R}^{m\times d} \times (\mathbb{R}^{m\times m})^{L-2} \times \mathbb{R}^{d\times m}\}$. The $\mathcal{S}$ denotes the function space of Eq. (3).

The $\phi(x) = \max(0, x)$ is the ReLU activation function. According to the property $\phi(x) = x\phi'(x)$ of ReLU, we have $\boldsymbol{s}_l = \boldsymbol{D}_l \boldsymbol{W}_l \boldsymbol{s}_{l-1}$, where $\boldsymbol{D}_l$ is a diagonal matrix defined as below.

Definition 1 (Diagonal sign matrix)**.** *For $l \in [L-1]$ and $k \in [m]$, the diagonal sign matrix $\boldsymbol{D}_l$ is defined as: $(\boldsymbol{D}_l)_{k,k} = 1\{(\boldsymbol{W}_l \boldsymbol{s}_{l-1})_k \ge 0\}$.*

Initialization: We make the standard random Gaussian initialization $[\boldsymbol{W}_l]_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$ for $l \in [L-1]$ and $[\boldsymbol{W}_L]_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$.

2.3 Causal discovery

In this paper, we follow the setting in Rolland et al. [2022] and consider the following causal model, a random variable $\boldsymbol{x} \in \mathbb{R}^d$ is generated by:

$$x^{(i)} = f_i(\text{PA}_i(\boldsymbol{x})) + \epsilon_i, \quad i \in [d], \tag{4}$$

where f_i is a non-linear function, $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\mathrm{PA}_i(\boldsymbol{x})$ represent the set of parents of $x^{(i)}$ in $\boldsymbol{x}$. Then we can write the probability distribution function of $\boldsymbol{x}$ as:

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} p(x^{(i)} | \mathrm{PA}_i(\boldsymbol{x})) \,. \tag{5}$$

For such non-linear additive Gaussian noise models Eq. (4), Rolland et al. [2022] provides Algorithm 1 to learn the topological order by score matching as follows:

Algorithm 1 SCORE matching causal order search (Adapted from Algorithm 1 in Rolland et al. [2022])

Input: training data $\{(\boldsymbol{x}_{(i)})_{i=1}^N\}$.
Initialize: $\pi = []$, nodes $= \{1, \ldots, d\}$
for $k = 1, \ldots, d$ **do**
  Estimate the score function $s_{\text{nodes}} = \nabla \log p_{\text{nodes}}$ by deep ReLU network with SGD.
  Estimate $V_j = \hat{\mathrm{Var}}_{\boldsymbol{x}_{\text{nodes}}} \left[\frac{\partial \boldsymbol{s}_j(\boldsymbol{x})}{\partial \boldsymbol{x}^{(j)}} \right]$.
  $l \leftarrow \text{nodes}[\arg\min_j V_j]$
  $\pi \leftarrow [l, \pi]$
  nodes $\leftarrow$ nodes $- \{l\}$
  Remove l-th element of $\boldsymbol{x}$
end for
Get the final DAG by pruning the full DAG associated with the topological order π.

2.4 Score-based generative modeling (SGM)

In this section, we give a brief overview of SGM following Song et al. [2021], Chen et al. [2023b].

2.4.1 Score-based generative modeling with SDEs

Forward process: The success of previous score-based generative modeling methods relies on perturbing data using multiple noise scales, and the proposal of the diffusion model is to expand upon this concept by incorporating an infinite number of noise scales. This will result in the evolution of perturbed data distributions as the noise intensity increases, which will be modeled through a stochastic differential equation (SDE).

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g_t \mathrm{d}\boldsymbol{w}, \quad \boldsymbol{x}_0 \sim p_0 \,. \tag{6}$$

The expression describes $\boldsymbol{x}_t$, where the standard Wiener process (also known as Brownian motion) is denoted as $\boldsymbol{w}$, the drift coefficient of $\boldsymbol{x}_t$ is represented by a vector-valued function called $\boldsymbol{f}$, and the diffusion coefficient of $\boldsymbol{x}_t$ is denoted as g_t, a scalar function. In this context, we will refer to the probability density of $\boldsymbol{x}_t$ as p_t, and the transition kernel from $\boldsymbol{x}_s$ to $\boldsymbol{x}_t$ as $p_{st}(\boldsymbol{x}_t|\boldsymbol{x}_s)$, where $0 \leq s < t \leq T$. The Ornstein–Uhlenbeck (OU) process is a Gaussian process that is both time-homogeneous and a Markov process. It is distinct in that its stationary distribution is equivalent to the standard Gaussian distribution γ^d on $\mathbb{R}^d$.

Reverse process: We can obtain samples of $\boldsymbol{x}_0 \sim p_0^{\text{SDE}}$ by reversing the process starting from samples of $\boldsymbol{x}_T \sim p_T^{\text{SDE}}$. An important finding is that the reversal of a diffusion process is a diffusion process as well. It operates in reverse time and is described by the reverse-time SDE:

$$\mathrm{d}\boldsymbol{x}_t = \left(\boldsymbol{f}(\boldsymbol{x}_t, t) - g_t^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t) \right) \mathrm{d}t + g_t \mathrm{d}\overline{\boldsymbol{w}} \,. \tag{7}$$

When time is reversed from T to 0, $\overline{\boldsymbol{w}}$ is a standard Wiener process with an infinitesimal negative timestep of $\mathrm{d}t$. The reverse diffusion process can be derived from Eq. (7) once the score of each marginal distribution, $\nabla \log p_t(\boldsymbol{x}_t)$, is known for all t. By simulating the reverse diffusion process, we can obtain samples from p_0^{SDE}.

Some special settings: In order to simplify the writing of symbols and proofs, in this work we choose that $\boldsymbol{f}(\boldsymbol{x}_t, t) = -\frac{1}{2}\boldsymbol{x}_t$ and $g(t) = 1$ which has been widely employed in prior research [Chen et al., 2023a,b, De Bortoli et al., 2021] for theoretical analysis in Ornstein–Uhlenbeck process in score-based generative modeling.

2.4.2 Score matching in diffusion model

We aim to minimize the equivalent objective for score matching:

$$\min_{\boldsymbol{s}\in\mathcal{S}} \int_0^T w(t)\mathbb{E}_{\boldsymbol{x}_0\sim p_0}\left[\mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\,\|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0) - \boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2\,\right]\right]\mathrm{d}t\,.$$

The transition kernel has an analytical form $\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0) = -\frac{\boldsymbol{x}_t-\alpha(t)\boldsymbol{x}_0}{h(t)}$, where $\alpha(t) = e^{-\frac{t}{2}}$ and $h(t) = 1-\alpha(t)^2 = 1-e^{-t}$.

The empirical score matching loss is:

$$\min_{\boldsymbol{s}\in\mathcal{S}} \hat{\mathcal{L}}(\boldsymbol{s}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{x}_{(i)};\boldsymbol{s})\,, \tag{8}$$

where the loss function $\ell(\boldsymbol{x}_{(i)};\boldsymbol{s})$ is defined as:

$$\ell(\boldsymbol{x}_{(i)};\boldsymbol{s}) = \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x}_{(i)})}\left[\,\left\|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x}_{(i)}) - \boldsymbol{s}(\boldsymbol{x}_t,t)\right\|_2^2\,\right]\mathrm{d}t\,.$$

Here we choose $w(t) = \frac{1}{T-t_0}$, and we define the expected loss $\mathcal{L}(\cdot) = \mathbb{E}_{\boldsymbol{x}\sim p_0}[\hat{\mathcal{L}}(\cdot)]$.

3 Theoretical results for causal discovery

In this section, we state the main theoretical results of this work. We present the assumptions on non-linear additive Gaussian noise causal models in Section 3.1. Then, we present the sample complexity bound for score matching in causal discovery in Section 3.2. In Section 3.3 we provide the upper bound on the error rate for causal discovery using the Algorithm 1. The full proofs of Theorem 1 and 2 are deferred to Appendix E and F, respectively.

3.1 Assumptions

Assumption 1 (Lipschitz property of score function)**.** *The score function $\nabla\log p(\cdot)$ is 1-Lipschitz.*

Remark: The Lipschitz property of the score function is a standard assumption commonly used in the existing literature [Block et al., 2020, Lee et al., 2022, Chen et al., 2023b,a]. However, for causal discovery, this assumption limits the family of mechanisms that we can cover.

Assumption 2 (Structural assumptions of causal model)**.** *Let p be the probability density function of a random variable $\boldsymbol{x}$ defined via a non-linear additive Gaussian noise model Eq.* (4)*. Then, $\forall i\in[d]$ the non-linear function is bounded, $|f_i|\le C_i$. And $\forall i,j\in[d]$, if j is one of the parents of i, i.e. $x^{(j)}\Rightarrow x^{(i)}$, then there exist a constant C_m that satisfy:*

$$\mathbb{E}_{p(\boldsymbol{x})}\left(\frac{\partial^2 f_i(PA_i(\boldsymbol{x}))}{\partial x^{(j)2}}^2\right) \ge C_m\sigma_i^2\,.$$

Remark: This is a novel assumption that we introduce, relating the average second derivative of a mechanism (related to its curvature) to the noise variance of the child variable. This will play a crucial yet intuitive role in our error bound: identifiability is easier when there is sufficient non-linearity of a mechanism with respect to the noise of the child variable. Consider the example of a quadratic mechanism, where the second derivative is the leading constant of the polynomial. If this constant is small (e.g., close to zero), the mechanism is almost linear and we may expect that the causal model should be harder to identify. Similarly, if the child variable has a very large variance, one may expect it to be more difficult to distinguish cause from effect, as the causal effect of the parent is small compared to the noise of the child. According to Assumption 2, we can derive the identified ability margin for leaf nodes and parent nodes.

Lemma 1. *If a non-linear additive Gaussian noise model Eq.* (4) *satisfies Assumption 2. Then, $\forall i,j\in[d]$, we have:*

$$i \text{ is a leaf} \Rightarrow Var\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) = 0,\ j \text{ is not a leaf} \Rightarrow Var\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) \ge C_m.$$

This lemma intuitively relates our identifiability margin with the decision rule of SCORE Rolland et al. [2022] to identify leaves. Non-leaf nodes should have the variance of their score Jacobian sufficiently far from zero. As one may expect, we will see in Theorem 2 that the closer C_m is to zero, the more likely it is that the result of the algorithm will be incorrect given finite samples.

3.2 Error bound for score matching in causal discovery

We are now ready to state the main result of the score matching in causal discovery. We provide the sample complexity bounds of the explicit score matching Eq. (1) that using denoising score matching Eq. (2) in Algorithm 1 for non-linear additive Gaussian noise models Eq. (4).

Theorem 1. *Given a DNN defined by Eq.* (3) *trained by SGD for minimizing empirical denoising score matching objective. Suppose Assumption 1 and 2 are satisfied. For any* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$*, if* $\sigma_i \eqsim \sigma$ *and* $\frac{C_i}{\sigma_i} \eqsim 1\,,\ \forall i \in [d]$*. Then with probability at least* $1 - 2\delta - 4\exp(-\frac{d}{32}) - 2L\exp(-\Omega(m)) - \frac{1}{nd}$ *over the randomness of initialization* $\boldsymbol{W}$*, noise* $\boldsymbol{\epsilon}$ *and* ϵ_i*, it holds that:*

$$J_{ESM}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) \lesssim \frac{\sigma^2 d \log nd}{n\varepsilon^2} \log \frac{\mathcal{N}_c(\frac{1}{n}, \mathcal{S})}{\delta} + \frac{1}{n} + d\varepsilon^2\,,$$

where the $\mathcal{N}_c(\frac{1}{n}, \mathcal{S})$ *is the covering number of the function space* $\mathcal{S}$ *for deep ReLU neural network.*

Remark:

1): To the best of our knowledge, our results present the first upper bound on the explicit sampling complexity of score matching for topological ordering Algorithm 1 in non-linear additive Gaussian noise causal models. This novel contribution provides valuable insights into the efficiency and effectiveness of utilizing score matching for topological ordering in non-linear additive Gaussian noise causal models.

2): By choosing $\varepsilon^2 = \frac{1}{\sqrt{n}}$, the bound is modified to $J_{\text{ESM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) \lesssim \frac{\sigma^2 d \log nd}{\sqrt{n}} \log \frac{\mathcal{N}_c(1/n, \mathcal{S})}{\delta}$. This expression demonstrates that the ℓ_2 estimation error converges at a rate of $\frac{\log n}{\sqrt{n}}$ when the sample size n is significantly larger than the number of nodes d.

3): The bound is also related to the number of nodes d, the variance of the noise in denoising score matching σ and causal model σ_i, the covering number of the function space $\mathcal{N}_c(\frac{1}{n}, \mathcal{S})$, and the upper bound of the data C_d. If these quantities increase, it is expected that the error of explicit score matching will also increase. This is due to the increased difficulty in accurately estimating the score function.

4): Theorem 1 is rooted in the generalization by sampling complexity bound. It is independent of the specific training algorithm used. The results are broadly applicable and can be seamlessly extended to encompass larger batch GD.

Next, we will establish a connection between score matching and the precise identification of the topological ordering.

3.3 Error bound for topological order in causal discovery

Based on the previously mentioned sample complexity bound of score matching, we establish an upper bound on the error rate of the topological ordering of the causal model obtained through Algorithm 1.

Theorem 2. *Given a DNN defined by Eq.* (3) *trained by SGD with a step size* $\eta = \mathcal{O}(\frac{1}{poly(n,L)m\log^2 m})$ *for minimizing empirical score matching objective. Then under Assumption 2, for* $m \geq poly(n, L)$*, with probability at least:*

$$1 - \exp(-\Theta(d)) - (L+1)\exp(-\Theta(m)) - 2n\exp(-\frac{nC_m^2 d^2}{2^{4L+5}(\log m)^2(m^2+d^2)})\,,$$

over the randomness of initialization $\boldsymbol{W}$ *and training data that Algorithm 1 can completely recover the correct topological order of the non-linear additive Gaussian noise model.*

Remark:

1): The foundation of Theorem 2 rests upon Theorem 1, it can be seen as an embodiment of applying the upper bound of score matching for causal discovery. To the best of our knowledge, our results provide the first upper bound on the error rate of topological ordering in non-linear additive Gaussian noise causal models using Algorithm 1.

2): Considering that when $m \approx d$ and $L \approx 1$ the probability degenerates to:

$$1 - \Theta(e^{-m}) - 2n \exp\left(- \Theta\big(\frac{nC_m^2}{(\log m)^2}\big)\right).$$

The first term of the error arises due to the initialization of the neural network. As for the second term of the error, if the number of training data n satisfies $\frac{n}{\log n} \gtrsim (\log m)^2$, then it will have that $2n \exp\big(-\Theta\big(\frac{nC_m^2}{(\log m)^2}\big)\big) \lesssim 1$. This implies that the second term of the error probability exhibits linear convergence towards 0 when n is sufficiently large. Therefore, when the sample size $\frac{n}{\log n} \gtrsim (\log m)^2$, the contribution of the second term to the full error becomes negligible.

3): The theorem reveals that a smaller value of the constant C_m increases the probability of algorithm failure. This observation further confirms our previous statement that a smaller average second derivative of the nonlinear function makes it more challenging to identify the causal relationship in the model. Additionally, when the causal relationship is linear, our theorem does not provide any guarantee for the performance of Algorithm 1.

4): Consider the two variables case. If a child node is almost a deterministic function of its parents, the constant C_m can take on arbitrarily large values, according to Assumption 2. Consequently, the second term of the error probability, $2n \exp\big(- \Theta\big(\frac{nC_m^2}{(\log m)^2}\big)\big)$, tends to zero. This implies that the errors in Algorithm 1 are primarily caused by the random initialization of the neural network. The identifiability of this setting is consistent with classical results Daniušis et al. [2010], Janzing et al. [2015]. Intuitively, as long as the non-linearity is chosen independently of the noise of the parent variable[2], the application of the non-linearity will increase the distance to the reference distribution of the parent variable (in our case Gaussian). Note that for the derivative in Assumption Assumption 2 to be defined, the parent node cannot be fully deterministic.

5): Instead of focusing on the kernel regime, we directly cover the more general neural network training. The kernel approach of Rolland et al. [2022] is a special case of our analysis. The basis of Theorem 2 lies in the proof of SGD/GD convergence of the neural network, These convergence outcomes also apply to BatchGD, as demonstrated in Jentzen and Kröger [2021]. Hence, Theorem 2 can naturally be expanded to incorporate Batch GD as well.

Proof sketch: The proof of Theorem 2 can be divided into three steps. The first and most important step is to derive the upper bound of $\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}$. Here, we utilize the properties of deep ReLU neural networks to derive the distribution relationship between features of adjacent layers, then accumulate them and combine it with the properties of Gaussian initialization, yielding the upper bound for $\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}$. The second step is to use the upper bound of $\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}$ obtained in the first step combined with the concentration inequality to derive the upper bound of the error of $\mathrm{Var}\big(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\big)$. The third step is to compare the upper bound in the second step with Lemma 1 to obtain the probability of successfully selecting leaf nodes in each step. After accumulation, we can obtain the probability that Algorithm 1 can completely recover the correct topological order of the non-linear additive Gaussian noise model.

4 Theoretical results for score-based generative modeling (SGM)

In this section, we present the additional assumption required for the theoretical analysis of score matching in score-based generative modeling. Then, we provide the sample complexity bound associated with score matching in this framework. The full proof in this section is deferred to Appendix G.

Assumption 3 (Bounded data)**.** *We assume that the input data satisfy* $\|\boldsymbol{x}\|_2 \leq C_d\,, \quad \boldsymbol{x} \sim p_0$.

Remark: Bounded data is standard in deep learning theory and also commonly used in practice [Du et al., 2019b,a, Allen-Zhu et al., 2019, Oymak and Soltanolkotabi, 2020, Malach et al., 2020].

[2] Daniušis et al. [2010], Janzing et al. [2015] have formalized independence of distribution and function via an information geometric orthogonality condition that refers to a reference distribution (e.g., Gaussian)

Theorem 3. *Given a DNN defined by Eq.* (3) *trained by SGD for minimizing empirical denoising score matching loss Eq.* (8)*. Suppose Assumption 1 and 3 are satisfied. For any* $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$*. Then with probability at least* $1-2\delta-2L\exp(-\Omega(m))$ *over the randomness of initialization* $\boldsymbol{W}$ *and noise* $\boldsymbol{\epsilon}$ *in denoising score matching, it holds:*

$$\frac{1}{T-t_0}\int_{t_0}^{T}\|\nabla \log p_t(\cdot)-\hat{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2\,\mathrm{d}t \lesssim \frac{1}{n\varepsilon^2}\left(\frac{d(T-\log(t_0))}{T-t_0}+C_d^2\right)\log\frac{\mathcal{N}_c(\frac{1}{n},\mathcal{S})}{\delta}+\frac{1}{n}+d\varepsilon^2\,,$$

where the $\mathcal{N}_c(\frac{1}{n},\mathcal{S})$ *is the covering number of the function space* $\mathcal{S}$ *for deep ReLU neural network.*

Remark:

1): Theorem 3 and Theorem 1 study similar problems between causal discovery and score-based generative modeling and share similar techniques drawn from statistical learning theory and deep learning theory. These two domains are connected by a common theoretical foundation centered on the upper bound of score matching.

2): Our result extends the results for score matching in diffusion models presented in Chen et al. [2023a] which rested on the assumption of low-dimensional data structures, employing this to decompose the score function and engineer specialized network architectures for the derivation of the upper bound. Our work takes a distinct route. Our conclusions are based on the general deep ReLU neural network instead of a specific encoder-decoder network and do not rely on the assumptions of low-dimensional data used in Chen et al. [2023a]. We harness the inherent traits and conventional techniques of standard deep ReLU networks to directly deduce the upper error bound. This broader scope allows for a more comprehensive understanding of the implications and applicability of score-based generative modeling in a wider range of scenarios.

3): Similar to Theorem 1, by choosing $\varepsilon^2 = \frac{1}{\sqrt{n}}$, we can obtain the best bound $\frac{1}{T-t_0}\int_{t_0}^{T}\|\nabla \log p_t(\cdot)-\hat{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2\,\mathrm{d}t \lesssim \frac{1}{\sqrt{n}}\left(\frac{d(T-\log(t_0))}{T-t_0}+C_d^2\right)\log\frac{\mathcal{N}_c(\frac{1}{n},\mathcal{S})}{\delta}$. This expression demonstrates that the ℓ_2 estimation error converges at a rate of $\frac{1}{\sqrt{n}}$ when the sample size n is significantly larger than the dimensionality d and time steps T.

4): The bound is also related to the data dimension d, the variance of the noise in denoising score matching σ, the covering number of the function space $\mathcal{N}_c(\frac{1}{n},\mathcal{S})$, and the upper bound of the data C_d. If these quantities increase, it is expected that the error of explicit score matching will also increase. This is due to the increased difficulty in accurately estimating the score function.

5): When $t_0=0$, the theorem lacks meaning. However, when $T \gg t_0 \eqsim 1$, the bound simplifies to $\frac{d+C_d^2}{\sqrt{n}}\log\frac{\mathcal{N}_c(\frac{1}{n},\mathcal{S})}{\delta}$. This indicates that when T is sufficiently large, the loss estimated by the score function in the diffusion model becomes independent of time steps T.

6): Similar to Theorem 1, the result of Theorem 3 is also broadly applicable and can be seamlessly extended to encompass larger batch GD.

5 Numerical evidence

We conducted a series of experiments to validate the theoretical findings presented in the paper. We took inspiration from the code provided inRolland et al. [2022] and employed the structural Hamming distance (SHD) between the generated output and the actual causal graph to assess the outcomes. The ensuing experimental outcomes for SHD, vary across causal model sizes d, sample sizes n, and C_m. The experimental results are shown in Tables 1 to 3

Table 1: Fixed model size $d=100$ and the number of sampling $n=100$, SHD results of causal discovery using Algorithm 1 for different C_m values (10 runs).

C_m	1	2	4	8	16
SHD	2941.0 ± 29.5	2905.7 ± 50.8	2900.6 ± 80.8	2637.1 ± 200.4	1512.4 ± 283.6
C_m	32	64	128	256	512
SHD	413.9 ± 93.4	55.0 ± 16.0	23.9 ± 4.6	21.2 ± 5.0	13.8 ± 1.8

Table 2: Fixed model size $d = 10$ and $C_m = 1$, SHD results of causal discovery using Algorithm 1 for the different number of sampling n (10 runs).

n	5	10	20	40	80	100	160
SHD	31.7 ± 2.1	27.8 ± 4.1	23.3 ± 2.7	23.0 ± 4.0	18.4 ± 3.3	16.5 ± 3.4	13.0 ± 4.0

Table 3: Fixed the number of sampling $n = 10$ and $C_m = 1$, SHD results of causal discovery using Algorithm 1 for the different model size d (10 runs).

d	5	10	20	40	80	100
SHD	4.5 ± 2.0	29.6 ± 2.2	124.3 ± 4.6	522.8 ± 11.6	1965.4 ± 18.7	2923.7 ± 38.5

Analyzing the experimental outcomes, we find a notable pattern: higher values of C_m, augmented sample sizes n, and reduced model size d all contribute to the performance of Algorithm 1 which is consistent with the insights from Theorem 2.

6 Related Work

Score matching: Score Matching was initially introduced by Hyvärinen [2005] and extended to energy-based models by Song and Ermon [2019]. Subsequently, Vincent [2011] proposed denoising score matching, which transforms the estimation of the score function for the original distribution into an estimation for the noise distribution, effectively avoiding the need for second derivative computations. Other methods, such as sliced score matching [Song et al., 2020], denoising likelihood score matching [Chao et al., 2022], and kernel-based estimators, have also been proposed for score matching. The relationship between score matching and Fisher information [Shao et al., 2019], as well as Langevin dynamics [Hyvarinen, 2007], has been explored. On the theoretical side, Wenliang and Kanagawa [2020] introduced the concept of "blindness" in score matching, while Koehler et al. [2023] compared the efficiency of maximum likelihood and score matching, although their results primarily focus on exponential family distributions. Our paper, for the first time, analyzes the sample complexity bounds of the score function estimating in causal inference.

Causal discovery: The application of score methods for causal inference for linear additive models began with Ghoshal and Honorio [2018], which proposed a causal structure recovery method based on topological ordering from the precision matrix (equivalent to the score in that setting). Under certain noise variance assumptions, their method can reliably recover the DAG in polynomial time and sample complexity.

In recent years, there have been numerous algorithms developed for causal inference in non-linear additive models. GraNDAG [Lachapelle et al., 2021] aims to maximize the likelihood of the observed data under this model and enforces a continuous constraint to ensure the acyclicity of the causal graph Rolland et al. [2022] proposed a novel approach for causal inference which utilize score matching algorithms as a foundation for topological ordering and then employ sparse regression techniques to prune the DAG. Subsequently, Montagna et al. [2023a] extended the method to non-Gaussian noise, Sanchez et al. [2023] proposed to use diffusion models to fit the score function, and Montagna et al. [2023b] proposed a new scalable score-based preliminary neighbor search techniques.

Although advances have been achieved in leveraging machine learning for causal discovery, there is generally a lack of further research on error bounds. Other studies concentrate on broader non-parametric models but depend on various assumptions like faithfulness, restricted faithfulness, or the sparsest Markov representation [Spirtes et al., 2000, Raskutti and Uhler, 2018, Solus et al., 2021]. These approaches employ conditional independence tests and construct a graph that aligns with the identified conditional independence relations [Zhang, 2008].

Theoretical analysis of score-based generative modeling: Existing work mainly focuses on two fundamental questions: "How do diffusion models utilize the learned score functions to estimate the data distribution?" [Chen et al., 2023b, De Bortoli et al., 2021, De Bortoli, 2022, Lee et al., 2022,

2023] and "Can neural networks effectively approximate and learn score functions? What are the convergence rate and bounds on the sample complexity?" [Chen et al., 2023a].

Specifically, De Bortoli et al. [2021] and Lee et al. [2022] studied the convergence guarantees of diffusion models under the assumptions that the score estimator is accurate under the ℓ_1 and ℓ_2 norms. Concurrently Chen et al. [2023b] and Lee et al. [2023] extended previous results to distributions with bounded moments. De Bortoli [2022] studied the distribution estimation guarantees of diffusion models for low-dimensional manifold data under the assumption that the score estimator is accurate under the ℓ_1 or ℓ_2 norms.

However, these theoretical results rely on the assumption that the score function is accurately estimated, while the estimation of the score function is largely untouched due to the non-convex training dynamics. Recently, Chen et al. [2023a] provided the first sample complexity bounds for score function estimation in diffusion models. However, their result is based on the assumption that the data distribution is supported on a low-dimensional linear subspace and they use a specialized Encoder-Decoder network instead of a general deep neural network. As a result, a complete theoretical picture of score-based generative modeling is still lacking.

7 Conclusion and Limitations

In this work, we investigate the sample complexity error bounds of Score Matching using deep ReLU neural networks under two different problem settings: causal discovery and score-based generative modeling. We provide a sample complexity analysis for the estimation of the score function in the context of causal discovery for nonlinear additive Gaussian noise models, with a convergence rate of $\frac{\log n}{n}$. Furthermore, we extend the sample complexity bounds for the estimation of the score function in the ScoreSDE method to general data and achieve a convergence rate of $\frac{1}{n}$. Additionally, we provide an upper bound on the error rate of the state-of-the-art causal discovery method SCORE [Rolland et al., 2022], showing that the error rate of this algorithm converges linearly with respect to the number of training data.

A core limitation of this work is limiting our results to the Gaussian noise assumption. In fact, non-linear mechanisms with additive non-gaussian noise are also identifiable under mild additional assumptions [Peters et al., 2014] and Montagna et al. [2023a] already extended the score-matching approach of Rolland et al. [2022] to that setting. Relaxing this assumption would also allow us to apply our bounds to interesting corner cases, such as linear non-gaussian [Ghoshal and Honorio, 2018], and non-gaussian deterministic causal relations [Daniušis et al., 2010, Janzing et al., 2015]. It may be possible for this assumption to be relaxed in future work, but we argue that the added challenge, the significant difference in algorithms, and the standalone importance of the non-linear Gaussian case justify our focus.

In addition, we make other assumptions that limit the general applicability of our bounds. In particular, the assumption of the Lipschitz property for the score function imposes a strong constraint on the model space. Further investigating the relationship between the noise, the properties of the nonlinear functions in the causal model Eq. (4), and the resulting Lipschitz continuity of the score function would be an interesting extension of this work.

Acknowledgements

We are thankful to the reviewers for providing constructive feedback and Kun Zhang and Dominik Janzing for helpful discussion on the special case of deterministic children. This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043). This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. Francesco Locatello did not contribute to this work at Amazon. Corresponding author: Zhenyu Zhu.

References

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

A. Block, Y. Mroueh, and A. Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling, 2020.

C.-H. Chao, W.-F. Sun, B.-W. Cheng, Y.-C. Lo, C.-C. Chang, Y.-L. Liu, Y.-L. Chang, C.-P. Chen, and C.-Y. Lee. Denoising likelihood score matching for conditional score-based data generation. In *International Conference on Learning Representations (ICLR)*, 2022.

M. Chen, W. Liao, H. Zha, and T. Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks, 2020.

M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning (ICML)*, 2023a.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023b.

D. M. Chickering. Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, 1996.

P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Uncertainty in Artificial Intelligence*, 2010.

V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis, 2022.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in neural information processing systems (NeurIPS)*, 2021.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems (NeurIPS)*, 2021.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019a.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019b.

M. Ghosh. Exponential tail bounds for chisquared random variables. *Journal of Statistical Theory and Practice*, 2021.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.

A. Hyvarinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on neural networks*, 2007.

D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf. Justifying information-geometric causal inference. *Measures of Complexity: Festschrift for Alexey Chervonenkis*, 2015.

A. Jentzen and T. Kröger. Convergence rates for gradient descent in the training of overparameterized artificial neural networks with biases, 2021.

F. Koehler, A. Heckett, and A. Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *International Conference on Learning Representations (ICLR)*, 2023.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations (ICLR)*, 2021.

H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in neural information processing systems (NeurIPS)*, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, 2023.

E. Malach, G. Yehudai, S. Shalev-Schwartz, and O. Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning (ICML)*, 2020.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *CLeaR*, 2023a.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In *CLeaR*, 2023b.

Q. Nguyen, M. Mondelli, and G. F. Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning (ICML)*, 2021.

S. Oymak and M. Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014.

G. Raskutti and C. Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 2018.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning (ICML)*, 2022.

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

P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 2022.

P. Sanchez, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *International Conference on Learning Representations (ICLR)*, 2023.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

S. Shao, P. E. Jacob, J. Ding, and V. Tarokh. Bayesian model comparison with the hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, 2019.

L. Solus, Y. Wang, and C. Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 2021.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in neural information processing systems (NeurIPS)*, 2019.

Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, 2020.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks, 2012.

H. R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 2016.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Taylor & Francis, 2018.

P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 2011.

X. Wang, Y. Du, S. Zhu, L. Ke, Z. Chen, J. Hao, and J. Wang. Ordering-based causal discovery with reinforcement learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021.

L. K. Wenliang and H. Kanagawa. Blindness of score-based methods to isolated components and mixing proportions, 2020.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 2008.

Z. Zhu, F. Liu, G. Chrysos, and V. Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in neural information processing systems (NeurIPS)*, 2022.

Appendix introduction

The Appendix is organized as follows:

- In Appendix A, we provide a summary of the symbols and notations used throughout this paper.
- In Appendix B, we provide some background to some of the content covered in this paper.
- In Appendix C, we present several relevant lemmas that are essential to the proofs in this paper.
- In Appendix D, we provide the proof of Lemma 1.
- In Appendix E, we provide the proof of Theorem 1.
- In Appendix F, we provide the proof of Theorem 2.
- In Appendix G, we provide the proof of Theorem 3.
- In Appendix H, we discuss the Assumption 1, the Lipschitz property of score function.
- Finally, in Appendix I, we discuss the broader impacts of this paper.

A Symbols and Notation

In the paper, vectors are indicated with bold small letters, and matrices with bold capital letters. To facilitate the understanding of our work, we include some core symbols and notation in Table 4.

Table 4: Core symbols and notations used in this project.

Symbol	Dimension(s)	Definition
$\mathcal{S}$	-	Function space
$\mathcal{N}_c(\cdot, \mathcal{S})$	$\mathbb{R}$	Covering number of function space $\mathcal{S}$
$\mathcal{N}(\mu, \sigma^2)$	-	Gaussian distribution with mean μ and variance σ^2
p	-	Probability density function of a probability distribution
$\mathbb{E}$	-	Expected value
$[L]$	-	Shorthand of $\{1, 2, \ldots, L\}$
$\mathcal{O}$, o, Ω and Θ	-	Standard Bachmann–Landau order notation
n	$\mathbb{R}$	Number of data
d	$\mathbb{R}$	Data dimension (number of variables in the causal model)
L	$\mathbb{R}$	Depth of the neural network
m	$\mathbb{R}$	Width of the neural network
ϕ	-	The ReLU activation function
$x^{(i)}$	$\mathbb{R}$	The i-th element of the vector $\boldsymbol{x}$
$\boldsymbol{x}_{(i)}$	$\mathbb{R}^d$	The i-th data point
$\boldsymbol{x}_t$	$\mathbb{R}^d$	The data point in time t in diffusion model
$\boldsymbol{W}_1$	$\mathbb{R}^{m \times d}$	Weight matrix for the input layer
$\boldsymbol{W}_l$	$\mathbb{R}^{m \times m}$	Weight matrix for the l-th hidden layer
$\boldsymbol{W}_L$	$\mathbb{R}^{d \times m}$	Weight matrix for the output layer
ϵ	$\mathbb{R}$	The noise introduced by denoising score matching
σ	$\mathbb{R}$	The standard deviation of Gaussian noise ϵ
ϵ_i	$\mathbb{R}$	The noise of i-th variable of causal model
σ_i	$\mathbb{R}$	The standard deviation of Gaussian noise ϵ_i
f_i	-	Non-linear function of i-th variable of causal model
$\mathrm{PA}_i(\boldsymbol{x})$	-	The set of parents of $x^{(i)}$ in $\boldsymbol{x}$
$\mathrm{CH}_j(\boldsymbol{x})$	-	The set of children of $x^{(j)}$ in $\boldsymbol{x}$

B More backgrounds

B.1 Covering number

The basic idea of covering number is to approximate a function space with an infinite number of elements by a finite number of elements. It is used to describe how many elements (or subsets) in a given metric space can be "covered" with a finite number of reference elements (or reference subsets) to ensure that the entire space is covered. It is defined as follows:

Definition 2. *We assume there exists $m = m(\epsilon)$ elements $f_1, \ldots, f_m$ such that for any $f \in \mathcal{F}, \exists i \in \{1, \ldots, m\}$ such that $d(f, f_i) \leq \epsilon$. The minimal possible number $m(\epsilon)$ is the covering number of $\mathcal{F}$ at precision ϵ.*

In learning theory, covering number can be used to bound the Rademacher complexity [Shalev-Shwartz and Ben-David, 2014] then it is related to generalization.

B.2 More backgrounds about Algorithm 1

The main source of inspiration of the Rolland et al. [2022] to design Algorithm 1 is the following lemma:

Lemma 2 (Adapted from Lemma 1 in Rolland et al. [2022])**.** *Let p be the probability density function of a random variable $\boldsymbol{x}$ defined via a non-linear additive Gaussian noise model Eq.* (4)*, and let $\boldsymbol{s}(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x})$ be the associated score function. Then, $\forall j \in [d]$, we have:*

1. *j is a leaf $\Leftrightarrow \forall \boldsymbol{x}, \frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}} = c$, with $c \in \mathbb{R}$ independent of $\boldsymbol{x}$, i.e., $Var\big(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\big) = 0$.*
2. *j is a leaf, i is a parent of $j \Leftrightarrow s_j(\boldsymbol{x})$ depends on $\boldsymbol{x}^{(i)}$, i.e., $Var\big(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(i)}}\big) \neq 0$.*

Lemma 2 reveals the important properties of the nonlinear additive Gaussian noise model: for nonlinear additive Gaussian noise models, leaf nodes (and only leaf nodes) have the property that the associated diagonal element in the score's Jacobian is a constant. Therefore, by repeating this method and always removing the identified leaves, we can estimate a full topological order. This procedure is summarized in Algorithm 1.

C Relevant Lemmas

Lemma 3 (Adapted from Lemma 10 in Chen et al. [2020])**.** *For any $\varepsilon \in (0, 1)$ and any target 1-Lipschitz function $\tilde{\boldsymbol{s}}$ that defined on $[0, 1]^d$ with $\tilde{\boldsymbol{s}}(0) = 0$, the architecture yields an approximation $\boldsymbol{s} \in \mathcal{S}$ satisfying $\|\boldsymbol{s} - \tilde{\boldsymbol{s}}\|_\infty \leq \varepsilon$.*

The configuration of network architecture is:

$$
\begin{aligned}
&\|\boldsymbol{s}_l\|_\infty \leq \sqrt{d}\,, \quad l \in [L]\,,\\
&\|\boldsymbol{W}_l\|_\infty \leq \mathcal{O}(1)\,, \quad l \in [L]\,,\\
&L = \mathcal{O}(\log \frac{1}{\varepsilon} + d)\,,\\
&m = \mathcal{O}(\frac{1}{\varepsilon^d})\,,\\
&\sum_{l=1}^{L} \|\boldsymbol{W}_l\|_0 \leq \mathcal{O}(\frac{1}{\varepsilon^d}(\log \frac{1}{\varepsilon} + d))\,.
\end{aligned}
$$

Lemma 4 (Adapted from Theorem 4.4.5 in Vershynin [2018])**.** *Let $\boldsymbol{W}$ be an $N \times n$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$, one has:*

$$
s(\boldsymbol{A})_{\max} \leq \sqrt{N} + \sqrt{n} + t\,,
$$

where the $s(\boldsymbol{W})_{\max}$ represent the largest singular value of $\boldsymbol{W}$.

Lemma 5. *If a causal model Eq.* (4) *satisfies Assumption 2. Then with probability at least* $1 - \frac{1}{n^2 d}$ *we have:*

$$\|\boldsymbol{x}\|_2^2 \leq \sum_{i=1}^{d} (C_i + 2\sigma_i \sqrt{\log nd})^2.$$

Proof. Firstly, we can derive that:

$$\|\boldsymbol{x}\|_2^2 = \sum_{i=1}^{d} (x^{(i)})^2 = \sum_{i=1}^{d} (f_i + \epsilon_i)^2 \leq \sum_{i=1}^{d} (C_i + |\epsilon_i|)^2 \,.$$

Since $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, according to the tail bound of Gaussian distribution, with probability at least $1 - \exp(-\frac{t_i^2}{2\sigma_i^2})$ we have $|\epsilon_i| \leq t_i$. Thus:

$$\|\boldsymbol{x}\|_2^2 \leq \sum_{i=1}^{d} (C_i + t_i)^2 \,,$$

with probability at least $1 - \sum_{i=1}^{d} \exp(-\frac{t_i^2}{2\sigma_i^2})$.

Choose $t_i = 2\sigma_i \sqrt{\log nd}$, then we have:

$$\|\boldsymbol{x}\|_2^2 \leq \sum_{i=1}^{d} (C_i + 2\sigma_i \sqrt{\log nd})^2,$$

with probability at least $1 - \frac{1}{n^2 d}$.

□

D Proof of Lemma 1

Proof. According to Eq. (5), we can derive that:

$$\begin{aligned} \log p(\boldsymbol{x}) &= \sum_{i=1}^{d} \log p(x^{(i)} | \mathrm{PA}_i(\boldsymbol{x})) \\ &= -\frac{1}{2} \sum_{i=1}^{d} \left(\frac{x^{(i)} - f_i(\mathrm{PA}_i(\boldsymbol{x}))}{\sigma_i} \right)^2 - \frac{1}{2} \sum_{i=1}^{d} \log(2\pi\sigma_i^2) \,. \end{aligned}$$

Then:

$$s_j(\boldsymbol{x}) = \frac{f_j(\mathrm{PA}_j(\boldsymbol{x})) - x^{(j)}}{\sigma_j^2} + \sum_{i \in \mathrm{CH}_j(\boldsymbol{x})} \frac{\partial f_i(\mathrm{PA}_i(\boldsymbol{x}))}{\partial x^{(j)}} \frac{\epsilon_i}{\sigma_i^2} \,. \tag{9}$$

If j is a leaf:

$$\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}} = -\frac{1}{\sigma_j^2} \,, \quad \mathrm{Var}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}} \right) = 0 \,. \tag{10}$$

If j is not a leaf:

$$\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}} = -\frac{1}{\sigma_j^2} + \sum_{i \in \mathrm{CH}_j(\boldsymbol{x})} \frac{\partial^2 f_i(\mathrm{PA}_i(\boldsymbol{x}))}{\partial x^{(j)2}} \frac{\epsilon_i}{\sigma_i^2} \,,$$

where the $\mathrm{PA}_i(\boldsymbol{x})$ represent the set of parents of $x^{(i)}$ in $\boldsymbol{x}$. Then, according to the independence of ϵ_i:

$$\begin{aligned}
\mathrm{Var}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) &= \sum_{i\in \mathrm{CH}_j(\boldsymbol{x})} \mathrm{Var}\left(\frac{\partial^2 f_i(\mathrm{PA}_i(\boldsymbol{x}))}{\partial x^{(j)2}}\frac{\epsilon_i}{\sigma_i^2}\right) \\
&\geq \mathrm{Var}\left(\frac{\partial^2 f_i(\mathrm{PA}_i(\boldsymbol{x}))}{\partial x^{(j)2}}\frac{\epsilon_i}{\sigma_i^2}\right) \quad \forall i \in \mathrm{CH}_j(\boldsymbol{x}) \\
&= \mathbb{E}_{p(\boldsymbol{x})}\left(\frac{\partial^2 f_i(\mathrm{PA}_i(\boldsymbol{x}))}{\partial x^{(j)2}}^2\right)\mathrm{Var}\left(\frac{\epsilon_i}{\sigma_i^2}\right) \quad \forall i \in \mathrm{CH}_j(\boldsymbol{x}) \\
&\geq C_m\,.
\end{aligned} \tag{11}$$

Combine Eqs. (10) and (11), which concludes the proof. □

E Proof of the error bound of score function estimate for the causal model (Theorem 1)

Proof. Firstly, we use oracle inequality to decompose $J_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x}))$, for any $a \in (0,1)$ and a fixed function $\overline{\boldsymbol{s}}$, we have:

$$\begin{aligned}
J_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) &= J_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) + (1+a)\hat{J}_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) \\
&= J_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) + (1+a)\inf_{\boldsymbol{s}\in\mathcal{S}}\hat{J}_{\mathrm{DSM}}(\boldsymbol{s}, p(\boldsymbol{x})) \\
&\leq J_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) \\
&+ (1+a)\big(\hat{J}_{\mathrm{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\mathrm{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) + (1+a)J_{\mathrm{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\big) \\
&= \Big(J_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\mathrm{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) \\
&+ (1+a)\Big(\hat{J}_{\mathrm{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\mathrm{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) + (1+a)^2 J_{\mathrm{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\,.
\end{aligned}$$

First term Firstly, we define that:

$$j_{\mathrm{DSM}}(\boldsymbol{s}, \boldsymbol{x}, p(\boldsymbol{x})) = \mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})}\left\|\boldsymbol{s}(\hat{\boldsymbol{x}}) - \frac{\partial \log p(\hat{\boldsymbol{x}}|\boldsymbol{x})}{\partial \hat{\boldsymbol{x}}}\right\|_2^2\,.$$

For any $\boldsymbol{s} \in \mathcal{S}$, we have:

$$\begin{aligned}
j_{\mathrm{DSM}}(\boldsymbol{s}, \boldsymbol{x}, p(\boldsymbol{x})) &= \mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})}\left\|\boldsymbol{s}(\hat{\boldsymbol{x}}) - \frac{\partial \log p(\hat{\boldsymbol{x}}|\boldsymbol{x})}{\partial \hat{\boldsymbol{x}}}\right\|_2^2 \\
&\leq 2\mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})}\left(\|\boldsymbol{s}(\hat{\boldsymbol{x}})\|_2^2 + \left\|\frac{\partial \log p(\hat{\boldsymbol{x}}|\boldsymbol{x})}{\partial \hat{\boldsymbol{x}}}\right\|_2^2\right) \\
&= 2\mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})}\left(\|\boldsymbol{s}(\hat{\boldsymbol{x}})\|_2^2 + \left\|\frac{\boldsymbol{x}-\hat{\boldsymbol{x}}}{\sigma^2}\right\|_2^2\right).
\end{aligned} \tag{12}$$

For the first part, recall that:

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})\,.$$

Then we have:

$$\left\|\frac{\hat{\boldsymbol{x}}-\boldsymbol{x}}{\sigma}\right\|_2^2 \sim \chi^2(d)\,.$$

According to the Bernstein's inequality [Vershynin, 2018] and choose $t = \frac{1}{2}$, we have:

$$\mathbb{P}\bigg(\bigg| \frac{\left\| \frac{\hat{\boldsymbol{x}}-\boldsymbol{x}}{\sigma} \right\|_2^2}{d} - 1 \bigg| \geq \frac{1}{2} \bigg) \leq 2\exp(-\frac{d}{32}) \,.$$

Then we have:

$$\begin{aligned}
\mathbb{P}\bigg(\left\| \hat{\boldsymbol{x}} - \boldsymbol{x} \right\|_2 \geq \sigma\sqrt{\frac{3d}{2}} \bigg) &= \mathbb{P}\bigg(\left\| \hat{\boldsymbol{x}} - \boldsymbol{x} \right\|_2^2 \geq \frac{3\sigma^2 d}{2} \bigg) \\
&= \mathbb{P}\bigg(\frac{\left\| \frac{\hat{\boldsymbol{x}}-\boldsymbol{x}}{\sigma} \right\|_2^2}{d} \geq \frac{3}{2} \bigg) \\
&\leq \mathbb{P}\bigg(\bigg| \frac{\left\| \frac{\hat{\boldsymbol{x}}-\boldsymbol{x}}{\sigma} \right\|_2^2}{d} - 1 \bigg| \geq \frac{1}{2} \bigg) \\
&\leq 2\exp(-\frac{d}{32}) \,.
\end{aligned}$$

By Lemma 5, we have:

$$\begin{aligned}
\left\| \hat{\boldsymbol{x}} \right\|_2 &\leq \left\| \hat{\boldsymbol{x}} - \boldsymbol{x} \right\|_2 + \left\| \boldsymbol{x} \right\|_2 \\
&\leq \sigma\sqrt{\frac{3d}{2}} + \sqrt{\sum_{i=1}^{d} (C_i + 2\sigma_i\sqrt{\log nd})^2} \,,
\end{aligned}$$

with probability at least $1 - 2\exp(-\frac{d}{32}) - \frac{1}{n^2 d}$ over the randomness of noise $\boldsymbol{\epsilon}$ and ϵ_i.
Then by Lemma 5 and Nguyen et al. [2021][Lemma C.1]:

$$2\mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})} \left\| \boldsymbol{s}(\hat{\boldsymbol{x}}) \right\|_2^2 \lesssim \sigma^2 d + \sum_{i=1}^{d} (C_i + 2\sigma_i\sqrt{\log nd})^2 \,. \tag{13}$$

with probability at least $1 - 2\exp(-\frac{d}{32}) - L\exp(-\Omega(m)) - \frac{1}{n^2 d}$ over the randomness of initialization $\boldsymbol{W}$, noise $\boldsymbol{\epsilon}$ and ϵ_i.

For the second part:

$$\begin{aligned}
2\mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})} \left\| \frac{\boldsymbol{x} - \hat{\boldsymbol{x}}}{\sigma^2} \right\|_2^2 &= 2\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\sigma^2\boldsymbol{I})} \left\| \frac{\boldsymbol{\epsilon}}{\sigma^2} \right\|_2^2 \\
&= 2\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\boldsymbol{I})} \left\| \boldsymbol{\epsilon}' \right\|_2^2 \\
&= 2\mathbb{E}_{\epsilon''\sim\chi^2(d)} \epsilon'' \\
&= 2d \,.
\end{aligned} \tag{14}$$

Combine Eqs. (12) to (14), we have:

$$\begin{aligned}
j_{\text{DSM}}(\boldsymbol{s}, \boldsymbol{x}, p(\boldsymbol{x})) &\leq 2\mathbb{E}_{\hat{\boldsymbol{x}}\sim p(\hat{\boldsymbol{x}}|\boldsymbol{x})} \bigg(\left\| \boldsymbol{s}(\hat{\boldsymbol{x}}) \right\|_2^2 + \left\| \frac{\boldsymbol{x} - \hat{\boldsymbol{x}}}{\sigma^2} \right\|_2^2 \bigg) \\
&\lesssim (\sigma^2 + 2)d + \sum_{i=1}^{d} (C_i + 2\sigma_i\sqrt{\log nd})^2 \,,
\end{aligned} \tag{15}$$

with probability at least $1-2\exp(-\frac{d}{32})-L\exp(-\Omega(m))-\frac{1}{n^2d}$ over the randomness of initialization $\boldsymbol{W}$, noise $\boldsymbol{\epsilon}$ and ϵ_i.

According to the Bernstein-type concentration inequality Chen et al. [2023a][Lemma 15], for $\delta \in (0,1)$, $a \leq 1$ and $\tau > 0$, we have:

$$J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) \lesssim \frac{1+3/a}{2n}\big((\sigma^2+2)d + \sum_{i=1}^{d}(C_i + 2\sigma_i\sqrt{\log nd})^2\big)\log\frac{\mathcal{N}_c(\tau, \mathcal{S})}{\delta} + (2+a)\tau\,,$$

with probability at least $1-\delta-2\exp(-\frac{d}{32})-L\exp(-\Omega(m))-\frac{1}{nd}$ over the randomness of initialization $\boldsymbol{W}$, noise $\boldsymbol{\epsilon}$ and ϵ_i.

Second term According to the Bernstein-type concentration inequality Chen et al. [2023a][Lemma 15] and Eq. (15), for $\delta \in (0,1)$, $a \leq 1$, $\tau > 0$ and a fixed function $\overline{\boldsymbol{s}}$, , we have:

$$\hat{J}_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) \lesssim \frac{1+3/a}{2n}\big((\sigma^2+2)d + \sum_{i=1}^{d}(C_i + 2\sigma_i\sqrt{\log nd})^2\big)\log\frac{1}{\delta} + (2+a)\tau\,,$$

with probability at least $1-\delta-2\exp(-\frac{d}{32})-L\exp(-\Omega(m))-\frac{1}{nd}$ over the randomness of initialization $\boldsymbol{W}$, noise $\boldsymbol{\epsilon}$ and ϵ_i.

Third term We can derive that:

$$J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) = J_{\text{ESM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) + J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - J_{\text{ESM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\,.$$

According to Lemma 3, since the error term is invariant with respect to translations on $\nabla\log p(\cdot)$ and the homogeneity of the ReLU neural network, we can omit $\nabla\log p(\mathbf{0}) = 0$ and rescale bound for the input data, for any $\varepsilon \in (0,1)$, there exists an approximation function $\overline{\boldsymbol{s}}$ satisfying $\|\nabla\log p(\cdot) - \overline{\boldsymbol{s}}(\cdot)\|_\infty \leq \varepsilon$, then we have:

$$J_{\text{ESM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) \leq \frac{d\varepsilon^2}{2}\,,$$

with probability at least $1-\frac{1}{nd}$ over the randomness of noise ϵ_i and satisfy the configuration of network architecture in Lemma 3.

According to Vincent [2011], we have:

$$J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - J_{\text{ESM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\hat{\boldsymbol{x}}\sim\phi(\boldsymbol{x}|\boldsymbol{x})}\big[\,\|\nabla_{\hat{\boldsymbol{x}}}\log\phi(\boldsymbol{x}|\boldsymbol{x})\|_2^2\,\big] - \frac{1}{2}\|\nabla\log p(\cdot)\|_{\ell^2(p)}^2\,.$$

which is an absolute value that does not depend on $\boldsymbol{s}$. So we can define that:

$$E_1 \coloneqq \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\hat{\boldsymbol{x}}\sim\phi(\boldsymbol{x}|\boldsymbol{x})}\big[\,\|\nabla_{\hat{\boldsymbol{x}}}\log\phi(\boldsymbol{x}|\boldsymbol{x})\|_2^2\,\big] - \frac{1}{2}\|\nabla\log p(\cdot)\|_{\ell^2(p)}^2\,.$$

So if we choose $\overline{\boldsymbol{s}}$ is the approximation function that Lemma 3 provide, then we have:

$$J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) \leq \frac{d\varepsilon^2}{2} + E_1\,.$$

Putting things together Combine all three terms, we have:

$$
\begin{aligned}
J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) &\leq \Big(J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) \\
&\quad + (1+a)\Big(\hat{J}_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) + (1+a)^2 J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) \\
&\lesssim \Big(J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) \\
&\quad + (1+a)\Big(\hat{J}_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) + (1+a)^2\Big(\frac{d\varepsilon^2}{2} + E_1\Big) \\
&= \Big(J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) \\
&\quad + (1+a)\Big(\hat{J}_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) + (1+a)^2\frac{d\varepsilon^2}{2} + (2a+a^2)E_1 + E_1\,.
\end{aligned}
$$

Then:

$$
\begin{aligned}
J_{\text{ESM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) &= J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - E_1 \\
&\lesssim \Big(J_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)\hat{J}_{\text{DSM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) \\
&\quad + (1+a)\Big(\hat{J}_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x})) - (1+a)J_{\text{DSM}}(\overline{\boldsymbol{s}}, p(\boldsymbol{x}))\Big) + (1+a)^2\frac{d\varepsilon^2}{2} + (2a+a^2)E_1 \\
&\lesssim \frac{1+3/a}{2n}\big((\sigma^2+2)d + \sum_{i=1}^{d}(C_i + 2\sigma_i\sqrt{\log nd})^2\big)\log\frac{\mathcal{N}_c(\tau, \mathcal{S})}{\delta} + (2+a)\tau \\
&\quad + (1+a)\Big(\frac{1+3/a}{2n}\big((\sigma^2+2)d + \sum_{i=1}^{d}(C_i + 2\sigma_i\sqrt{\log nd})^2\big)\log\frac{1}{\delta} + (2+a)\tau\Big) \\
&\quad + (1+a)^2\frac{d\varepsilon^2}{2} + (2a+a^2)E_1\,,
\end{aligned}
$$

with probability at least $1 - 2\delta - 4\exp(-\frac{d}{32}) - 2L\exp(-\Omega(m)) - \frac{1}{nd}$ over the randomness of initialization $\boldsymbol{W}$, noise $\boldsymbol{\epsilon}$ and ϵ_i.

Let $a = \varepsilon^2$, $\tau = \frac{1}{n}$, $\sigma_i \eqsim \sigma$ and $\frac{C_i}{\sigma_i} \eqsim 1\,,\ \forall i \in [d]$. Then we have:

$$
J_{\text{ESM}}(\hat{\boldsymbol{s}}, p(\boldsymbol{x})) \lesssim \frac{\sigma^2 d\log nd}{n\varepsilon^2}\log\frac{\mathcal{N}_c(\frac{1}{n}, \mathcal{S})}{\delta} + \frac{1}{n} + d\varepsilon^2\,,
$$

with probability at least $1 - 2\delta - 4\exp(-\frac{d}{32}) - 2L\exp(-\Omega(m)) - \frac{1}{nd}$ over the randomness of initialization $\boldsymbol{W}$, noise $\boldsymbol{\epsilon}$ and ϵ_i. □

F Proof of the error bound of topological ordering using the SCORE algorithm in a causal model (Theorem 2)

Proof. We set the weights of the neural network after training are $\widehat{\boldsymbol{W}}$. i.e.

$$
\boldsymbol{s}(\boldsymbol{x}) = \widehat{\boldsymbol{W}}_L\phi(\widehat{\boldsymbol{W}}_{L-1}\cdots\phi(\widehat{\boldsymbol{W}}_1\boldsymbol{x})\cdots)\,.
$$

According to the standard chain rule and Zhu et al. [2022][Lemma 3], we have:

$$
\nabla_{\boldsymbol{x}}\boldsymbol{s}(\boldsymbol{x})^\top = \widehat{\boldsymbol{W}}_L\widehat{\boldsymbol{D}}_{L-1}\widehat{\boldsymbol{W}}_{L-1}\cdots\widehat{\boldsymbol{D}}_1\widehat{\boldsymbol{W}}_1\,.
$$

Let $\boldsymbol{v}_i$ be a one-hot vector with length d, with the i-th element is 1 and the rest of the elements are 0, then we have:

$$\begin{aligned}
\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}} &= \boldsymbol{v}_i \widehat{\boldsymbol{W}}_L \widehat{\boldsymbol{D}}_{L-1} \widehat{\boldsymbol{W}}_{L-1} \cdots \widehat{\boldsymbol{D}}_1 \widehat{\boldsymbol{W}}_1 \boldsymbol{v}_i \\
&\leq \|\boldsymbol{v}_i\|_2 \left\| \widehat{\boldsymbol{W}}_L \widehat{\boldsymbol{D}}_{L-1} \widehat{\boldsymbol{W}}_{L-1} \cdots \widehat{\boldsymbol{D}}_1 \right\|_2 \left\| \widehat{\boldsymbol{W}}_1 \right\|_2 \|\boldsymbol{v}_i\|_2 \\
&= \left\| \widehat{\boldsymbol{W}}_L \widehat{\boldsymbol{D}}_{L-1} \widehat{\boldsymbol{W}}_{L-1} \cdots \widehat{\boldsymbol{D}}_1 \right\|_2 \left\| \widehat{\boldsymbol{W}}_1 \right\|_2 \\
&= \left(\|\boldsymbol{W}_L \boldsymbol{D}_{L-1} \boldsymbol{W}_{L-1} \cdots \boldsymbol{D}_1\|_2 + \left\| \widehat{\boldsymbol{W}}_L \widehat{\boldsymbol{D}}_{L-1} \widehat{\boldsymbol{W}}_{L-1} \cdots \widehat{\boldsymbol{D}}_1 - \boldsymbol{W}_L \boldsymbol{D}_{L-1} \boldsymbol{W}_{L-1} \cdots \boldsymbol{D}_1 \right\|_2 \right) \\
&\times \left(\|\boldsymbol{W}_1\|_2 + \left\| \widehat{\boldsymbol{W}}_1 - \boldsymbol{W}_1 \right\|_2 \right) \\
&\coloneqq (T_1 + T_2) \times (T_3 + T_4) \,.
\end{aligned} \tag{16}$$

Firstly, we focus on T_1. Define $\boldsymbol{t}_l(\boldsymbol{v}) = \boldsymbol{D}_l \boldsymbol{W}_l \cdots \boldsymbol{D}_1 \boldsymbol{v}$, then for any vector $\boldsymbol{v}$ that satisfy $\|\boldsymbol{v}\|_2 = 1$:

$$\begin{aligned}
\|\boldsymbol{W}_L \boldsymbol{D}_{L-1} \boldsymbol{W}_{L-1} \cdots \boldsymbol{D}_1 \boldsymbol{v}\|_2 &= \|\boldsymbol{W}_L \boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2 \\
&= \sqrt{\|\boldsymbol{W}_L \boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2} \\
&= \sqrt{\frac{\|\boldsymbol{W}_L \boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2} \frac{\|\boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_{L-2}(\boldsymbol{v})\|_2^2} \cdots \frac{\|\boldsymbol{t}_2(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_1(\boldsymbol{v})\|_2^2} \|\boldsymbol{t}_1(\boldsymbol{v})\|_2^2} \,.
\end{aligned} \tag{17}$$

According to Zhu et al. [2022][Lemma 2], we have:

$$\frac{\|\boldsymbol{t}_l(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_{l-1}(\boldsymbol{v})\|_2^2} \sim \frac{2}{m} \chi^2(\varrho), \quad \forall l = 2, \cdots, L-1 \,,$$

where $\varrho \sim \mathrm{Ber}(m, 1/2)$.

According to Ghosh [2021], with probability at least $1 - \exp(-\Theta(m))$ over the randomness of initialization $\boldsymbol{W}_l$, we have:

$$\frac{\|\boldsymbol{t}_l(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_{l-1}(\boldsymbol{v})\|_2^2} \leq 4, \quad \forall l = 2, \cdots, L-1 \,. \tag{18}$$

By the definition of chi-square distribution, we have:

$$\frac{\|\boldsymbol{W}_L \boldsymbol{t}_{L-1}\|_2^2}{\|\boldsymbol{t}_{L-1}\|_2^2} \sim \frac{\chi^2(d)}{d} \,,$$

Similar, according to Ghosh [2021], with probability at least $1 - \exp(-\Theta(d))$ over the randomness of initialization $\boldsymbol{W}_L$, we have:

$$\frac{\|\boldsymbol{W}_L \boldsymbol{t}_{L-1}\|_2^2}{\|\boldsymbol{t}_{L-1}\|_2^2} \leq 2 \,. \tag{19}$$

And we can derive that:

$$\|\boldsymbol{t}_1(\boldsymbol{v})\|_2^2 = \|\boldsymbol{D}_1 \boldsymbol{v}\|_2^2 \leq \left(\|\boldsymbol{D}_1\|_2 \|\boldsymbol{v}\|_2 \right)^2 \leq 1 \,. \tag{20}$$

Combine Eqs. (17) to (20), we have:

$$\|\boldsymbol{W}_L\boldsymbol{D}_{L-1}\boldsymbol{W}_{L-1}\cdots\boldsymbol{D}_1\boldsymbol{v}\|_2=\sqrt{\frac{\|\boldsymbol{W}_L\boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2}\frac{\|\boldsymbol{t}_{L-1}(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_{L-2}(\boldsymbol{v})\|_2^2}\cdots\frac{\|\boldsymbol{t}_2(\boldsymbol{v})\|_2^2}{\|\boldsymbol{t}_1(\boldsymbol{v})\|_2^2}\|\boldsymbol{t}_1(\boldsymbol{v})\|_2^2}\leq 2^{\frac{2L-1}{2}}\,,$$

with probability at least $1-\exp(-\Theta(d))-(L-2)\exp(-\Theta(m))$ over the randomness of initialization $\boldsymbol{W}$.

i.e.

$$T_1=\|\boldsymbol{W}_L\boldsymbol{D}_{L-1}\boldsymbol{W}_{L-1}\cdots\boldsymbol{D}_1\|_2\leq 2^{\frac{2L-1}{2}}\,,\tag{21}$$

with probability at least $1-\exp(-\Theta(d))-(L-2)\exp(-\Theta(m))$ over the randomness of initialization $\boldsymbol{W}$.

For a perturbation matrices satisfy $T_4=\left\|\widehat{\boldsymbol{W}}_l-\boldsymbol{W}_l\right\|_2\leq\omega=\mathcal{O}(\frac{1}{L^{3/2}})\,,\ \forall l\in[L]$, by Allen-Zhu et al. [2019, Lemma 8.7], we obtain that for any integer $s=\mathcal{O}(m\omega^{2/3}L)$ and $d\leq\mathcal{O}(\frac{m}{L\log m})$, with probability at least $1-\exp\left(-\Omega(m\log m\omega^{2/3}L)\right)$ over the randomness of initialization $\boldsymbol{W}$, it holds that:

$$T_2=\left\|\widehat{\boldsymbol{W}}_L\widehat{\boldsymbol{D}}_{L-1}\widehat{\boldsymbol{W}}_{L-1}\cdots\widehat{\boldsymbol{D}}_1-\boldsymbol{W}_L\boldsymbol{D}_{L-1}\boldsymbol{W}_{L-1}\cdots\boldsymbol{D}_1\right\|_2\leq\mathcal{O}\bigg(\frac{\omega^{1/3}L^2\sqrt{m\log m}}{\sqrt{d}}\bigg)\,.\tag{22}$$

For T_3, according to Lemma 4, we have that for every $t\geq 0$, with probability at least $1-2\exp(-t^2/2)$ over the randomness of initialization $\boldsymbol{W}_1$, one has:

$$T_3=\|\boldsymbol{W}_1\|_2\leq\sqrt{\frac{2}{m}}(\sqrt{m}+\sqrt{d}+t)\,.\tag{23}$$

Combine Eqs. (16) and (21) to (23), choose $t=\sqrt{m}$ we have:

$$\begin{aligned}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}&\leq(T_1+T_2)\times(T_3+T_4)\\&\lesssim\bigg(2^{\frac{2L-1}{2}}+\frac{\omega^{1/3}L^2\sqrt{m\log m}}{\sqrt{d}}\bigg)\times\bigg(\frac{1}{L^{3/2}}+\sqrt{\frac{2}{m}}(2\sqrt{m}+\sqrt{d})\bigg)\\&\lesssim\frac{2^L\sqrt{\log m}(\sqrt{m}+\sqrt{d})}{\sqrt{d}}\,,\end{aligned}\tag{24}$$

with probability at least $1-\exp(-\Theta(d))-L\exp(-\Theta(m))-\exp\left(-\Omega(m\log m)\right)$ over the randomness of initialization $\boldsymbol{W}$.

Then, for $\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)$, we have that:

$$\bigg(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\bigg)^2\lesssim\frac{2^{2L}\log m(m+d)}{d}\,,\tag{25}$$

with probability at least $1-\exp(-\Theta(d))-L\exp(-\Theta(m))-\exp\left(-\Omega(m\log m)\right)$ over the randomness of initialization $\boldsymbol{W}$.

According to Hoeffding's inequality for bounded random variables [Vershynin, 2018][Thmorem 2.2.6], we have that:

$$\left|\frac{1}{n}\sum_{i=1}^n\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}-\mathbb{E}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right|\leq\frac{C_m}{12\mathbb{E}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}}\,,$$

with probability at least $1 - \exp(-\Theta(d)) - L\exp(-\Theta(m)) - \exp\left(-\Omega(m\log m)\right) - 2\exp(-\Omega(\frac{nC_m^2 d^2}{2^{4L+5}(\log m)^2(m^2+d^2)}))$, and

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2 - \mathbb{E}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2\right| \leq \frac{C_m}{4}\,,$$

with probability at least $1 - \exp(-\Theta(d)) - L\exp(-\Theta(m)) - \exp\left(-\Omega(m\log m)\right) - 2\exp(-\frac{nC_m^2 d^2}{2^{4L+5}(\log m)^2(m^2+d^2)})$.

Then we have:

$$\begin{aligned}
\left|\mathrm{Var}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) - \hat{\mathrm{Var}}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)\right| &= \left|\mathbb{E}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2 - \left(\mathbb{E}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2 - \sum_{i=1}^{n}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2 + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2\right| \\
&\leq \left|\mathbb{E}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2 - \sum_{i=1}^{n}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2\right| + \left|-\left(\mathbb{E}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2 + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right)^2\right| \\
&\leq \frac{C_m}{4} + \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}} - \mathbb{E}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right|\left|\mathbb{E}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right| \\
&\leq \frac{C_m}{2}\,,
\end{aligned}$$

□

with probability at least $1 - \exp(-\Theta(d)) - L\exp(-\Theta(m)) - \exp\left(-\Theta(m\log m)\right) - 2\exp(-\frac{nC_m^2 d^2}{2^{4L+5}(\log m)^2(m^2+d^2)})$.

Thus, for i is a leaf and j is not a leaf, according to Assumption 2 and Lemma 1, we have:

$$\mathrm{Var}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) - \mathrm{Var}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) \geq C_m\,.$$

Then:

$$\begin{aligned}
\hat{\mathrm{Var}}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) &= \hat{\mathrm{Var}}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) - \mathrm{Var}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) + \mathrm{Var}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) \\
&\leq \frac{C_m}{2} + \mathrm{Var}\left(\frac{\partial s_i(\boldsymbol{x})}{\partial x^{(i)}}\right) \\
&\leq \mathrm{Var}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) - \frac{C_m}{2} \\
&= \mathrm{Var}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) - \hat{\mathrm{Var}}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) + \hat{\mathrm{Var}}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) - \frac{C_m}{2} \\
&\leq \frac{C_m}{2} + \hat{\mathrm{Var}}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right) - \frac{C_m}{2} \\
&= \hat{\mathrm{Var}}\left(\frac{\partial s_j(\boldsymbol{x})}{\partial x^{(j)}}\right).
\end{aligned}$$

with probability at least $1 - \exp(-\Theta(d)) - L\exp(-\Theta(m)) - \exp\left(-\Theta(m\log m)\right) - 2\exp(-\frac{nC_m^2 d^2}{2^{4L+5}(\log m)^2(m^2+d^2)})$. Considering all variables, then with probability at least:

$$1 - \exp(-\Theta(d)) - (L+1)\exp(-\Theta(m)) - 2n\exp(-\frac{nC_m^2 d^2}{2^{4L+5}(\log m)^2(m^2+d^2)})\,,$$

that Algorithm 1 can completely recover the correct topological order of the non-linear additive Gaussian noise model.

G Proof of the error bound of score function estimate for the score-based generative modeling (Theorem 3)

Proof. Firstly, we use oracle inequality to decompose $\mathcal{L}(\hat{\boldsymbol{s}})$, for any $a \in (0,1)$ and a fixed function $\overline{\boldsymbol{s}}$, we have:

$$
\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{s}}) &= \mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}}) + (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}}) \\
&= \mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}}) + (1+a)\inf_{\boldsymbol{s}\in\mathcal{S}} \hat{\mathcal{L}}(\boldsymbol{s}) \\
&\leq \mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}}) + (1+a)\big(\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}}) + (1+a)\mathcal{L}(\overline{\boldsymbol{s}})\big) \\
&= \Big(\mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}})\Big) + (1+a)\Big(\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}})\Big) + (1+a)^2\mathcal{L}(\overline{\boldsymbol{s}})\,.
\end{aligned}
$$

First term For any $\boldsymbol{s} \in \mathcal{S}$, we have:

$$
\begin{aligned}
\ell(\boldsymbol{x};\boldsymbol{s}) &= \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})}\big[\,\|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x}) - \boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2\,\big]\mathrm{d}t \\
&= \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})}\left(\left\|\frac{\boldsymbol{x}_t-\alpha(t)\boldsymbol{x}}{h(t)} + \boldsymbol{s}(\boldsymbol{x}_t,t)\right\|_2^2\right)\mathrm{d}t \\
&\leq \frac{3}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})}\left[\left(\left\|\frac{\boldsymbol{x}_t}{h(t)}\right\|_2^2 + \left\|\frac{\alpha(t)\boldsymbol{x}}{h(t)}\right\|_2^2 + \|\boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2\right)\right]\mathrm{d}t \\
&= \frac{3}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})}\left(\left\|\frac{\boldsymbol{x}_t}{h(t)}\right\|_2^2\right)\mathrm{d}t \\
&+ \frac{3}{T-t_0}\int_{t_0}^{T}\left(\left\|\frac{\alpha(t)\boldsymbol{x}}{h(t)}\right\|_2^2\right)\mathrm{d}t \\
&+ \frac{3}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})}\left(\|\boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2\right)\mathrm{d}t\,.
\end{aligned}
\tag{26}
$$

For the first part, for forward process SDE Eq. (6) we can easily derive that $p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0) \sim \mathcal{N}\big(\alpha(t)\boldsymbol{x}_0, h(t)I_d\big)$, where $\alpha(t) = e^{-\frac{t}{2}}$ and $h(t) = 1-\alpha(t)^2$.

$$
\begin{aligned}
&\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})}\left(\left\|\frac{\boldsymbol{x}_t}{h(t)}\right\|_2^2\right)\mathrm{d}t \\
=&\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t\sim\mathcal{N}\left(\alpha(t)\boldsymbol{x}_0,h(t)I_d\right)}\left(\left\|\frac{\boldsymbol{x}_t}{h(t)}\right\|_2^2\right)\mathrm{d}t \\
=&\int_{t_0}^{T}\left(\sum_{i=1}^{d}\big[\mathbb{E}_{x_t^{(i)}\sim\mathcal{N}\left(\alpha(t)x_0^{(i)},h(t)\right)}\left(\frac{x_t^{(i)}}{h(t)}\right)^2\big]\right)\mathrm{d}t \\
=&\int_{t_0}^{T}\left(\sum_{i=1}^{d}\big[\frac{\alpha(t)^2}{h(t)^2}(x_0^{(i)})^2 + \frac{1}{h(t)}\big]\right)\mathrm{d}t \\
=&\sum_{i=1}^{d}(x_0^{(i)})^2\int_{t_0}^{T}\frac{\alpha(t)^2}{h(t)^2}\mathrm{d}t + \int_{t_0}^{T}\frac{d}{h(t)}\mathrm{d}t \\
\leq&\frac{T-t_0}{Tt_0}C_d^2 + d(T-\log(t_0))\,.
\end{aligned}
\tag{27}
$$

For the second part:

$$\int_{t_0}^{T} \left\| \frac{\alpha(t)\boldsymbol{x}}{h(t)} \right\|_2^2 \mathrm{d}t \leq C_d^2 \int_{t_0}^{T} \frac{\alpha(t)^2}{h(t)^2} \mathrm{d}t = C_d^2 \frac{T-t_0}{Tt_0} \,. \tag{28}$$

For the third part, by Nguyen et al. [2021][Lemma C.1]:

$$\int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t \sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})} \left(\|\boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2 \right) \mathrm{d}t \eqsim C_d^2 (T-t_0) \,, \tag{29}$$

with probability at least $1 - L\exp(-\Omega(m))$ over the randomness of initialization $\boldsymbol{W}$.

Combine Eqs. (26) to (29), we have:

$$\begin{aligned}
\ell(\boldsymbol{x};\boldsymbol{s}) &= \frac{3}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t \sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})} \left(\left\| \frac{\boldsymbol{x}_t}{h(t)} \right\|_2^2 \right) \mathrm{d}t \\
&+ \frac{3}{T-t_0} \int_{t_0}^{T} \left(\left\| \frac{\alpha(t)\boldsymbol{x}}{h(t)} \right\|_2^2 \right) \mathrm{d}t \\
&+ \frac{3}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\boldsymbol{x}_t \sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x})} \left(\|\boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2 \right) \mathrm{d}t \\
&\lesssim \frac{3}{T-t_0} \left(\frac{T-t_0}{Tt_0} C_d^2 + d(T - \log(t_0)) + C_d^2 \frac{T-t_0}{Tt_0} + C_d^2 (T-t_0) \right) \\
&= \frac{3d(T-\log(t_0))}{T-t_0} + 3C_d^2 + \frac{6C_d^2}{Tt_0} \,,
\end{aligned} \tag{30}$$

with probability at least $1 - L\exp(-\Omega(m))$ over the randomness of initialization $\boldsymbol{W}$.

According to the Bernstein-type concentration inequality Chen et al. [2023a][Lemma 15], for $\delta \in (0,1)$, $a \leq 1$ and $\tau > 0$, we have:

$$\mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}}) \lesssim \frac{1+3/a}{n} \left(\frac{d(T-\log(t_0))}{T-t_0} + C_d^2 \right) \log \frac{\mathcal{N}_c(\tau, \mathcal{S})}{\delta} + (2+a)\tau \,,$$

with probability at least $1 - \delta - L\exp(-\Omega(m))$ over the randomness of initialization $\boldsymbol{W}$.

Second term According to the Bernstein-type concentration inequality Chen et al. [2023a][Lemma 15] and Eq. (30), for $\delta \in (0,1)$, $\tau > 0$ and a fixed function $\overline{\boldsymbol{s}}$, , we have:

$$\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}}) \lesssim \frac{1+3/a}{n} \left(\frac{d(T-\log(t_0))}{T-t_0} + C_d^2 \right) \log \frac{1}{\delta} + (2+a)\tau \,,$$

with probability at least $1 - \delta - L\exp(-\Omega(m))$ over the randomness of initialization $\boldsymbol{W}$.

Third term We can derive that:

$$\begin{aligned}
\mathcal{L}(\overline{\boldsymbol{s}}) &= \frac{1}{T-t_0} \int_{t_0}^{T} \|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \, \mathrm{d}t \\
&+ \mathcal{L}(\overline{\boldsymbol{s}}) - \frac{1}{T-t_0} \int_{t_0}^{T} \|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \, \mathrm{d}t
\end{aligned}$$

For the first part, according to Lemma 3, since the error term is invariant with respect to translations on $\nabla \log p_t(\cdot)$ and the homogeneity of the ReLU neural network, we can omit $\nabla \log p_t(\mathbf{0}) = 0$ and rescale bound for the input data, for any $\varepsilon \in (0,1)$, there exist an approximation function $\overline{\boldsymbol{s}}$ satisfying $\|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_\infty \leq \varepsilon$, then we have:

$$\frac{1}{T-t_0}\int_{t_0}^{T} \|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \,\mathrm{d}t \leq d\varepsilon^2\,,$$

that satisfy the configuration of network architecture in Lemma 3.

For the second part:

$$\begin{aligned}
&\mathcal{L}(\overline{\boldsymbol{s}}) - \frac{1}{T-t_0}\int_{t_0}^{T} \|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \,\mathrm{d}t \\
&= \frac{1}{T-t_0}\int_{t_0}^{T} \left(\mathbb{E}_{\boldsymbol{x}_0\sim p_0}\mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\, \|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0) - \boldsymbol{s}(\boldsymbol{x}_t,t)\|_2^2 \,\right] - \|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \right)\mathrm{d}t\,.
\end{aligned}$$

According to Vincent [2011], we have:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_0\sim p_0}\mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x}_{(i)})}\left[\, \left\|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0=\boldsymbol{x}_{(i)}) - \boldsymbol{s}(\boldsymbol{x}_t,t)\right\|_2^2 \,\right] - \|\nabla \log p_t(\cdot) - \overline{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \\
=&\mathbb{E}_{\boldsymbol{x}_0\sim p_0}\mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\, \|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2 \,\right] - \|\nabla \log p_t(\cdot)\|_{\ell^2(p_t)}^2\,,
\end{aligned}$$

which is an absolute value that does not depend on $\boldsymbol{s}$. So we can define that:

$$E_2 \coloneqq \mathbb{E}_{\boldsymbol{x}_0\sim p_0}\mathbb{E}_{\boldsymbol{x}_t\sim p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\, \|\nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2 \,\right] - \|\nabla \log p_t(\cdot)\|_{\ell^2(p_t)}^2\,.$$

So if we choose $\overline{\boldsymbol{s}}$ is the approximation function that Lemma 3 provide, then we have:

$$\mathcal{L}(\overline{\boldsymbol{s}}) \leq d\varepsilon^2 + E_2\,.$$

Putting things together Combine all three terms, we have:

$$\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{s}}) &\leq \left(\mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}})\right) + (1+a)\left(\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}})\right) + (1+a)^2\mathcal{L}(\overline{\boldsymbol{s}}) \\
&\leq \left(\mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}})\right) + (1+a)\left(\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}})\right) + (1+a)^2(d\varepsilon^2 + E_2) \\
&= \left(\mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}})\right) + (1+a)\left(\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}})\right) + (1+a)^2 d\varepsilon^2 + (2a+a^2)E_2 + E_2
\end{aligned}$$

Then:

$$\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T} \|\nabla \log p_t(\cdot) - \hat{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2 \,\mathrm{d}t \\
=&\mathcal{L}(\hat{\boldsymbol{s}}) - E_2 \\
=&\left(\mathcal{L}(\hat{\boldsymbol{s}}) - (1+a)\hat{\mathcal{L}}(\hat{\boldsymbol{s}})\right) + (1+a)\left(\hat{\mathcal{L}}(\overline{\boldsymbol{s}}) - (1+a)\mathcal{L}(\overline{\boldsymbol{s}})\right) + (1+a)^2 d\varepsilon^2 + (2a+a^2)E_2 \\
\lesssim&\left(\frac{1+3/a}{n}\left(\frac{d(T-\log(t_0))}{T-t_0} + C_d^2\right)\log\frac{\mathcal{N}_c(\tau,\mathcal{S})}{\delta} + (2+a)\tau\right) \\
&+(1+a)\left(\frac{1+3/a}{n}\left(\frac{d(T-\log(t_0))}{T-t_0} + C_d^2\right)\log\frac{1}{\delta} + (2+a)\tau\right) \\
&+(1+a)^2 d\varepsilon^2 + (2a+a^2)E_2\,,
\end{aligned}$$

with probability at least $1-2\delta-2L\exp(-\Omega(m))$ over the randomness of initialization $\boldsymbol{W}$.

Let $a=\varepsilon^2$ and $\tau=\frac{1}{n}$, then we have:

$$\frac{1}{T-t_0}\int_{t_0}^{T}\|\nabla\log p_t(\cdot)-\hat{\boldsymbol{s}}(\cdot,t)\|_{\ell^2(p_t)}^2\,\mathrm{d}t\lesssim\frac{1}{n\varepsilon^2}\left(\frac{d(T-\log(t_0))}{T-t_0}+C_d^2\right)\log\frac{\mathcal{N}_c(\frac{1}{n},\mathcal{S})}{\delta}+\frac{1}{n}+d\varepsilon^2\,,$$

with probability at least $1-2\delta-2L\exp(-\Omega(m))$ over the randomness of initialization $\boldsymbol{W}$. □

H Discussion of Lipschitz property of score function

Here we provide an example to illustrate how the Lipschitz constant of the score function in a causal model is related to the model's nonlinear functions.

Here we give an example with $d=3$, the causality is $x^{(1)}\Rightarrow x^{(2)}\Rightarrow x^{(3)}$.

According to Eq. (9), we have that:

$$s_1(\boldsymbol{x})=-\frac{x^{(1)}}{\sigma_1^2}+\frac{\partial f_2(x^{(1)})}{\partial x^{(1)}}\frac{\epsilon_2}{\sigma_2^2},\quad s_2(\boldsymbol{x})=\frac{f_2(x^{(1)})-x^{(2)}}{\sigma_2^2}+\frac{\partial f_3(x^{(2)})}{\partial x^{(2)}}\frac{\epsilon_3}{\sigma_3^2},\quad s_3(\boldsymbol{x})=\frac{f_3(x^{(2)})-x^{(3)}}{\sigma_3^2}\,.$$

Then we can derive that:

$$\frac{\partial s_1(\boldsymbol{x})}{\partial x^{(2)}}=\frac{\partial s_1(\boldsymbol{x})}{\partial x^{(3)}}=\frac{\partial s_2(\boldsymbol{x})}{\partial x^{(3)}}=\frac{\partial s_3(\boldsymbol{x})}{\partial x^{(1)}}=\frac{\partial s_3(\boldsymbol{x})}{\partial x^{(2)}}=0\,,$$

$$\frac{\partial s_1(\boldsymbol{x})}{\partial x^{(1)}}=-\frac{1}{\sigma_1^2}+\frac{\partial^2 f_2(x^{(1)})}{\partial x^{(1)2}}\frac{\epsilon_2}{\sigma_2^2}\,,$$

$$\frac{\partial s_2(\boldsymbol{x})}{\partial x^{(2)}}=-\frac{1}{\sigma_2^2}+\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)2}}\frac{\epsilon_3}{\sigma_3^2}\,,$$

$$\frac{\partial s_2(\boldsymbol{x})}{\partial x^{(1)}}=\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)}\partial x^{(1)}}\frac{\epsilon_3}{\sigma_3^2}\,,$$

$$\frac{\partial s_3(\boldsymbol{x})}{\partial x^{(3)}}=-\frac{1}{\sigma_3^2}\,.$$

We denote $\boldsymbol{J}$ as the Jacobian of the score function. Then we can derive:

$$\begin{aligned}\|\boldsymbol{J}\|_{\ell_\infty}&=\max\left(\left|-\frac{1}{\sigma_1^2}+\frac{\partial^2 f_2(x^{(1)})}{\partial x^{(1)2}}\frac{\epsilon_2}{\sigma_2^2}\right|,\left|-\frac{1}{\sigma_2^2}+\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)2}}\frac{\epsilon_3}{\sigma_3^2}\right|+\left|\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)}\partial x^{(1)}}\frac{\epsilon_3}{\sigma_3^2}\right|,\frac{1}{\sigma_3^2}\right)\\&\le\max\left(\frac{1}{\sigma_1^2}+\left|\sup\frac{\partial^2 f_2(x^{(1)})}{\partial x^{(1)2}}\frac{\epsilon_2}{\sigma_2^2}\right|,\frac{1}{\sigma_2^2}+\left|\sup\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)2}}\frac{\epsilon_3}{\sigma_3^2}\right|+\left|\sup\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)}\partial x^{(1)}}\frac{\epsilon_3}{\sigma_3^2}\right|,\frac{1}{\sigma_3^2}\right).\end{aligned}$$

Then we have that for any L satisfy:

$$L\ge\max\left(\frac{1}{\sigma_1^2}+\left|\sup\frac{\partial^2 f_2(x^{(1)})}{\partial x^{(1)2}}\frac{\epsilon_2}{\sigma_2^2}\right|,\frac{1}{\sigma_2^2}+\left|\sup\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)2}}\frac{\epsilon_3}{\sigma_3^2}\right|+\left|\sup\frac{\partial^2 f_3(x^{(2)})}{\partial x^{(2)}\partial x^{(1)}}\frac{\epsilon_3}{\sigma_3^2}\right|,\frac{1}{\sigma_3^2}\right),$$

then the L is one of the Lipschitz constants of the score function.

According to the previous analysis, we can obtain the Lipschitz property of the score function by imposing some assumptions on the nonlinear function and noise of the model. For causal models with more complex DAG, the relationship between Lipshcitz and the model will be more complicated, but as long as the second derivatives of nonlinear functions are bounded and the variance of the noise is non-zero, the score function has Lipschitz property, and the value of the Lipschitz constant depends on the nonlinear function and noise of the model.

I Broader Impacts

This is a theoretical work that provides theoretical analysis for causal inference based on score matching. As such, we do not expect our work to have negative societal bias, as we do not focus on obtaining state-of-the-art results in a particular task. On the contrary, our work can have various benefits for the community:

- Causal inference is crucial in fields such as medicine, social sciences, and economics for understanding the essence of phenomena and formulating effective intervention measures. The outcomes of this work not only provide researchers in these fields with more reliable and interpretable theoretical insights into causal inference, driving scientific advancements and societal development, but also the score matching-based causal inference methods can help uncover hidden causal effects and mechanisms, providing a scientific foundation for decision-making in areas such as social equity, educational policies, and medical interventions.
- The theoretical framework and methods developed in this work can inspire and inform other causal inference approaches, fostering interdisciplinary research and collaboration, and expanding the application scope of causal inference in different domains.