Score-based Causal Representation Learning:
Linear and General Transformations

Burak Varıcı*† BVARICI@ANDREW.CMU.EDU
Machine Learning Department
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Emre Acartürk* ACARTE@RPI.EDU
Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Karthikeyan Shanmugam   KARTHIKEYANVS@GOOGLE.COM
Google DeepMind India
Bengaluru 560043, India

Abhishek Kumar‡ ABHISHEK.MLWORK@GMAIL.COM
Amazon AGI, USA

Ali Tajer   TAJER@ECSE.RPI.EDU
Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Abstract

This paper addresses intervention-based causal representation learning (CRL) under a general nonparametric latent causal model and an unknown transformation that maps the latent variables to the observed variables. Linear and general transformations are investigated. The paper addresses both the *identifiability* and *achievability* aspects. Identifiability refers to determining algorithm-agnostic conditions that ensure the recovery of the true latent causal variables and the underlying latent causal graph. Achievability refers to the algorithmic aspects and addresses designing algorithms that achieve identifiability guarantees. By drawing novel connections between *score functions* (i.e., the gradients of the logarithm of density functions) and CRL, this paper designs a *score-based class of algorithms* that ensures both identifiability and achievability. First, the paper focuses on *linear* transformations and shows that one stochastic hard intervention per node suffices to guarantee identifiability. It also provides partial identifiability guarantees for soft interventions, including identifiability up to mixing with parents for general causal models and perfect recovery of the latent graph for sufficiently nonlinear causal models. Secondly, it focuses on *general* transformations and demonstrates that two stochastic hard interventions per node are sufficient for identifiability. This is achieved by defining a differentiable loss function whose global optima ensure identifiability for general CRL. Notably, one does *not* need to know which pair of interventional environments has the same node intervened. Finally, the theoretical results are empirically validated via experiments on structured synthetic data and image data.

Keywords: causal representation learning, causality, interventions

---

*. Equal contribution
†. The work was completed when B.V. was at Rensselaer Polytechnic Institute.
‡. Contributions of A.K. were done when the author was at Google Brain.

1. Overview

Causal representation learning (CRL) aims to form a causal understanding of the world by learning appropriate representations that support causal interventions, reasoning, and planning (Schölkopf et al., 2021). Specifically, CRL considers a data-generating process in which high-level latent causally-related variables are mapped to low-level, generally high-dimensional observed data through an *unknown* transformation. Formally, consider a causal Bayesian network (Pearl, 2009) encoded by a directed acyclic graph (DAG) $\mathcal{G}$ with n nodes and generating *causal* random variables $\mathbf{Z} \triangleq [Z_1, \ldots, Z_n]^\top$. These random variables are transformed by an *unknown* function $g : \mathbb{R}^n \to \mathbb{R}^d$ to generate the d-dimensional *observed* random variables $\mathbf{X} \triangleq [X_1, \ldots, X_d]^\top$ according to:

$$\mathbf{X} = g(\mathbf{Z}) \ . \tag{1}$$

CRL is the process of using the observed data $\mathbf{X}$ and recovering (i) **the causal structure** $\mathcal{G}$ and (ii) **the latent causal variables Z**. Achieving these implicitly involves another objective of recovering the unknown transformation g as well. Addressing CRL consists of two central questions:

- **Identifiability,** which refers to determining the necessary and sufficient conditions under which $\mathcal{G}$ and $\mathbf{Z}$ can be recovered. The scope of identifiability (e.g., perfect or partial) critically depends on the extent of information available about the data, the underlying causal structure, and the transformation. The nature of the identifiability results can be algorithm-agnostic and non-constructive without specifying how to recover $\mathcal{G}$ and $\mathbf{Z}$. In particular, this is the case when considering CRL under a general transform g without parametric assumptions. Furthermore, the study of identifiability also investigates necessary conditions, e.g., which type of data is required for identifiability, regardless of the algorithmic approach.

- **Achievability**, which complements identifiability and pertains to designing algorithms that can recover $\mathcal{G}$ and $\mathbf{Z}$ while maintaining identifiability guarantees. Achievability hinges on forming reliable estimates for the transformation g.

CRL from interventions. Identifiability is known to be impossible without additional supervision or sufficient statistical diversity among the samples of the observed data $\mathbf{X}$. As shown in (Hyvärinen and Pajunen, 1999; Locatello et al., 2019), this is the case even for the simpler settings in which the latent variables are statistically independent (i.e., graph $\mathcal{G}$ has no edges). On the other hand, using data generated under interventions in addition to the observational data generated by the underlying latent model creates useful statistical diversity. Specifically, an intervention on a set of causal variables alters the causal mechanisms that generate those variables. Note that these causal mechanisms capture the effect of parents on the child variable. Such interventions, even when imposing sparse changes in the statistical models, create variations in the observed data sufficient for learning latent causal representations. This has led to CRL via intervention as an important class of CRL problems, which in its general form has remained an open problem (Schölkopf et al., 2021). This paper addresses this open problem by drawing a novel connection to score functions, i.e., the gradients of the logarithm of density functions.

We note that we use the interventions as a weak form of supervision via having access to *only* the pair of distributions before and after an intervention. Such supervision can be quite flexible and bodes well for practical applications in genomics (Tejada-Lapuerta et al., 2025) and robotics (Lee et al., 2021). For instance, genomics experiments often involve interventions, which can be modeled as deterministic or stochastic interventions, depending on the experimental mechanism. In robotics, let us consider the causal variables to be the joint angles of a robotic arm. A stochastic intervention corresponds to setting the joint angle to take values within a suitable support. We note that the interventions here can be soft since the feasible support for the joint may still depend on the other joint angles after the intervention. In this setting, the observations are simply images of the entire arm captured by a camera positioned at a specific location.

Contributions.   This paper provides both identifiability and achievability results for CRL from interventions under general latent causal models and general transformations. We establish these results by uncovering hitherto unknown connections between *score functions* (i.e., the gradients of the logarithm of density functions) and CRL. We leverage these connections to design CRL algorithms that serve as constructive proofs for the identifiability and achievability results. We do not make any parametric assumption on the latent causal model, i.e., the relationships among elements of $\mathbf{Z}$ take any arbitrary form. For the transformation g, we consider a diffeomorphism (i.e., bijective such that both g and g^{-1} are continuously differentiable) onto its image. Our results are categorized into two main groups based on the form of g: (i) **linear** transformation, and (ii) **general** (nonparametric) transformation. We consider both stochastic hard and soft interventions, and our contributions are summarized below.

Linear transformations.   We first focus on linear transformations and investigate various extents of results (perfect and partial) given *one* intervention per node. In this setting, we consider both hard and soft interventions.

- ✔ On **identifiability from hard interventions**, we show that *one* hard intervention per node suffices to guarantee perfect identifiability (Theorem 16).

- ✔ On **identifiability from soft interventions**, we show that *one* soft intervention per node suffices to guarantee identifiability up to ancestors – transitive closure of the latent DAG is recovered and latent variables are recovered up to a linear function of their ancestors (Theorem 14).

- ✔ While establishing these results, we also establish partial identifiability of a latent variable given an intervention on *single* node (Theorem 11).

- ✔ We further tighten the previous results and show that when the latent causal model is sufficiently nonlinear, perfect DAG recovery becomes possible using soft interventions. Furthermore, we recover a latent representation that is Markov with respect to the latent DAG, preserving the true conditional independence relationships (Theorem 18).

- ✔ On **achievability**, we design an algorithm referred to as **L**inear **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (LSCALE-I), which achieves the identifiability guarantees under both soft and hard interventions. LSCALE-I first forms an encoder estimate using correlation matrices of the observed score differences and then uses the

learned encoder to estimate the transitive closure of the latent DAG. Then, in the case of hard interventions, it refines these outputs without requiring any conditional independence test.

General transformations. In this setting, we do not have any restriction on the transformation from the latent space to the observed space. In this general setting, our contributions are as follows.

✔ On **identifiability**, we show that observational data and *two* distinct hard interventions per node suffice to guarantee perfect identifiability (Theorem 29). This result generalizes the recent results in the literature in two ways. First, we do not require the commonly adopted *faithfulness* assumption on latent causal models. Secondly, we assume the learner does not know which pair of environments intervene on the same node. While proving these results, we also establish identifiability of a single variable given two hard interventions *only* for the said variable (Theorem 22).

✔ More importantly, on **achievability**, we design the first provably correct algorithm that recovers $\mathcal{G}$ and $\mathbf{Z}$ perfectly. This algorithm is referred to as **G**eneralized **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (GSCALE-I). We note that GSCALE-I requires only the score functions of observed variables as its inputs and computes those of the latent variables by leveraging the Jacobian of the decoders.

✔ We also establish new results that shed light on the role of observational data. Specifically, when two interventional environments per node are given in pairs, observational data is only needed for recovering the latent DAG (Theorem 25). Furthermore, we show that observational data can be dispensed with under a weak faithfulness condition on the latent causal model (Theorem 26). Remarkably, our results are tight in terms of the identifiability objectives, which we will discuss in Section 3.4.

✔ Finally, we show that extrapolation to unseen combinations of single-node interventions can be achieved *without performing CRL*, just by using score functions of the *observed* space (Section 6.4).

Before describing our novel methodology, we recall the general approach to solving CRL. Recovering the latent causal variables hinges on finding the inverse of g based on the observed data $\mathbf{X}$, which facilitates recovering $\mathbf{Z}$ via $\mathbf{Z} = g^{-1}(\mathbf{X})$. In other words, referring to (g^{-1}, g) as the *true* encoder-decoder pair, we search for an encoder that takes observed variables back to the true latent space. A valid encoder should necessarily have an associated decoder to ensure a perfect reconstruction of observed variables from estimated latent variables. However, infinitely many valid encoder-decoder pairs satisfy the reconstruction property. The purpose of using interventions is to inject variations into the observed data, which can help us distinguish among these solutions.

Score-based methodology. We start by showing that an intervention on a latent node induces changes only in the score function's coordinates corresponding to the intervened node and its parents. This is because the intervention changes only the causal mechanism (i.e., conditional distribution) of the intervened node, which is a function of the latent node and its parents. Furthermore, when we consider two hard interventions on the same node, such changes

will be limited only to the intervened node (parents intact). This implies that score changes of the latent variables $\mathbf{Z}$ are generally sparse across different environments. Furthermore, these score changes contain all the information about the latent causal structure. Motivated by these key properties, we formalize a score-based CRL framework, based on which we design provably correct distinct algorithms, LSCALE-I and GSCALE-I, for linear and general transformations, respectively, presented in Section 5 and Section 6. We briefly describe the key technique in GSCALE-I for general transformation via two interventions. LSCALE-I involves exploiting the linearity of the transformation to achieve identifiability using only one intervention per node.

Algorithm sketch for general transformations. Consider two interventional environments in which the same node is intervened on. As described in the preceding paragraph, we show that the score functions of the latent variables under these two environments differ only at the coordinate of the intervened node. Subsequently, the key idea of the score-based framework is that tracing these sparse changes in the score functions of the latent variables can guide finding reliable estimates for the inverse of the transformation g, which in turn facilitates estimating $\mathbf{Z}$. In particular, we will seek encoders that result in sparse score variations among the estimated latent variables, thereby matching the properties of the true latent variables. To this end, we consider a pair of interventional environments for each node and find an encoder that minimizes the variations in the score function across all pairs. We show that the encoder obtained via this procedure perfectly recovers $\mathcal{G}$ and $\mathbf{Z}$. An important process in this methodology is projecting the score changes in observed data to the latent space so that we do not need to estimate latent scores for each possible encoder. We show that this can be done by multiplying the observed score difference by the Jacobian of the decoder associated with the encoder. Therefore, recovering Z_i is facilitated by solving the following problem:

$$\text{For } i\text{-th int. env. pair}: \quad \min \left\| \mathbb{E}\left[\left| \text{Jac. of decoder} \times \big(\text{score difference of } \mathbf{X}\big) \right| \right] - \mathbf{e}_i \right\|^2, \quad (2)$$

in which $\mathbf{e}_i$ denotes i-th standard basis vector, reflecting the sparsity property. Score differences of observed variables are computed across two environments with the same intervened node. When we have two interventions for *each* latent variable, we solve this problem for $i \in [n]$ simultaneously to recover complete $\mathbf{Z}$, and subsequently, the graph $\mathcal{G}$. Finally, the minimization is performed over the set of valid encoder-decoder pairs that ensure perfect reconstruction of $\mathbf{X}$.

Organization. The rest of the paper is organized as follows. Section 2 provides an overview of the literature with the main focus on CRL via interventions. Section 3 provides the preliminaries for formulating the problem and specifies the notations and definitions used throughout the paper. Section 4 establishes the properties of score functions under interventions and presents the key lemmas and their implications, which will be used in our CRL framework. We present our CRL algorithms and results for linear transformations in Section 5, and general transformations in Section 6. Proofs of the building blocks of the results are presented in the main body of the paper, and the rest of the proofs are provided in the appendices. In Section 7, we empirically assess the performance of the proposed CRL algorithms for recovering the latent causal variables and the latent causal graph on both structured synthetic data, biological data, and image data. Section 8 concludes the paper with a discussion of the results and the future directions.
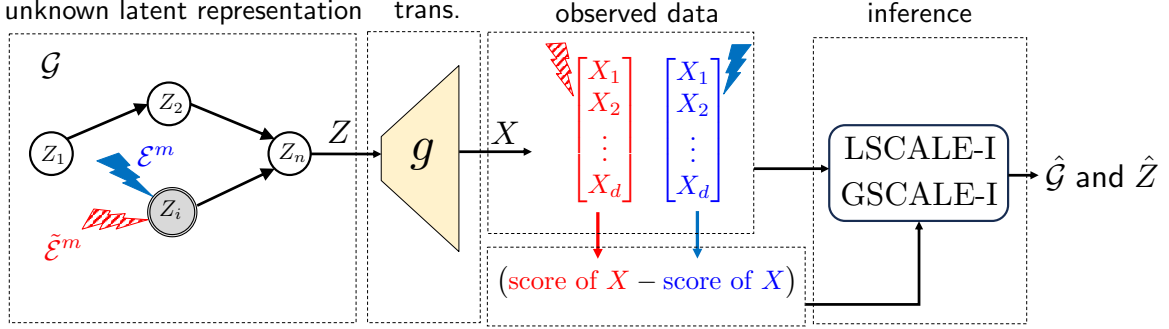
Figure 1: **An overview of LSCALE-I and GSCALE-I algorithms**. Corresponding to each latent variable Z_i, there are two interventional mechanisms, denoted by red and blue. For each pair of environments, the score functions of observed variables are fed into the GSCALE-I algorithm. Then, GSCALE-I uses this input to compute *latent score differences* for a given encoder and returns the encoder that minimizes these latent score differences. This encoder is used to compute estimates $\hat{\mathcal{G}}$ and $\hat{\mathbf{Z}}$.

2. Related Work

In this paper, we study CRL in different interventional environments where the interventions act on the latent space. We first provide an overview of the literature that investigates CRL from interventional data, with the main results of the most closely related work summarized in Table 1 and Table 2. Then, we discuss the other relevant lines of work.

Interventional causal representation learning. The majority of the rapidly growing literature on CRL from interventions focuses on *parametric* settings, i.e., a parametric form is assumed for the latent model, the transformation, or both. Among the related studies, that in (Ahuja et al., 2023) mainly considers polynomial transformations without restrictions on the latent causal model and shows identifiability under deterministic *do* interventions. It also establishes identifiability under general transformations, albeit requiring a combinatorial number of *do* interventions. Finally, it shows identifiability under soft interventions with independent support assumptions on latent variables. The study in (Squires et al., 2023) considers a linear latent model with a linear mapping to observations and proves identifiability under hard interventions. It also shows the impossibility of perfect identifiability under soft interventions and proves identifiability up to ancestors. The study in (Buchholz et al., 2023) focuses on linear Gaussian latent models and extends the results in (Squires et al., 2023) to prove identifiability for general transformations. Zhang et al. (2023) consider polynomial transformations under nonlinearity assumptions on latent models and prove identifiability up to ancestors under soft interventions. If the latent graph is restricted to polytrees, they further prove perfect identifiability. Identifying the non-intervened variables from the intervened variables by using single-node and multi-node soft interventions is studied in (Ahuja et al., 2024). It also considers a new setting in which the latent DAG can change across data points and relies on the support invariance of non-intervened variables to identify them from the rest. The study in (Bing et al., 2024) considers a nonlinear latent model under linear transformation and uses multi-target *do* interventions to prove identifiability under certain sparsity assumptions. Linear transformation and linear non-Gaussian latent

Table 1: **Parametric Settings.** Comparison of the results to prior studies in *parametric* settings. Only the main results of the most closely related studies are listed, and standard shared assumptions are omitted. Formal definitions of identifiability measures are provided in Section 3.

Work	Transform	Latent Model	Int. Data (one env. per node)	Identifiability result
Squires et al. (2023)	Linear	Linear	Hard	perfect
	Linear	Linear	Soft	up to ancestors
Ahuja et al. (2023)	Polynomial	General	*do*	perfect
	Polynomial	Bounded RV	Soft	perfect
Buchholz et al. (2023)	General	Lin. Gaussian	Hard	perfect
	General	Lin. Gaussian	Soft	up to ancestors
Zhang et al. (2023)	Polynomial	nonlinear	Soft	up to ancestors
	Polynomial	nonlinear (polytree)	Soft	perfect
Theorem 16	Linear	General	Hard	perfect
Theorem 14	Linear	General	Soft	up to ancestors
Theorem 18	Linear	nonlinear	Soft	perfect DAG and mixing w. surrounding

models are studied in (Jin and Syrgkanis, 2024), establishing identifiability up to surrounding parents using soft interventions. It also establishes sufficient conditions for multi-target soft interventions to ensure identifiability. The study in (Saengkyongam et al., 2024) takes a different approach and considers the task of intervention extrapolation. In this formulation, interventions are applied to exogenous action variables (e.g., instrumental variables) that linearly affect the latent variables.

On the fully *nonparametric* setting, von Kügelgen et al. (2023) provide the most closely related identifiability results to ours. Specifically, they show that two *coupled* hard interventions per node suffice for identifiability under the faithfulness assumption on latent causal models. Our results have two significant differences: 1) We address achievability via a provably correct algorithm whereas von Kügelgen et al. (2023) focus mainly on identifiability (e.g., no algorithm for recovery of the latent variables), 2) we dispense with the restrictive assumptions on identifiability results, namely, we do not require to know which two environments share the same intervention target (hence, *uncoupled* interventions), and do not require faithfulness on the latent models. While the importance of this relaxation is not fully apparent for single-node interventions, being able to work without knowing intervention targets is critical for investigating CRL under more realistic cases of multi-node interventions. Among the other studies on the nonparametric setting, Jin and Syrgkanis (2024) provide analogous results to (von Kügelgen et al., 2023) by considering two coupled soft interventions and identifying latent variables up to mixing with surrounding variables. Jiang and Aragam (2023) consider identifying the latent DAG without recovering latent variables, where it is shown that a restricted class of DAGs can be recovered. Finally, some studies employ stronger supervision signals, such as annotations of the ground-truth causal variables (Shen et al., 2022), or knowledge of the causal graph to recover the latent variables under hard interventions (Liang et al., 2023).

Multi-view causal representation learning. A commonly used form of weak supervision in CRL involves multi-view data, non-i.i.d. samples (or views) are observed, typically

generated by the same or closely related realizations of underlying latent variables. Multi-view scenarios can involve various data settings. Earlier studies have mostly considered paired data, in which pre- and post-intervention observations are generated from the same set of latent variables (Locatello et al., 2020; von Kügelgen et al., 2021; Brehmer et al., 2022). The more generalized and relaxed formulation typically involves partial observability, where multiple views are concurrently generated by an overlapping subset of latent variables, such as observing different camera angles of the same scene (Sturma et al., 2023). Yao et al. (2024) generalizes the multi-view approach via a unified framework, which also allows partial observability with nonlinear transforms. We also note that Yao et al. (2025) presents an even more general unifying view to interpret different CRL approaches, including interventional and multi-view, as special ways of aligning the representations to known data symmetries via leveraging the invariance principle.

Temporal causal representation learning.   Temporal CRL is particularly motivated by applications in domains where time-series data is readily available, such as robotics and control systems, which has led to a growing interest in this area. Some studies in this category assume that the latent causal variables are mutually independent, meaning there are no instantaneous causal effects between variables at the same time step (Lippe et al., 2022; Yao et al., 2022; Lachapelle et al., 2024). However, more recent studies have relaxed this assumption, allowing instantaneous causal effects when sufficient diversity is present in the observed or interventional data (Lippe et al., 2023a,b; Li et al., 2024b).

Other approaches.   In some other approaches to CRL, Zhang et al. (2024) consider general mixing functions and general SCMs under sparsity constraints; Liu et al. (2024) leverage nonlinear ICA principles to work with nonlinear mixing functions and polynomial latent SCMs; Morioka and Hyvärinen (2024) consider disjoint groups of observational variables, with general mixing functions and pairwise latent causal relationships; Welch et al. (2024) use score functions for linear mixing functions and nonlinear Gaussian SCMs to derive coarser identifiability results; and Li et al. (2024a) study "domains" and interventions in a combined way for non-Markovian causal systems when the graph is known.

Identifiable representation learning.   As a special case of CRL, where the latent variables are independent, there is extensive literature on identifying latent representations. Some representative approaches include leveraging knowledge of the mechanisms that govern the system's evolution (Ahuja et al., 2022) and using weak supervision with auxiliary information (Shu et al., 2020). Nonlinear independent component analysis (ICA) also uses side information, in the form of structured time series, to exploit temporal information (Hyvärinen and Morioka, 2017; Hälvä and Hyvärinen, 2020) or knowledge of auxiliary variables that make latent variables conditionally independent (Khemakhem et al., 2020a,b; Hyvärinen et al., 2019). Morioka and Hyvärinen (2024) imposes additional constraints on observational mixing and the causal model to prove identifiability. Kivva et al. (2022) studies the identifiability of deep generative models without auxiliary information.

Score functions for causal discovery within observed variables.   Score matching has recently gained attention in the causal discovery of observed variables. Rolland et al. (2022) use score matching to recover nonlinear additive Gaussian noise models. Montagna et al. (2023b) focus on the same setting, recover the full graph from Jacobian scores, and

Table 2: **General (Nonparametric) Settings.** Comparison of the results to prior studies in the *general* setting. Formal definitions of identifiability measures are provided in Section 3. The shared assumptions (interventional discrepancy) and additional assumptions (faithfulness) are discussed in Section 6.

Work	Transform and Latent Model	Obs. Data	Int. Data (env. per node)	Faithfulness	Identifiability result	Provable algorithm
von Kügelgen et al. (2023)	General	No	2 coupled hard	Yes	perfect	✘
Jin and Syrgkanis (2024)	General	No	2 coupled soft	No	perfect DAG and mix w. surrounding	✘
Theorem 25	General	Yes	2 coupled hard	No	perfect	✔
Theorem 26	General	No	2 coupled hard	Yes	perfect	✔
Theorem 29	General	Yes	2 uncoupled hard	No	perfect	✔

dispense with the computationally expensive pruning stage of the algorithm in (Rolland et al., 2022). Montagna et al. (2023a) empirically demonstrate the robustness of score matching approaches against the assumption violations in causal discovery. Zhu et al. (2023) establish bounds on the error rate of score matching-based causal discovery methods. All of these studies are limited to observed causal variables, whereas in our case, we have a causal model in the latent space.

3. Preliminaries and Definitions

Notations. For a vector $\mathbf{a} \in \mathbb{R}^m$, the i-the entry is denoted by a_i. Random vectors are denoted by bold upper-case letters and their realizations are denoted by bold lower-case letters, e.g., $\mathbf{X}$ and $\mathbf{x}$. Matrices are denoted by bold upper-case letters, e.g., $\mathbf{A}$, where $\mathbf{A}_i$ denotes the i-th row of $\mathbf{A}$ and $\mathbf{A}_{i,j}$ denotes the entry at row i and column j. For matrices $\mathbf{A}$ and $\mathbf{B}$ with the same shapes, $\mathbf{A} \preccurlyeq \mathbf{B}$ denotes component-wise inequality. We denote the indicator function by $\mathbb{1}$, and for a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we use the convention that $\mathbb{1}\{\mathbf{A}\} \in \{0,1\}^{m \times n}$, where the entries are specified by $[\mathbb{1}\{\mathbf{A}\}]_{i,j} = \mathbb{1}\{\mathbf{A}_{i,j} \neq 0\}$. For a positive integer $n \in \mathbb{N}$, we define $[n] \triangleq \{1, \ldots, n\}$. The permutation matrix associated with any permutation π of $[n]$ is denoted by $\mathbf{P}_\pi$, i.e., $[\pi_1 \ \pi_2 \ \ldots \ \pi_n]^\top = \mathbf{P}_\pi \cdot [1 \ 2 \ \ldots \ n]^\top$. The n-dimensional identity matrix is denoted by $\mathbf{I}_{n \times n}$, and the Hadamard product is denoted by $\odot$. We use $\mathrm{im}(f)$ to denote the image of the function f. We let $\mathbf{e}_i$ denote the i-th standard basis vector of $\mathbb{R}^n$. Given a function $f : \mathbb{R}^s \to \mathbb{R}^r$ that has first-order partial derivatives on $\mathbb{R}^s$, we denote the Jacobian of f at $\mathbf{z} \in \mathbb{R}^s$ by $J_f(\mathbf{z}) \in \mathbb{R}^{r \times s}$ with entries $[J_f(\mathbf{z})]_{i,j} = \partial f(\mathbf{z})_i / \partial \mathbf{z}_j$.

3.1 Latent Causal Structure

Consider latent causal random variables $\mathbf{Z} \triangleq [Z_1, \ldots, Z_n]^\top$. An *unknown* transformation $g : \mathbb{R}^n \to \mathbb{R}^d$ generates the observable random variables $\mathbf{X} \triangleq [X_1, \ldots, X_d]^\top$ from the latent variables $\mathbf{Z}$ according to:

$$\mathbf{X} = g(\mathbf{Z}) \ . \tag{3}$$

We assume that $d \geq n$, and the transformation g is continuously differentiable and a diffeomorphism onto its image (otherwise, identifiability is ill-posed). We denote the image of g by $\mathcal{X} \triangleq \mathrm{im}(g) \subseteq \mathbb{R}^d$. The probability density functions (pdfs) of $\mathbf{Z}$ and $\mathbf{X}$ are denoted by p and $p_\mathbf{X}$, respectively. We assume that p is absolutely continuous with respect to the

n-dimensional Lebesgue measure. Subsequently, $p_{\mathbf{X}}$, which is defined on the image manifold $\text{im}(g)$, is absolutely continuous with respect to the n-dimensional Hausdorff measure rather than d-dimensional Lebesgue measure[1]. The distribution of latent variables $\mathbf{Z}$ factorizes with respect to a DAG that consists of n nodes and is denoted by $\mathcal{G}$. Node $i \in [n]$ of $\mathcal{G}$ represents Z_i and p factorizes according to:

$$p(\mathbf{z}) = \prod_{i=1}^{n} p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) \ , \tag{4}$$

where $\text{pa}(i)$ denotes the set of parents of node i and $p_i(z_i \mid \mathbf{z}_{\text{pa}(i)})$ is the conditional pdf of z_i given the variables of its parents. We use $\text{ch}(i)$, $\text{an}(i)$, and $\text{de}(i)$ to denote the children, ancestors, and descendants of node i, respectively. Accordingly, for each node $i \in [n]$ we also define

$$\overline{\text{pa}}(i) \triangleq \text{pa}(i) \cup \{i\}, \ \ \overline{\text{ch}}(i) \triangleq \text{ch}(i) \cup \{i\}, \ \ \overline{\text{an}}(i) \triangleq \text{an}(i) \cup \{i\}, \ \text{and} \ \overline{\text{de}}(i) \triangleq \text{de}(i) \cup \{i\} \ . \tag{5}$$

We denote the transitive closure and transitive reduction of $\mathcal{G}$ by $\mathcal{G}_{\text{tc}}$ and $\mathcal{G}_{\text{tr}}$, respectively[2]. The parental relationships in these graphs are denoted by pa_{tc} and pa_{tr}, and other graphical relationships are denoted similarly. Based on the modularity property, a change in the causal mechanism of node i does not affect those of the other nodes. We also assume that all conditional pdfs $\{p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) : i \in [n]\}$ are continuously differentiable with respect to all variables and $p(\mathbf{z}) \neq 0$ for all $\mathbf{z} \in \mathbb{R}^n$. We consider the general structural causal models (SCMs) based on which for each $i \in [n]$,

$$Z_i = f_i(\mathbf{Z}_{\text{pa}(i)}, N_i) \ , \tag{6}$$

where $\{f_i : i \in [n]\}$ are general functions that capture the dependence of node i on its parents and $\{N_i : i \in [n]\}$ account for the exogenous noise terms that we assume to have pdfs with full support. We specialize some of the results to additive noise SCMs, in which (6) becomes

$$Z_i = f_i(\mathbf{Z}_{\text{pa}(i)}) + N_i \ . \tag{7}$$

Next, we provide several definitions that we will use frequently throughout the paper to formalize the framework and analyze it.

Definition 1 (Valid Causal Order) *We refer to a permutation $(\pi_1, \ldots, \pi_n)$ of $[n]$ as a valid causal order[3] if $\pi_i \in \text{pa}(\pi_j)$ indicates that $i < j$.*

In this paper, without loss of generality, we assume that $(1, \ldots, n)$ is a valid causal order. We also define a graphical notion that will be useful for presenting our results and analysis on CRL under a linear transformation.

---

1. For details of where this has been used, see Appendix A.2.
2. Transitive closure of a DAG $\mathcal{G}$, denoted by $\mathcal{G}_{\text{tr}}$, is a DAG with parents denoted by $\text{pa}_{\text{tr}}(i) = \text{an}(i)$ for each node i. The transitive reduction of $\mathcal{G}$ is the DAG with the fewest edges that preserves the same reachability relation as $\mathcal{G}$.
3. It is also called *topological ordering* or *topological sort* in the literature.

Figure 2: Sample latent DAGs $\mathcal{G}$. **(Left)** The valid causal orders are $(1, 2, 3, 4)$ and $(1, 3, 2, 4)$, and only surrounded node is 4 with $\mathrm{sur}(4) = \{2, 3\}$. **(Right)** The only valid causal order is $(1, 2, 3)$, and the surrounded nodes are $\mathcal{S} = \{2, 3\}$ with $\mathrm{sur}(2) = \{1\}, \mathrm{sur}(3) = \{1, 2\}$.

Definition 2 (Surrounded Node) *Node $i \in [n]$ in DAG $\mathcal{G}$ is said to be* surrounded *if there exists another node $j \in [n]$ such that $\overline{\mathrm{ch}}(i) \subseteq \mathrm{ch}(j)$. We denote the set of nodes that surround $i \in [n]$ by $\mathrm{sur}(i)$, and the set of all nodes that are surrounded by $\mathcal{S}$, i.e.,*

$$\mathrm{sur}(i) \triangleq \{j \in [n] : j \neq i \ , \ \overline{\mathrm{ch}}(i) \subseteq \mathrm{ch}(j)\} \ , \quad and \quad \mathcal{S} \triangleq \{i \in [n] : \mathrm{sur}(i) \neq \emptyset\} \ . \quad (8)$$

The intuition behind surrounded nodes is that, the effect of i on its children $\mathrm{ch}(i)$ can be dominated by the effect of its surrounding node j.[4] Specifically, any effect of i on a node $k \in \mathrm{ch}(i)$ can also be interpreted as the effect of node j since $k \in \mathrm{ch}(j)$ as well. This effect causes ambiguities in the recovery of Z_i in the case of soft interventions.

3.2 Score Functions

The *score function* associated with a pdf is defined as the gradient of its logarithm. The score function associated with p is denoted by

$$\boldsymbol{s}(\mathbf{z}) \triangleq \nabla_{\mathbf{z}} \log p(\mathbf{z}) \ . \quad (9)$$

Noting the connection $\mathbf{X} = g(\mathbf{Z})$, the density of $\mathbf{X}$ under $\mathcal{E}^0$, denoted by $p_{\mathbf{X}}$, is supported on an n-dimensional manifold $\mathcal{X}$ *embedded* in $\mathbb{R}^d$. Hence, specifying the score function of $\mathbf{X}$ requires notions from differential geometry. For this purpose, we denote the tangent space of the manifold $\mathcal{X}$ at point $\mathbf{x} \in \mathcal{X}$ by $T_{\mathbf{x}}\mathcal{X}$. Tangent vectors $\mathbf{v} \in T_x \mathcal{X}$ are equivalence classes of continuously differentiable curves $\gamma \colon (-1, 1) \to \mathcal{X} \subseteq \mathbb{R}^d$ with $\gamma(0) = \mathbf{x}$ and $\gamma'(0) = \mathbf{v}$. Furthermore, given a function $f \colon \mathcal{X} \to \mathbb{R}$, denote its directional derivative at point $\mathbf{x} \in \mathcal{X}$ along a tangent vector $\mathbf{v} \in T_{\mathbf{x}}\mathcal{X}$ by $D_{\mathbf{v}} f(\mathbf{x})$, which is defined as

$$D_{\mathbf{v}} f(\mathbf{x}) \triangleq \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t) \Big|_{t=0} \ , \quad (10)$$

for any curve γ in equivalence class v. The differential of f at point $\mathbf{x} \in \mathcal{X}$, denoted by $\mathrm{d}f_{\mathbf{x}}$, is the linear operator mapping tangent vector $\mathbf{v} \in T_{\mathbf{x}}\mathcal{X}$ to $D_{\mathbf{v}} f_{\mathbf{x}}$ (Simon, 2014, p. 57), i.e.,

$$\mathrm{d}f_x \colon T_{\mathbf{x}}\mathcal{X} \ni v \mapsto D_{\mathbf{v}} f(\mathbf{x}) \in \mathbb{R} \ . \quad (11)$$

---

4. Surrounded node concept is first defined by (Varıcı et al., 2023), and later adopted by Jin and Syrgkanis (2024) when considering soft interventions.

Let $\mathbf{B} \in \mathbb{R}^{d \times n}$ be a matrix for which the columns of $\mathbf{B}$ form an orthonormal basis for $T_{\mathbf{x}}\mathcal{X}$. Denote the directional derivative of f along the i-th column of $\mathbf{B}$ by $D_i f$ for all $i \in [n]$. Then, the differential operator can be expressed by the vector

$$D f_{\mathbf{x}} \triangleq \mathbf{B} \cdot \left[D_1 f_{\mathbf{x}} \ldots D_n f_{\mathbf{x}} \right]^{\top} \in \mathbb{R}^d \ , \tag{12}$$

such that

$$\mathrm{d} f_{\mathbf{x}}(\mathbf{v}) = \mathbf{v}^{\top} \cdot D f_{\mathbf{x}} \ , \qquad \forall \mathbf{x} \in \mathcal{X} \ , \quad \forall \mathbf{v} \in T_{\mathbf{x}}\mathcal{X} \ . \tag{13}$$

Note that the differential operator $\mathrm{d} f$ is a generalization of the gradient. Hence, we can generalize the definition of the score function using the differential operator by setting f to the logarithm of pdf. Therefore, the score function of $\mathbf{X}$ under $\mathcal{E}^0$ is specified as follows:

$$s_{\mathbf{X}}(\mathbf{x}) \triangleq D \log p_{\mathbf{X}}(\mathbf{x}) \ , \qquad \forall \mathbf{x} \in \mathcal{X} \ . \tag{14}$$

3.3 Intervention Mechanisms

We consider two types of interventions. A *soft* intervention on node i (also referred to as *imperfect* intervention in literature), changes the conditional distribution $p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$ to a distinct conditional distribution, which we denote by $q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$. A soft intervention does not necessarily remove the functional dependence of an intervened node on its parents but rather alters it to a different mechanism. A stochastic *hard* intervention on node i (also referred to as *perfect* intervention) is stricter than a soft intervention and removes the edges incident on i. A hard intervention on node changes $p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$ to $q_i(z_i)$ that emphasizes the lack of dependence of z_i on $\mathbf{z}_{\mathrm{pa}(i)}$. Finally, we note that in some settings, we assume two hard interventions per node, in which case the two hard interventional mechanisms for node i are denoted by two distinct pdfs $q_i(z_i)$ and $\tilde{q}_i(z_i)$.

Interventional environments. We consider atomic *interventional* environments in which each environment one node is intervened in, as is customary in the closely related CRL literature (Squires et al., 2023; Ahuja et al., 2023; Buchholz et al., 2023). In some settings (linear transformation), we will have *one* interventional environment per node and denote the interventional environments by $\mathcal{E} \triangleq \{\mathcal{E}^1, \ldots, \mathcal{E}^n\}$, where we call $\mathcal{E}$ the *atomic* environment set. We denote the node intervened in environment $\mathcal{E}^m$ by $I^m \in [n]$. For other settings (general transformation), we will have *two* interventional environments per node and denote the second atomic environment set by $\tilde{\mathcal{E}} = \{\tilde{\mathcal{E}}^1, \ldots, \tilde{\mathcal{E}}^n\}$. Similarly, we denote the intervened node in $\tilde{\mathcal{E}}^m$ by $\tilde{I}^m$ for each $m \in [n]$. We assume that node-environment pairs are *unspecified*, i.e., the ordered intervention sets $\mathcal{I} \triangleq (I^1, \ldots, I^n)$ and $\tilde{\mathcal{I}} \triangleq (\tilde{I}^1, \ldots, \tilde{I}^n)$ are two *unknown* permutations of $[n]$. We also adopt the convention that $\mathcal{E}^0$ is the *observational* environment and $I^0 \triangleq \emptyset$. Next, we define the notion of coupling between the environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$.

Definition 3 (Coupled/Uncoupled Environments) *The two environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$ are said to be **coupled** if for the unknown permutations $\mathcal{I}$ and $\tilde{\mathcal{I}}$ we know that $\mathcal{I} = \tilde{\mathcal{I}}$, i.e., the same node is intervened in environments $\mathcal{E}^i$ and $\tilde{\mathcal{E}}^i$. The two environment sets are said to be **uncoupled** if $\tilde{\mathcal{I}}$ is an unknown permutation of $\mathcal{I}$.*

Next, we define p^m as the pdf of $\mathbf{Z}$ in environment $\mathcal{E}^m$. Hence, under soft and hard intervention for each $m \in [n]$, p^m can be factorized as follows.

$$\text{soft int. in } \mathcal{E}^m : \quad p^m(\mathbf{z}) = q_\ell(z_\ell \mid \mathbf{z}_{\text{pa}(\ell)}) \prod_{i \neq \ell} p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) , \qquad \text{where} \quad \ell = I^m , \quad (15)$$

$$\text{hard int. in } \mathcal{E}^m : \quad p^m(\mathbf{z}) = q_\ell(z_\ell) \prod_{i \neq \ell} p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) , \qquad \text{where} \quad \ell = I^m . \quad (16)$$

Similarly, we define $\tilde{p}^m$ as the pdf of $\mathbf{Z}$ in $\tilde{\mathcal{E}}^m$, which can be factorized similarly to (16) with $\tilde{q}_\ell$ replaced with q_ℓ. Hence, the score functions associated with p^m and $\tilde{p}^m$ are specified as follows.

$$\boldsymbol{s}^m(\mathbf{z}) \triangleq \nabla_{\mathbf{z}} \log p^m(\mathbf{z}) , \quad \text{and} \quad \tilde{\boldsymbol{s}}^m(\mathbf{z}) \triangleq \nabla_{\mathbf{z}} \log \tilde{p}^m(\mathbf{z}) . \quad (17)$$

We denote the score functions of the observed variables $\mathbf{X}$ under $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ by $\boldsymbol{s}_{\mathbf{X}}^m$ and $\tilde{\boldsymbol{s}}_{\mathbf{X}}^m$, respectively. Note that the score functions change across different environments, which is induced by the changes in the distribution of $\mathbf{Z}$. Specifically, following (4) and (15), the latent scores $\boldsymbol{s}(\mathbf{z})$ and $\boldsymbol{s}^m(\mathbf{z})$ are decomposed as

$$\boldsymbol{s}(\mathbf{z}) = \nabla_{\mathbf{z}} \log p_\ell(z_\ell \mid \mathbf{z}_{\text{pa}(\ell)}) + \sum_{i \neq \ell} \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) , \quad (18)$$

$$\text{and} \quad \boldsymbol{s}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_\ell(z_\ell \mid \mathbf{z}_{\text{pa}(\ell)}) + \sum_{i \neq \ell} \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) . \quad (19)$$

where $\ell = I^m$. Hence, $\boldsymbol{s}(\mathbf{z})$ and $\boldsymbol{s}^m(\mathbf{z})$ differ in only the causal mechanism of the intervened node in environment $\mathcal{E}^m$. In Section 4, we investigate these discrepancies between $\boldsymbol{s}$ and $\boldsymbol{s}^m$ (or $\tilde{\boldsymbol{s}}^m$) and characterize the relationship between the scores in the observational and interventional environments.

3.4 Identifiability and Achievability Objectives

The objective of CRL is to use observations $\mathbf{X}$ generated by the observational and interventional environments and estimate the true latent variables $\mathbf{Z}$ and causal relations among them captured by $\mathcal{G}$. The first objective is *identifiability*, which pertains to determining algorithm-agnostic sufficient conditions under which $\mathbf{Z}$ and $\mathcal{G}$ can be recovered uniquely up to a permutation and element-wise transform, which is the strongest form of recovery in CRL from interventions as shown in (von Kügelgen et al., 2023). The second objective is *achievability*, which refers to designing algorithms that are amenable to practical implementation and generate provably correct estimates for $\mathbf{Z}$ and $\mathcal{G}$, foreseen by the identifiability guarantees. In this subsection, we provide the definitions needed for formalizing these objectives.

We denote a generic estimator of $\mathbf{Z}$ given $\mathbf{X}$ by $\hat{\mathbf{Z}}(\mathbf{X}) : \mathbb{R}^d \to \mathbb{R}^n$. We also consider a generic estimate of $\mathcal{G}$ denoted by $\hat{\mathcal{G}}$. To assess the fidelity of the estimates $\hat{\mathbf{Z}}(\mathbf{X})$ and $\hat{\mathcal{G}}$ with respect to the ground truth $\mathbf{Z}$ and $\mathcal{G}$, we provide the following identifiability measures, which will be achieved when we have a complete set of atomic interventions.

Definition 4 (Latent Graph Identifiability) *For identifiability of the latent graph, we define:*

1. **Perfect DAG recovery:** *DAG recovery is said to be perfect if $\hat{\mathcal{G}}$ is isomorphic to $\mathcal{G}$.*

2. **Transitive closure recovery:** *DAG recovery is said to maintain transitive closure if $\hat{\mathcal{G}}$ and $\mathcal{G}$ have the same ancestral relationships, i.e., $\hat{\mathcal{G}}_{\text{tr}}$ is isomorphic to $\mathcal{G}_{\text{tr}}$.*

Definition 5 (Latent Variable Identifiability) *For identifiability of* all *latent variables, we define:*

1. **Componentwise latent recovery:** *The estimator $\hat{\mathbf{Z}}(\mathbf{X})$ satisfies componentwise latent recovery if $\hat{\mathbf{Z}}(\mathbf{X})$ is a componentwise diffeomorphism of a permutation of $\mathbf{Z}$, i.e., there exists a permutation π of $[n]$ and a set diffeomorphisms $\{\phi_i : \mathbb{R} \to \mathbb{R} : i \in [n]\}$ such that we have*

$$\hat{\mathbf{Z}}(\mathbf{X}) = (\pi \circ \phi)(\mathbf{Z}) \ , \qquad \forall \mathbf{z} \in \mathbb{R}^n \ , \tag{20}$$

   where $\phi(\mathbf{Z}) \triangleq (\phi_1(Z_1), \ldots, \phi_n(Z_n))$.

2. **Scaling consistency:** *The estimator $\hat{\mathbf{Z}}(\mathbf{X})$ satisfies scaling consistency if there exists a permutation π of $[n]$ and a constant* diagonal *matrix $\mathbf{C}_{\text{scale}} \in \mathbb{R}^{n \times n}$ such that*

$$\hat{\mathbf{Z}}(\mathbf{X}) = \mathbf{P}_\pi \cdot \mathbf{C}_{\text{scale}} \cdot \mathbf{Z} \ , \qquad \forall \mathbf{z} \in \mathbb{R}^n \ . \tag{21}$$

3. **Consistency up to mixing with parents:** *The estimator $\hat{\mathbf{Z}}(\mathbf{X})$ satisfies consistency up to mixing with parents if there exists a permutation π of $[n]$ and a constant matrix $\mathbf{C}_{\text{pa}} \in \mathbb{R}^{n \times n}$ such that*

$$\hat{\mathbf{Z}}(\mathbf{X}) = \mathbf{P}_\pi \cdot \mathbf{C}_{\text{pa}} \cdot \mathbf{Z} \ , \qquad \forall \mathbf{z} \in \mathbb{R}^n \ , \tag{22}$$

   where $\mathbf{C}_{\text{pa}}$ has nonzero diagonal entries and for all $j \notin \overline{\text{pa}}(i)$, $[\mathbf{C}_{\text{pa}}]_{i,j} = 0$. Equivalently, $\hat{Z}_{\pi_i}$ is a linear function of $\{Z_j : j \in \overline{\text{pa}}(i)\}$ for all $i \in [n]$.

4. **Consistency up to mixing with surrounding parents:** *The estimator $\hat{\mathbf{Z}}(\mathbf{X})$ satisfies consistency up to mixing with surrounding parents if there exists a permutation π of $[n]$ and a constant matrix $\mathbf{C}_{\text{sur}} \in \mathbb{R}^{n \times n}$ such that*

$$\hat{\mathbf{Z}}(\mathbf{X}) = \mathbf{P}_\pi \cdot \mathbf{C}_{\text{sur}} \cdot \mathbf{Z} \ , \qquad \forall \mathbf{z} \in \mathbb{R}^n \ , \tag{23}$$

   where $\mathbf{C}_{\text{sur}}$ has nonzero diagonal entries and for all $j \notin \overline{\text{sur}}(i)$, $[\mathbf{C}_{\text{sur}}]_{i,j} = 0$. Equivalently, $\hat{Z}_{\pi_i}$ is a linear function of $\{Z_j : j \in \overline{\text{sur}}(i)\}$ for all $i \in [n]$.

We note that scaling consistency is a special case of component-wise latent recovery in which the diffeomorphism is restricted to scaling. Next, we provide *node-level partial identifiability* definitions which measure the recovery level of a *single* latent variable. These will be useful to assess the identifiability guarantees of the algorithms under an incomplete set of interventions.

Definition 6 (Node-level Partial Identifiability) *For identifiability of a* single *latent variable, we define:*

1. **Partial latent recovery:** *The estimator* $\hat{\mathbf{Z}}(\mathbf{X})$ *satisfies partial latent recovery* **for node** i *if*

$$[\hat{\mathbf{Z}}(\mathbf{X})]_i = \phi_k(Z_k) \ , \tag{24}$$

   for some $k \in [n]$ *where* $\phi_k : \mathbb{R} \to \mathbb{R}$ *is a diffeomorphism.*

2. **Partial consistency up to mixing with parents:** *The estimator* $\hat{\mathbf{Z}}(\mathbf{X})$ *satisfies partial consistency up to mixing with parents* **for node** i *if*

$$[\hat{\mathbf{Z}}(\mathbf{X})]_i = \sum_{j \in \overline{\mathrm{pa}}(k)} c_k \cdot Z_k \ , \tag{25}$$

   for some $k \in [n]$ *and constants* $\{c_j : j \in \overline{\mathrm{pa}}(k)\}.$

Tightness of the identifiability definitions.   For general transformations, von Kügelgen et al. (2023, Proposition 3.1) show that identifiability up to component-wise diffeomorphisms is the best possible result under interventions, without additional assumptions. For linear transformations, Squires et al. (2023, Proposition 2) and Buchholz et al. (2023, Remark 2) show that under hard interventions, identifiability up to scaling consistency and permutation ambiguity is the best achievable outcome without further information on $\mathbf{Z}$. Furthermore, for soft interventions, transitive closure recovery is the best latent graph identifiability result without additional assumptions on the causal model (Squires et al., 2023, Appendix J). Finally, Jin and Syrgkanis (2024, Theorem 3) show that, under some non-degeneracy assumptions, consistency up to mixing with surrounding parents is the optimal result for soft interventions. Therefore, given a complete set of atomic interventions, the identifiability definitions presented in this section represent the ultimate objectives when using interventions.

3.5   Algorithm-related Definitions

For formalizing the achievability results and designing the associated algorithms, generating the estimates $\hat{\mathbf{Z}}(\mathbf{X})$ and $\hat{\mathcal{G}}$ is facilitated by estimating the inverse of g based on the observed data $\mathbf{X}$. Specifically, an estimate of g^{-1}, where g^{-1} denotes the inverse of g, facilitates recovering $\mathbf{Z}$ via $\mathbf{Z} = g^{-1}(\mathbf{X})$. Throughout the rest of this paper, we refer to g^{-1} as the *true encoder*. To formalize the procedures of estimating g^{-1}, we define $\mathcal{H}$ as the set of possible valid encoders, i.e., candidates for g^{-1}. A function $h \in \mathcal{H}$ can be such a candidate if it is invertible, that is, there exists an associated decoder h^{-1} such that $(h^{-1} \circ h)(\mathbf{X}) = \mathbf{X}$. Hence, the set of valid encoders is specified by

$$\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathbb{R}^n \ : \ \exists h^{-1} : \mathbb{R}^n \to \mathbb{R}^d \ \text{such that} \ \ (h^{-1} \circ h)(\mathbf{x}) = \mathbf{x} \ , \ \ \forall \mathbf{x} \in \mathcal{X} \} \ . \tag{26}$$

Next, corresponding to any pair of observation $\mathbf{X}$ and valid encoder $h \in \mathcal{H}$, we define $\hat{\mathbf{Z}}(\mathbf{X}; h)$ as an *auxiliary* estimate of $\mathbf{Z}$ generated by applying the valid encoder h on $\mathbf{X}$, i.e.,

$$\hat{\mathbf{Z}}(\mathbf{X}; h) \triangleq h(\mathbf{X}) = (h \circ g)(\mathbf{Z}) \ , \qquad \forall \, h \in \mathcal{H} \ . \tag{27}$$

The estimate $\hat{\mathbf{Z}}(\mathbf{X}; h)$ inherits its randomness from $\mathbf{X}$, and its statistical model is governed by that of $\mathbf{Z}$ under the chosen h. To emphasize the dependence on h, we denote the

score functions associated with the pdfs of $\hat{\mathbf{Z}}(\mathbf{X}; h)$ under environments $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$, respectively, by

$$s_{\hat{\mathbf{Z}}}(\cdot; h) , \quad s_{\hat{\mathbf{Z}}}^m(\cdot; h), \quad \text{and} \quad \tilde{s}_{\hat{\mathbf{Z}}}^m(\cdot; h) . \tag{28}$$

We will be addressing both general and linear transformations g. In the linear transformation setting, the true linear transformation g is denoted by the matrix $\mathbf{G} \in \mathbb{R}^{d \times n}$. Accordingly, we denote a valid linear encoder by $\mathbf{H} \in \mathbb{R}^{n \times d}$. For a given valid encoder $\mathbf{H}$, the associated valid decoder is given by its Moore-Penrose inverse, i.e., $\mathbf{H}^\dagger \triangleq \mathbf{H}^\top \cdot (\mathbf{H} \cdot \mathbf{H}^\top)^{-1}$.

4. Properties of Score Functions under Interventions

Score functions and their variations across different interventional environments play pivotal roles in our approach to identifying latent representations. In this section, we present the key properties of the score functions that will be leveraged in Section 5 and Section 6 to construct identifiability results, along with the corresponding algorithms.

We first investigate score variations across pairs of environments, such as the observational environment and an interventional one (under both soft and hard atomic interventions) or two interventional environments, either coupled or uncoupled. The key insight is that an intervention causes changes in only certain coordinates of the score function, as indicated by the sparse changes in the decompositions of the score functions in (18) and (19). These sparse changes further reflect the graph structure in the score differences. For instance, for a single-node intervention $I^m = \{i\}$ in m-th environment, j-th coordinate of the score difference $s - s^m$ becomes

$$\left[s(\mathbf{z}) - s^m(\mathbf{z}) \right]_j = \frac{\partial}{\partial z_j} \log p_i(z_i \mid \mathbf{z}_{\text{pa}(i)})) - \frac{\partial}{\partial z_j} \log q_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) , \tag{29}$$

which is zero if $j \notin \overline{\text{pa}}(i)$. The following lemma formalizes this property for all relevant cases.

Lemma 7 (Score Changes under Interventions) *Consider the observational environment $\mathcal{E}^0$ and an interventional environment $\mathcal{E}^m$ with unknown intervention target I^m. Then, for any causal model and intervention type, $\mathbb{E}\left[\left| s(\mathbf{Z}) - s^m(\mathbf{Z}) \right|_i \right] \neq 0$ implies that $i \in \overline{\text{pa}}(I^m)$. For the reverse direction, we have the following results specified for each relevant case.*

(i) **Hard interventions:** *If the intervention in $\mathcal{E}^m$ (or $\tilde{\mathcal{E}}^m$) is hard, then score functions s and s^m (or $\tilde{s}^m$) differ in their i-th coordinate if and only if node i or one of its children is intervened in $\mathcal{E}^m$ (or in $\tilde{\mathcal{E}}^m$).*

$$\mathbb{E}\left[\left| s(\mathbf{Z}) - s^m(\mathbf{Z}) \right|_i \right] \neq 0 \quad \Longleftrightarrow \quad i \in \overline{\text{pa}}(I^m) , \tag{30}$$

$$\text{and} \quad \mathbb{E}\left[\left| s(\mathbf{Z}) - \tilde{s}^m(\mathbf{Z}) \right|_i \right] \neq 0 \quad \Longleftrightarrow \quad i \in \overline{\text{pa}}(\tilde{I}^m) . \tag{31}$$

(ii) **Soft interventions:** *If the intervention in $\mathcal{E}^m$ is soft and the latent causal model is an additive noise model, then score functions s and s^m differ in their i-th coordinate if and only if node i or one of its children is intervened in $\mathcal{E}^m$.*

$$\mathbb{E}\left[\left| s(\mathbf{Z}) - s^m(\mathbf{Z}) \right|_i \right] \neq 0 \quad \Longleftrightarrow \quad i \in \overline{\text{pa}}(I^m) . \tag{32}$$

16

(iii) **Coupled environments** $I^m = \tilde{I}^m$: *In the coupled environment setting, $\boldsymbol{s}^m$ and $\tilde{\boldsymbol{s}}^m$ differ in their i-th coordinate if and only if i is intervened.*

$$\mathbb{E}\left[\left|\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\right|_i\right] \neq 0 \quad \Longleftrightarrow \quad i = I^m \ . \tag{33}$$

(iv) **Uncoupled environments** $I^m \neq \tilde{I}^m$: *Consider two interventional environments $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ with different intervention targets $I^m \neq \tilde{I}^m$, and consider additive noise models specified in (7). Given that $p(\mathbf{Z})$ is twice differentiable, the score functions $\boldsymbol{s}^m$ and $\tilde{\boldsymbol{s}}^m$ differ in their i-th coordinate if and only if node i or one of its children is intervened.*

$$\mathbb{E}\left[\left|\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\right|_i\right] \neq 0 \quad \Longleftrightarrow \quad i \in \overline{\mathrm{pa}}(I^m, \tilde{I}^m) \ . \tag{34}$$

Proof: See Appendix A.1.

Theorem 7 provides the necessary and sufficient conditions for the invariance of the coordinates of the score functions of latent variables. The core idea of the score-based framework is that tracing these sparse changes in the score functions of the latent variables guides finding reliable estimates for the inverse of the transformation g, which in turn facilitates estimating $\mathbf{Z}$. Intuitively, we will look for the encoders $h \in \mathcal{H}$ such that the variations between the score estimates $\boldsymbol{s}_{\hat{\mathbf{Z}}}(\hat{\mathbf{z}}; h)$, $\boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h)$, and $\tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h)$ will be similar to the true score variations given by Theorem 7. However, the scores of the latent variables are not directly accessible. To circumvent this, we need to understand the connection between score functions of $\mathbf{X}$ and $\mathbf{Z}$. In the following lemma, we leverage the change of variables formula for injective mappings and establish this relationship for *any* injective mapping f from latent to observed space.

Lemma 8 (Score Difference Transformation) *Consider random vectors $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}^r$ and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^s$ that are related through*

$$\mathbf{Y}_1 = f(\mathbf{W}_1) \ , \quad and \quad \mathbf{Y}_2 = f(\mathbf{W}_2) \ , \tag{35}$$

such that $r \geq s$, probability measures of $\mathbf{W}_1, \mathbf{W}_2$ are absolutely continuous with respect to the $\boldsymbol{s}$-dimensional Lebesgue measure, and $f : \mathbb{R}^s \to \mathbb{R}^r$ is an injective and continuously differentiable function. The difference of the score functions of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, and that of $\mathbf{W}_1$ and $\mathbf{W}_2$ are related as

$$s_{\mathbf{W}_1}(\mathbf{w}) - s_{\mathbf{W}_2}(\mathbf{w}) = \left[J_f(\mathbf{w})\right]^\top \cdot \left[s_{\mathbf{Y}_1}(\mathbf{y}) - s_{\mathbf{Y}_2}(\mathbf{y})\right] \ , \quad where \ \mathbf{y} = f(\mathbf{w}) \ , \tag{36}$$

where $J_f(\mathbf{w})$ denotes the Jacobian of f at point w. Furthermore, for the reverse direction, we have

$$s_{\mathbf{Y}_1}(\mathbf{y}) - s_{\mathbf{Y}_2}(\mathbf{y}) = \left[\left[J_f(\mathbf{w})\right]^\dagger\right]^\top \cdot \left[s_{\mathbf{W}_1}(\mathbf{w}) - s_{\mathbf{W}_2}(\mathbf{w})\right] \ , \quad where \ \mathbf{y} = f(\mathbf{w}) \ . \tag{37}$$

Proof: See Appendix A.2.

For CRL under linear transformations in Section 5, we use the following corollary of Theorem 8.

Corollary 9 *In Theorem 8, if f is a linear transform, that is $\mathbf{Y} = \mathbf{F} \cdot \mathbf{W}$ for a full-rank matrix $\mathbf{F} \in \mathbb{R}^{r \times s}$, then the score functions of $\mathbf{Y}$ and $\mathbf{W}$ are related through $\boldsymbol{s}_{\mathbf{W}}(\mathbf{w}) = \mathbf{F}^\top \cdot \boldsymbol{s}_{\mathbf{Y}}(\mathbf{y})$ and $\boldsymbol{s}_{\mathbf{Y}}(\mathbf{y}) = \left[\mathbf{F}^\dagger\right]^\top \cdot \boldsymbol{s}_{\mathbf{W}}(\mathbf{w})$, where $\mathbf{y} = \mathbf{F} \cdot \mathbf{w}$.*

For CRL under general transformations in Section 6, we customize Theorem 8 as follows. Consider a candidate encoder $h \in \mathcal{H}$. Recall that $\hat{\mathbf{Z}}(\mathbf{X}; h) = h(\mathbf{X}) = (h \circ g)(\mathbf{Z})$. Then, by letting $f = (h \circ g)$, Theorem 8 gives

$$\text{between } \mathcal{E}^0 \text{ and } \mathcal{E}^m : \quad s_{\hat{\mathbf{Z}}}(\hat{\mathbf{z}}; h) - s_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h) = J_f^{-\top}(\mathbf{z}) \cdot \left[s(\mathbf{z}) - s^m(\mathbf{z}) \right] , \tag{38}$$

$$\text{between } \mathcal{E}^0 \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{\mathbf{Z}}}(\hat{\mathbf{z}}; h) - \tilde{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h) = J_f^{-\top}(\mathbf{z}) \cdot \left[s(\mathbf{z}) - \tilde{s}^m(\mathbf{z}) \right] , \tag{39}$$

$$\text{between } \mathcal{E}^m \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h) - \tilde{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h) = J_f^{-\top}(\mathbf{z}) \cdot \left[s^m(\mathbf{z}) - \tilde{s}^m(\mathbf{z}) \right] . \tag{40}$$

5. CRL under Linear Transformations

In this section, we consider CRL under linear transformation, in which the general transformation model in (3) becomes:

$$\mathbf{X} = \mathbf{G} \cdot \mathbf{Z} , \tag{41}$$

where $\mathbf{G} \in \mathbb{R}^{d \times n}$ is an *unknown* full-rank matrix mapping the latent variables to the observed ones. We will present steps for leveraging the properties of the score functions presented in Section 4 to design an algorithm that identifies the true encoder and recovers the true causal representations. The algorithm is referred to as **L**inear **S**core-based **C**ausal **L**atent **E**stimation via **I**nterventions (LSCALE-I). The theoretical guarantees associated with the algorithm steps will also be presented, serving as constructive proof of identifiability.

LSCALE-I consists of three stages outlined in Algorithm 1. In the first stage, we will learn an encoder $\hat{\mathbf{H}}$, row by row, where each row is an estimate of a row of the true encoder $\mathbf{G}^{\dagger}$. This encoder estimate will recover each Z_i up to mixing with $\mathbf{Z}_{\text{pa}(i)}$. In the second stage, we will use the estimated latent variables $\hat{\mathbf{H}}$ to obtain the transitive closure of the latent graph $\mathcal{G}$. Finally, as an optional third stage for *hard* interventions, we will leverage the independence statements implied by the hard interventions to refine our encoder estimate. We will demonstrate that this refinement yields scaling consistency and perfect DAG recovery. We will also discuss the relevance and distinctions of our results vis-à-vis the results in the existing literature.

Statistical diversity. Similarly to the existing identifiability results, it is necessary to have some regularity conditions [5] on the probability distributions of observational and interventional environments. We adopt the following assumption for the case of linear transformations.

Assumption 1 *For every possible pair (i, k) where $i \in [n], k \in \text{pa}(i)$, the following term cannot be a constant function in $\mathbf{z}$,*

$$\frac{\partial}{\partial z_k} \left(\log \frac{p_i(z_i \mid \mathbf{z}_{\text{pa}(i)})}{q_i(z_i \mid \mathbf{z}_{\text{pa}(i)})} \right) \left[\frac{\partial}{\partial z_i} \log \frac{p_i(z_i \mid \mathbf{z}_{\text{pa}(i)})}{q_i(z_i \mid \mathbf{z}_{\text{pa}(i)})} \right]^{-1} . \tag{42}$$

Essentially, $\frac{\partial}{\partial z_k} \left(\log \frac{p_i(z_i | \mathbf{z}_{\text{pa}(i)})}{q_i(z_i | \mathbf{z}_{\text{pa}(i)})} \right)$ captures the effect of intervening on node i on the *score* associated with node k. Therefore, Assumption 1 ensures that an intervention sufficiently differentiates the target variable and its parents in the score function. We note that this

---

5. Some examples include *generic interventions* in (Squires et al., 2023), *no pure shift interventions* condition in (Buchholz et al., 2023), and the *genericity condition* in (von Kügelgen et al., 2023).

is a very mild assumption and holds for a wide range of commonly used models (including additive noise models under hard interventions) and is discussed in more detail in Section 5.5.

Before we explain the details of LSCALE-I and establish identifiability results in the following subsections, we first present the rationale and derive the partial identifiability result, which will serve as the building block. In the rest of the paper, for brevity, we denote the score difference functions for each $m \in [n]$ by

$$d_{\mathbf{Z}}^m(\mathbf{z}) \triangleq s^m(\mathbf{z}) - s(\mathbf{z}) \quad \text{and} \quad d_{\mathbf{X}}^m(\mathbf{x}) \triangleq s_{\mathbf{X}}^m(\mathbf{x}) - s_{\mathbf{X}}(\mathbf{x}) . \tag{43}$$

5.1 Rationale for LSCALE-I and Partial Identifiability

To construct an algorithm that achieves identifiability under linear transformations, we first analyze the score differences for the linear transformation. Note that for linear transformation $\mathbf{X} = \mathbf{G} \cdot \mathbf{Z}$, Jacobian $J_{\mathbf{G}}(\mathbf{z})$ is independent of $\mathbf{z}$ and equal to matrix $\mathbf{G}$. Also, we have $\mathbf{z} = \mathbf{G}^\dagger \cdot \mathbf{x}$ for all $\mathbf{x} \in \text{col}(\mathbf{G})$. Then, using Theorem 9 we have

$$d_{\mathbf{X}}^m(\mathbf{x}) = [\mathbf{G}^\dagger]^\top \cdot d_{\mathbf{Z}}^m(\mathbf{z}) . \tag{44}$$

Let us denote the correlation matrices of $d_{\mathbf{X}}^m(\mathbf{x})$ and $d_{\mathbf{Z}}^m(\mathbf{z})$, respectively, by

$$\mathbf{R}_{\mathbf{X}}^m \triangleq \mathbb{E}\big[d_{\mathbf{X}}^m(\mathbf{x}) \cdot [d_{\mathbf{X}}^m(\mathbf{x})]^\top\big] \quad \text{and} \quad \mathbf{R}_{\mathbf{Z}}^m \triangleq \mathbb{E}\big[d_{\mathbf{Z}}^m(\mathbf{z}) \cdot [d_{\mathbf{Z}}^m(\mathbf{z})]^\top\big] . \tag{45}$$

Note that (44) implies that $\mathbf{R}_{\mathbf{X}}^m = [\mathbf{G}^\dagger]^\top \cdot \mathbf{R}_{\mathbf{Z}}^m \cdot \mathbf{G}^\dagger$. The next lemma specifies that the structure of the column space of correlation matrix $\mathbf{R}_{\mathbf{X}}^m$ is heavily constrained by the true encoder $\mathbf{G}^\dagger$, the graph $\mathcal{G}$, and the intervention target I^m in m-th environment.

Lemma 10 *For any interventional environment $\mathcal{E}^m$, $\text{col}(\mathbf{R}_{\mathbf{X}}^m) \subseteq \text{span}\big\{[\mathbf{G}_i^\dagger]^\top : i \in \overline{\text{pa}}(I^m)\big\}$.*

Proof: Using the sparse score changes property in (29), we know that $\big[d_{\mathbf{Z}}^m(\mathbf{z})\big]_i = 0$ for all $i \notin \overline{\text{pa}}(I^m)$. Then, using (44), for any $\mathbf{x} = \mathbf{G} \cdot \mathbf{z}$ we have

$$d_{\mathbf{X}}^m(\mathbf{x}) = \sum_{i \in \overline{\text{pa}}(I^m)} [\mathbf{G}_i^\dagger]^\top \cdot \big[d_{\mathbf{Z}}^m(\mathbf{z})\big]_i \in \text{span}\big\{ [\mathbf{G}_i^\dagger]^\top : i \in \overline{\text{pa}}(I^m) \big\} . \tag{46}$$

Since this holds for any $\mathbf{x} \in \text{col}(\mathbf{G})$, we have

$$\text{span}\big\{ d_{\mathbf{X}}^m(\mathbf{x}) : \mathbf{x} \in \text{col}(\mathbf{G}) \big\} \subseteq \text{span}\big\{ [\mathbf{G}_i^\dagger]^\top : i \in \overline{\text{pa}}(I^m) \big\} . \tag{47}$$

By definition of $\mathbf{R}_{\mathbf{X}}^m$, we have $\text{col}(\mathbf{R}_{\mathbf{X}}^m) = \text{span}\{d_{\mathbf{X}}^m(\mathbf{x}) : \mathbf{x} \in \text{col}(\mathbf{G})\}$ which concludes the proof. $\qquad\square$

Theorem 10 crucially implies that we can achieve partial identifiability of a single latent variable given that we have an environment in which the said variable is intervened.

Theorem 11 (Linear − Node-level partial identifiability) *A single-node soft interventional environment $\mathcal{E}^m$ with intervention target $\ell = I^m$ is sufficient to recover Z_ℓ up to mixing with its parents.*

Proof: Let us pick $\mathbf{y}$ uniformly at random from unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and let $\mathbf{h} = \mathbf{R}_{\mathbf{X}}^m \cdot \mathbf{y}$. Note that $\left[d_{\mathbf{Z}}^m(\mathbf{z}) \right]_i = 0$ for all $i \notin \overline{\text{pa}}(\ell)$ implies that $\left[\mathbf{R}_{\mathbf{Z}}^m \right]_i = \mathbf{0}$ for all $i \notin \overline{\text{pa}}(\ell)$. Then, using $\mathbf{R}_{\mathbf{X}}^m = [\mathbf{G}^\dagger]^\top \cdot \mathbf{R}_{\mathbf{Z}}^m \cdot \mathbf{G}^\dagger$, we have

$$\mathbf{h} = \mathbf{R}_{\mathbf{X}}^m \cdot \mathbf{y} = [\mathbf{G}^\dagger]^\top \cdot \mathbf{R}_{\mathbf{Z}}^m \cdot \mathbf{G}^\dagger \cdot \mathbf{y} \tag{48}$$

$$= \sum_{i \in \overline{\text{pa}}(\ell)} \left[\mathbf{R}_{\mathbf{Z}}^m \right]_i \cdot \mathbf{G}^\dagger \cdot \mathbf{y} \cdot [\mathbf{G}_i^\dagger]^\top \tag{49}$$

Denote $c_i = \left[\mathbf{R}_{\mathbf{Z}}^m \right]_i \cdot \mathbf{G}^\dagger \cdot \mathbf{y}$ for each $i \in \overline{\text{pa}}(\ell)$. Then, using $\mathbf{G}_i^\dagger \cdot \mathbf{G} = \mathbf{e}_i$, i-th standard basis vector, we have $[\mathbf{h}^\top \cdot \mathbf{G}]_i = c_i$ for $i \in \overline{\text{pa}}(\ell)$. As such, we can use $\mathbf{h}$ to obtain

$$\mathbf{h}^\top \cdot \mathbf{X} = \mathbf{h}^\top \cdot \mathbf{G} \cdot \mathbf{Z} = c_\ell \cdot Z_\ell + \sum_{i \in \text{pa}(\ell)} c_i \cdot Z_i \ . \tag{50}$$

Finally, note that $\left[\mathbf{R}_{\mathbf{Z}}^m \right]_\ell$ is not a zero vector since $[d_{\mathbf{Z}}^m]_\ell \neq 0$. Therefore, since rows of $\mathbf{G}^\dagger$ are linearly independent, we know that $\left[\mathbf{R}_{\mathbf{Z}}^m \right]_i \cdot \mathbf{G}^\dagger$ is a nonzero vector. Thus, $\mathbf{y} \in \mathbb{S}^{d-1}$ ensures that c_ℓ is nonzero with probability 1. Then, $\mathbf{h}^\top \cdot \mathbf{X}$ is an estimate of Z_ℓ that satisfies consistency up to mixing with parents. $\qquad\square$

5.2 Identifiability via Soft Interventions

After obtaining the partial identifiability result from a single interventional environment, our goal is to construct an algorithm that uses the complete set of interventional environments $\mathcal{E} = \{\mathcal{E}^1, \ldots, \mathcal{E}^n\}$ with targets $\cup_{m \in [n]} I^m = [n]$ to form estimates of the true encoder $\mathbf{G}^\dagger$ and the latent graph $\mathcal{G}$.

Encoder estimation. In Stage L1 of Algorithm 1, we form an encoder estimate $\hat{\mathbf{H}}$ via using the column spaces of $\{\mathbf{R}_{\mathbf{X}}^m : m \in [n]\}$. Specifically, for each $m \in [n]$ we select a vector randomly from $\text{col}(\mathbf{R}_{\mathbf{X}}^m)$ and assign its transpose to m-th row of $\hat{\mathbf{H}}$.

Lemma 12 (Linear − Encoder via Soft Interventions) *Under soft interventions, output $\hat{\mathbf{H}}$ of Algorithm 1 achieves identifiability up to mixing with parents. Specifically, $\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}}) = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\text{pa}} \cdot \mathbf{Z}$ such that diagonal entries of $\mathbf{C}_{\text{pa}}$ are nonzero and $\left[\mathbf{C}_{\text{pa}} \right]_{i,j} = 0$ for all $j \notin \overline{\text{pa}}(i)$.*

Proof: Following Theorem 11 for all $m \in [n]$ immediately implies the desired result. Specifically, according to (50), $[\hat{\mathbf{H}}_m]^\top \in \text{col}(\mathbf{R}_{\mathbf{X}}^m)$ satisfies

$$[\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}})]_m = \hat{\mathbf{H}}_m \cdot \mathbf{X} = c_\ell \cdot Z_\ell + \sum_{i \in \text{pa}(\ell)} c_i \cdot Z_i \ , \qquad \text{where } \ell = I^m \ , \tag{51}$$

for some constants $\{c_i : i \in \overline{\text{pa}}(\ell)\}$ where $c_\ell \neq 0$ with probability 1. Then, combining all n identities, we have

$$\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}}) = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\text{pa}} \cdot \mathbf{Z} \ , \tag{52}$$

where $\mathbf{P}_{\mathcal{I}}$ is the permutation matrix of the intervention order $(I^1, \ldots, I^n)$, and $\mathbf{C}_{\text{pa}}$ has nonzero diagonal entries and satisfies $\left[\mathbf{C}_{\text{pa}} \right]_{i,j} = 0$ for all $j \notin \overline{\text{pa}}(i)$. $\qquad\square$

Algorithm 1 **L**inear **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (LSCALE-I)

1: **Input:** $\mathbf{R}_{\mathbf{X}}^m$ for all $m \in [n]$ ▷ compute using (45)

2: **Stage L1: Encoder estimation**
3: $\hat{\mathbf{H}} \leftarrow \mathbf{0}_{n \times d}$
4: **for** $m \in (1, \ldots, n)$ **do**
5:     Select any $\mathbf{h} \in \mathrm{col}(\mathbf{R}_{\mathbf{X}}^m)$
6:     $\hat{\mathbf{H}}_m \leftarrow \mathbf{h}^\top$

7: **Stage L2: Latent graph estimation**
8: Construct latent DAG estimate $\hat{\mathcal{G}}$ with parent sets

$$\hat{\mathrm{pa}}(m) \leftarrow \left\{ i \neq m : \mathbb{E}\left[\left| \boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) \big|_i \right| \right] \neq 0 \right\}, \quad \forall m \in [n]$$

9: **if** the interventions are hard **then**
10:     **Stage L3: Unmixing procedure**
11:     $\pi \leftarrow$ a topological order of $\hat{\mathcal{G}}$
12:     **for** $m \in (\pi_1, \ldots, \pi_n)$ **do** ▷ refine rows of $\hat{\mathbf{H}}$ sequentially
13:         $\hat{\mathbf{Z}} \leftarrow \hat{\mathbf{Z}}^m(\mathbf{X}; \hat{\mathbf{H}})$
14:         $\mathbf{u} \leftarrow \mathrm{Cov}(\hat{Z}_m, \hat{\mathbf{Z}}_{\hat{\mathrm{pa}}(m)}) \cdot \left[\mathrm{Cov}(\hat{\mathbf{Z}}_{\hat{\mathrm{pa}}(m)}) \right]^{-1}$
15:         $\hat{\mathbf{H}}_m \leftarrow \hat{\mathbf{H}}_m - \mathbf{u} \cdot \hat{\mathbf{H}}_{\hat{\mathrm{pa}}(m)}$
16:     Refine $\hat{\mathcal{G}}$ according to

$$\hat{\mathrm{pa}}(m) \leftarrow \left\{ i \neq m : \mathbb{E}\left[\left| \boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) \big|_i \right| \right] \neq 0 \right\}, \quad \forall m \in [n]$$

17: **Return** $\hat{\mathcal{G}}$, $\hat{\mathbf{H}}$, and $\hat{\mathbf{Z}}$

Latent graph estimation. Next, we use the encoder estimate $\hat{\mathbf{H}}$ to recover the transitive closure of the latent graph $\mathcal{G}$. Motivated by Theorem 7(ii), we form $\hat{\mathcal{G}}$ with parent sets

$$\hat{\mathrm{pa}}(m) \triangleq \left\{ i \neq m : \mathbb{E}\left[\left| \boldsymbol{s}_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) - \boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) \big|_i \right| \right] \neq 0 \right\}. \tag{53}$$

Using the result that $\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}})$ recovers $\mathbf{Z}$ up to mixing with parents, we show that the transitive closures of the estimate $\hat{\mathcal{G}}$ and the true graph $\mathcal{G}$ are the same under a graph isomorphism.

Lemma 13 (Linear – Graph via Soft Interventions) *Under Assumption 1 and soft interventions, output $\hat{\mathcal{G}}$ of Algorithm 1 recovers the transitive closure of $\mathcal{G}$.*

Proof: Let $\hat{\mathbf{Z}}$ denote $\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}}) = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\mathrm{pa}} \cdot \mathbf{Z}$. Recall that $\left[\mathbf{C}_{\mathrm{pa}} \right]_{i,j} = 0$ for all $j \notin \overline{\mathrm{pa}}(i)$, and since nodes $(1, \ldots, n)$ are topologically ordered, $\mathbf{C}_{\mathrm{pa}}$ is a lower triangular matrix. By using

standard linear algebra for the expansion of an inverse of a lower triangular matrix, we can show that $\left[\mathbf{C}_{\mathrm{pa}}^{-1}\right]_{i,j} = 0$ for all $j \notin \overline{\mathrm{an}}(i)$ (see Theorem 33 in Appendix B.1 for the detailed steps). Then, using Theorem 9 for the scores of $\mathbf{Z}$, we have

$$\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}) = \mathbf{P}_{\mathcal{I}} \cdot [\mathbf{C}_{\mathrm{pa}}^{-1}]^\top \cdot \boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{Z}) \ . \tag{54}$$

Next, taking the transpose of $\mathbf{C}_{\mathrm{pa}}^{-1}$, we have $\left[\mathbf{C}_{\mathrm{pa}}^{-\top}\right]_{i,j} = 0$ for all $j \notin \overline{\mathrm{de}}(i)$. Combining this observation with the fact that $[\boldsymbol{d}_{\mathbf{Z}}^m]_j = 0$ for all $j \notin \overline{\mathrm{pa}}(I^m)$, we obtain

$$\left[\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}})\right]_k = \left[\mathbf{C}_{\mathrm{pa}}^{-\top}\right]_{I^k} \cdot \boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{Z}) = \sum_{j \in \overline{\mathrm{de}}(I^k) \cap \overline{\mathrm{pa}}(I^m)} \left[\mathbf{C}_{\mathrm{pa}}^{-\top}\right]_{I^k,j} \cdot [\boldsymbol{d}_{\mathbf{Z}}^m]_j \tag{55}$$

Subsequently, $\mathbb{E}\left[|\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}})|_k\right] \neq 0$ implies that $\overline{\mathrm{de}}(I^i) \cap \overline{\mathrm{pa}}(I^m) \neq \emptyset$, and $I^k \notin \overline{\mathrm{an}}(I^m)$. Therefore, the estimated parent set in (53) satisfies $\hat{\mathrm{pa}}(m) \subseteq \overline{\mathrm{an}}(I^m)$, and the transitive closure of $\mathcal{G}$ is a supergraph of $\hat{\mathcal{G}}$ under relabeling of the nodes with permutation $\mathcal{I}$. Next, we show that $\hat{\mathcal{G}}$ contains all the edges in the transitive reduction of $\mathcal{G}$. Let $I^k \to I^m$ be an edge in the transitive reduction of $\mathcal{G}$, i.e., there is no other directed path between nodes I^k and I^m in $\mathcal{G}$. This implies that $\overline{\mathrm{de}}(I^k) \cap \overline{\mathrm{pa}}(I^m) = \{I^k, I^m\}$, and (54) becomes

$$\left[\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}})\right]_k = \left[\mathbf{C}_{\mathrm{pa}}^{-\top}\right]_{I^k,I^k} \cdot [\boldsymbol{d}_{\mathbf{Z}}^m]_{I^k} + \left[\mathbf{C}_{\mathrm{pa}}^{-\top}\right]_{I^k,I^m} \cdot [\boldsymbol{d}_{\mathbf{Z}}^m]_{I^m} \ . \tag{56}$$

Assumption 1 ensures that $\left[\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}})\right]_k$ is not constantly zero, which means that $\hat{\mathcal{G}}$ contains $k \to m$ edge, preserving $I^k \to I^m$ in $\mathcal{G}$. Therefore, $\hat{\mathcal{G}}$ preserves all the edges in the transitive reduction of $\mathcal{G}$, and taking its transitive closure gives $\mathcal{G}_{\mathrm{tc}}$, under relabeling of the nodes with permutation $\mathcal{I}$. □

Theorem 14 (Linear − Soft Interventions for General SCMs) *Under Assumption 1 for linear transformations, using observational data and interventional data from one* soft *intervention per node suffice to identify (i) the transitive closure of the latent DAG $\mathcal{G}$ and (ii) the latent variables $\mathbf{Z}$ with consistency up to mixing with parents. Specifically, Algorithm 1 achieves these identifiability guarantees.*

Proof: Combining Theorem 12 and Theorem 13 gives the statements of the theorem. □

We note that the existing literature on CRL with linear transformations and one soft intervention requires the latent causal model to be either linear Gaussian (Squires et al., 2023; Buchholz et al., 2023) or satisfy nonlinearity conditions (Zhang et al., 2023)[6]. In contrast, Theorem 14 achieves identifiability guarantees for soft interventions without imposing any restrictions (distributional or structural) on the latent causal model. Furthermore, recovering $\mathbf{Z}$ up to mixing with *parents* improves upon the existing guarantees of recovering up to mixing with *ancestors* (Squires et al., 2023; Buchholz et al., 2023). Finally, we note that $\mathcal{G}$ cannot be identified beyond its transitive closure when using soft interventions without making additional assumptions (Squires et al., 2023, Appendix J). Therefore, the graph identifiability result in Theorem 14 is tight.

---

6. The "linear interventional faithfulness" (Zhang et al., 2023, Assumption 2) implies nonlinearity, which we elaborate in Appendix D.3.

5.3 Identifiability via Hard Interventions

Next, we investigate hard interventions (i.e., perfect interventions) for general latent causal models. Hard interventions are special cases of soft interventions, in which the intervened node becomes functionally independent of its parents. Consequently, the identifiability guarantees for hard interventions are usually stronger, as we show in this section. Our analysis builds on leveraging the following property, which is exclusive to hard interventions.

Proposition 15 *For the environment $\mathcal{E}^m$ in which node $\ell = I^m$ is hard intervened, we have*

$$Z_\ell^m \perp\!\!\!\perp Z_j^m , \quad \forall j \in \mathrm{nd}(\ell) , \tag{57}$$

where $\mathrm{nd}(\ell)$ is the set of non-descendants of ℓ in $\mathcal{G}$.

The Markov property readily implies this property, that is, each variable in a DAG is independent of its non-descendants, given its parents. When node ℓ is hard-intervened, it has no parents, and the statement follows directly. Motivated by this property, we want to ensure that the estimated latent variables conform to Theorem 15. To this end, we aim to update rows of the encoder estimate $\hat{\mathbf{H}}$ to remove the effects of $\mathbf{Z}_{\mathrm{pa}(i)}$ from the estimate of Z_i. We achieve this objective as follows.

Note that Theorem 15 provides us with random variable pairs that are supposed to be independent, and the covariance of two independent random variables is necessarily zero. In Stage L3 of Algorithm 1, for each row $\hat{\mathbf{H}}_m$, we consider $\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}})$ in the corresponding environment in which $\hat{Z}_m$ is intervened, and compute an unmixing vector $\mathbf{u}$ via a linear minimum mean square error (MMSE) estimator,

$$\mathbf{u} = \mathrm{Cov}(\hat{Z}_m, \hat{\mathbf{Z}}_{\hat{\mathrm{pa}}(m)}) \cdot \left[\mathrm{Cov}(\hat{\mathbf{Z}}_{\hat{\mathrm{pa}}(m)}) \right]^{-1} . \tag{58}$$

Then, we update the corresponding row of the encoder as $\hat{\mathbf{H}}_m \leftarrow \hat{\mathbf{H}}_m - \mathbf{u} \cdot \hat{\mathbf{H}}_{\hat{\mathrm{pa}}(m)}$. Finally, for the perfect DAG recovery, we compute the latent score differences under the final encoder estimate and reconstruct the graph $\hat{\mathcal{G}}$ with parent sets

$$\hat{\mathrm{pa}}(m) \triangleq \left\{ i \neq m : \mathbb{E}\left[\left| \boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) \right|_i \right] \neq 0 \right\} . \tag{59}$$

Theorem 16 (Linear – Hard Interventions) *Under Assumption 1 for linear transformations, using observational data and interventional data from one* hard *intervention per node suffices to identify (i) the latent DAG $\mathcal{G}$ perfectly and (ii) the latent variables $\mathbf{Z}$ with scaling consistency. Specifically, Algorithm 1 achieves these identifiability guarantees.*

Proof: See Appendix B.2.

Similar to the restrictions in the existing results for soft interventions, the identifiability results for hard interventions in the existing literature restrict the latent causal model, e.g., to linear causal models (Squires et al., 2023; Buchholz et al., 2023). In contrast, Theorem 16 does not impose any restriction on the latent causal model and shows that one stochastic hard intervention per node is sufficient for the identifiability of general latent causal models.

Finally, we comment on the differences between the identifiability results for soft and hard interventions under missing intervention targets and discuss the identifiability of latent subgraphs under special cases.

Remark 17 *Let $\mathcal{J} \subset [n]$ denote a non-complete set of intervention targets, i.e., $\mathcal{J} \neq [n]$. We note the immediate implications of our results as follows.*

1. ***Partial identifiability of latent variables.*** *Theorem 11 immediately implies that for all $i \in \mathcal{J}$ we can recover Z_i up to mixing with parents. However, the unmixing procedure for hard interventions resolves the mixing with parents sequentially. Specifically, to identify Z_i up to scaling, we use the fact that the parents of Z_i are already identified up to scaling. Therefore, even if we have a hard intervention on Z_i, we cannot identify Z_i up to scaling consistency when its parents are not identified up to scaling.*

2. ***Ancestrally closed set of interventions.*** *Let $\mathcal{G}_{\mathcal{J}}$ denote the induced subgraph of $\mathcal{G}$ over the set of nodes $\mathcal{J}$. If $\mathcal{J}$ is not ancestrally closed, i.e., $\mathrm{pa}(\mathcal{J}) \nsubseteq \mathcal{J}$, then the marginal distribution of $Z_{\mathcal{J}}$ is not a causally sufficient model. Conversely, if $\mathcal{J}$ is an ancestrally closed set, then the marginal of $Z_{\mathcal{J}}$ comes from a causally sufficient model with graph $\mathcal{G}_{\mathcal{J}}$. Subsequently, using Stage L2 of Algorithm 1 for score differences of $\hat{\mathbf{Z}}_{\mathcal{J}}$, we obtain the transitive closure of $\mathcal{G}_{\mathcal{J}}$. Finally, performing the unmixing procedure over the ancestrally closed set of* hard *interventions, Stage L3 of Algorithm 1 ensures the perfect recovery of $\mathcal{G}_{\mathcal{J}}$.*

5.4 Identifiability up to Surrounding Parents

For linear transformations, we finally investigate the conditions under which soft interventions are guaranteed to achieve identifiability results stronger than transitive closure and mixing up to parents. In particular, we specify one condition on the rank of the score function differences $\{\boldsymbol{d}_{\mathbf{Z}}^m : m \in [n]\}$, which is formalized next.

Assumption 2 (Full-rank Score Difference) *For all interventional environments $\mathcal{E}^m \in \mathcal{E}$ we have*

$$\mathrm{rank}(\mathbf{R}_{\mathbf{Z}}^m) = |\overline{\mathrm{pa}}(I^m)| \ . \tag{60}$$

For insight into this assumption, it can be readily verified that for linear Gaussian latent models $\mathrm{rank}(\mathbf{R}_{\mathbf{Z}}^m) \leq 2$ and on the other hand, for sufficiently nonlinear causal models, $\mathrm{rank}(\mathbf{R}_{\mathbf{Z}}^m)$ is $\overline{\mathrm{pa}}(I^m)$. This assumption is stronger than Assumption 1 since it implies that the effects of an intervention on *all* parents of the target variable are different. We will provide more discussions on this assumption in Section 5.5. Under this condition, Algorithm 2 identifies the latent graph and the latent variables with the following key intuitions.

Latent graph estimation. Recall that $\mathbf{R}_{\mathbf{X}}^m = [\mathbf{G}^\dagger]^\top \cdot \mathbf{R}_{\mathbf{Z}}^m \cdot \mathbf{G}^\dagger$. Since $\mathbf{G}^\dagger$ has full row-rank n, Assumption 2 implies that $\mathrm{rank}(\mathbf{R}_{\mathbf{X}}^m) = |\overline{\mathrm{pa}}(I^m)|$ for all $m \in [n]$, which further implies that Theorem 10 becomes an equality as well, i.e.,

$$\mathrm{col}(\mathbf{R}_{\mathbf{X}}^m) = \mathrm{span}\{[\mathbf{G}_i^\dagger]^\top : i \in \overline{\mathrm{pa}}(I^m)\} \ . \tag{61}$$

Then, for any $t, k \in [n]$, we have

$$\mathrm{col}(\mathbf{R}_{\mathbf{X}}^t) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k) = \mathrm{span}\{[\mathbf{G}_i^\dagger]^\top : i \in \overline{\mathrm{pa}}(I^t) \cap \overline{\mathrm{pa}}(I^k)\} \ , \tag{62}$$

$$\dim\left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^t) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k)\right) = |\overline{\mathrm{pa}}(I^t) \cap \overline{\mathrm{pa}}(I^k)| \tag{63}$$

$$= |\mathrm{pa}(I^t) \cap \mathrm{pa}(I^k)| + \mathbb{1}\{\{I^t \in \mathrm{pa}(I^k)\} \vee \{I^k \in \mathrm{pa}(I^t)\}\} \ . \tag{64}$$

Algorithm 2 LSCALE-I for sufficiently nonlinear latent causal models

---

1: **Input:** $\mathbf{R}_{\mathbf{X}}^m$ for all $m \in [n]$, topological order π from Algorithm 1
2: **Latent graph estimation**
3: Initialize $\hat{\mathcal{G}}$ with empty graph
4: **for** $k \in (\pi_1, \ldots, \pi_n)$ **do**
5:     **for** $t \in (\pi_1, \ldots, \pi_{k-1})$ **do**
6:         **if** $\dim \left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^t) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k) \right) > |\hat{\mathrm{pa}}(t) \cap \hat{\mathrm{pa}}(k)|$ **then**
7:             Update $\hat{\mathrm{pa}}(k) \leftarrow \hat{\mathrm{pa}}(k) \cup \{t\}$
8: **Encoder estimation**
9: Initialize $\hat{\mathbf{H}} \leftarrow \mathbf{0}_{n \times d}$
10: **for** $m \in (1, \ldots, n)$ **do**
11:     Select any $\mathbf{h} \in \bigcap\limits_{k \in \hat{\mathrm{ch}}(m) \cup m} \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k)$
12:     $\hat{\mathbf{H}}_m \leftarrow \mathbf{h}^\top$
13: **Return** $\hat{\mathcal{G}}$, $\hat{\mathbf{H}}$, and $\hat{\mathbf{Z}}$

In Algorithm 2, we first obtain a topological order of the true DAG using the graph estimate from Algorithm 1 and initialize the new estimate $\hat{\mathcal{G}}$ with the empty graph. Consider a pair (t, k) such that $I^t \in \mathrm{pa}(I^k)$. If $\hat{\mathrm{pa}}(k)$ does not contain node t in $\hat{\mathcal{G}}$, then by (64) we have $\dim \left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^t) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k) \right) > |\hat{\mathrm{pa}}(t) \cap \hat{\mathrm{pa}}(k)|$. We use this observation to sequentially identify the parents of each node, starting from the root node(s) and gradually advancing to the leaf node(s). In the end, resulting $\hat{\mathcal{G}}$ equals $\mathcal{G}$ under permutation $\mathcal{I}$.

Encoder estimation. In Algorithm 2, we refine the encoder estimation step of Algorithm 1 as follows. Extending (62) and using the graph recovery result, we have

$$\bigcap_{k \in \hat{\mathrm{ch}}(m) \cup m} \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k) = \mathrm{span}\left\{ [\mathbf{G}_i^\dagger]^\top : i \in \bigcap_{k \,:\, I^k \in \overline{\mathrm{ch}}(I^m)} \overline{\mathrm{pa}}(I^k) \right\} . \tag{65}$$

Now, consider $I^t \notin \mathrm{sur}(I^m)$. By the definition of surrounding parents, there exists $I^j \in \overline{\mathrm{ch}}(I^m)$ such that $I^j \notin \mathrm{ch}(I^t)$. Then, I^j is not in the intersection of the parent sets in the right-hand-side of (65), which means $[\mathbf{G}_j^\dagger]^\top$ is not in the column space of the left-hand-side of (65). Therefore, by choosing $\hat{\mathbf{H}}_m$ from this column space, we obtain for all $m \in [n]$,

$$[\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}})]_m = \hat{\mathbf{H}}_m \cdot \mathbf{X} = c_\ell \cdot Z_\ell + \sum_{i \in \mathrm{sur}(\ell)} c_i \cdot Z_i , \qquad \text{where } \ell = I^m , \tag{66}$$

$$\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}}) = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\mathrm{pa}} \cdot \mathbf{Z} , \tag{67}$$

where $\mathbf{P}_{\mathcal{I}}$ is the permutation matrix of the intervention order $(I^1, \ldots, I^n)$, $\mathbf{C}_{\mathrm{sur}}$ has nonzero diagonal entries and satisfies $\left[\mathbf{C}_{\mathrm{sur}} \right]_{i,j} = 0$ for all $j \notin \overline{\mathrm{sur}}(i)$. $\square$

Theorem 18 (Linear – Soft Interventions for Nonlinear SCMs) *Under Assumption 2 for linear transformations, using observational data and interventional data from one soft intervention per node suffice to identify (i) the latent DAG $\mathcal{G}$ perfectly, and (ii) the latent variables $\mathbf{Z}$ with consistency up to mixing with surrounding parents, and recovered latent*

variables satisfy Markov property with respect to $\mathcal{G}$. Specifically, Algorithm 2 achieves these identifiability guarantees.

Proof: Proof of the consistency of latent variables up to mixing with surrounding parents is given above in complete detail. For the proof of perfect graph recovery and $\hat{\mathbf{Z}}$ satisfies Markov property with respect to $\hat{\mathcal{G}}$, see Appendix B.3. □

Theorem 18 has two important implications. First, the latent DAG can be identified using *only* soft interventions under mild nonlinearity assumptions on the latent causal model. To our knowledge, this is the first result in the literature for fully recovering latent DAG with soft interventions without restricting the graphical structure, e.g., Zhang et al. (2023) require *linear faithfulness* assumption to achieve similar results, which is only shown to hold for nonlinear latent models with polytree structure. Secondly, the estimated latent variables reveal the true conditional independence relationships since they satisfy the Markov property with respect to the estimated latent graph, which is isomorphic to the latent graph. Recalling that the motivation of CRL is to learn useful representations that preserve causal relationships, our result shows that this can be achieved without perfect identifiability for a large class of models. Furthermore, Jin and Syrgkanis (2024, Theorem 3) establish that under some nondegeneracy assumptions, mixing consistency up to surrounding parents is the best possible result when using single-node soft interventions. Hence, the results in Theorem 18 are tight for the considered setting.

5.5 Discussion on Assumptions 1 and 2

In this subsection, we elaborate on Assumption 1 and Assumption 2, which are relevant to Theorem 14 and Theorem 18, respectively. Assumption 1 essentially states that score changes in the coordinates of the intervened node and a parent of the intervened node are linearly independent. This property holds for (but is not limited to) the widely adopted additive noise models specified in (7) when we apply hard interventions.

Lemma 19 (Proof in Appendix D.1) *Assumption 1 is satisfied for additive noise models under hard interventions.*

When soft interventions are applied, the transitive closure of $\mathcal{G}$ cannot be identified without making assumptions about the effect of the interventions. Specifically, for linear latent causal models, Buchholz et al. (2023) prove impossibility results for pure shift interventions and Squires et al. (2023) show that a *genericity condition* is necessary for identifying the transitive closure of the latent DAG. Therefore, Assumption 1 can be interpreted as the counterpart of the commonly adopted assumptions in the literature on soft interventions adapted to the setting of general latent causal models. Finally, the next example demonstrates the working of Assumption 1 on a linear Gaussian latent model.

Example 1 *Consider a linear Gaussian latent model with $\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$. Score function of $\mathbf{Z}$ is given by $\boldsymbol{s}(\mathbf{z}) = -\Sigma^{-1} \cdot \mathbf{z}$. Let $Z_i = \mathbf{w} \cdot \mathbf{Z}_{\mathrm{pa}(i)} + N_i$ where $N_i \sim \mathcal{N}(0, \sigma_i^2)$ for the node i. Consider an intervention on node i on environment $\mathcal{E}^m$ such that $Z_i^m = \bar{\mathbf{w}} \cdot \mathbf{Z}_{\mathrm{pa}(i)}^m + \bar{N}_i$, where $Z_m \sim \mathcal{N}(0, \bar{\Sigma})$, and $\bar{N}_i \sim \mathcal{N}(0, \bar{\sigma}_i^2)$, which yields $\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z}) = -(\Sigma^{-1} - \bar{\Sigma}^{-1}) \cdot \mathbf{z}$. Then, for*

a node $k \in \mathrm{pa}(i)$, we obtain

$$[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_i = \left(\frac{1}{\sigma_i^2} - \frac{1}{\bar{\sigma}_i^2}\right) z_i - \left(\frac{\mathbf{w}}{\sigma_i^2} - \frac{\bar{\mathbf{w}}}{\bar{\sigma}_i^2}\right) \mathbf{z}_{\mathrm{pa}(i)} , \tag{68}$$

$$[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_k = -\left(\frac{w_k}{\sigma_i^2} - \frac{\bar{w}_k}{\bar{\sigma}_i^2}\right) z_i + \left(w_k \frac{\mathbf{w}}{\sigma_i^2} - \bar{w}_k \frac{\bar{\mathbf{w}}}{\bar{\sigma}_i^2}\right) \mathbf{z}_{\mathrm{pa}(i)} , \tag{69}$$

where w_k and $\bar{w}_k$ correspond to weights of the parent Z_k in observational and interventional models, respectively. Note that, for Assumption 1 to be violated, there must exists a constant $\kappa \in \mathbb{R}$ such that

$$\kappa \cdot \left[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})\right]_i = \left[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})\right]_k , \quad \forall \mathbf{z} \in \mathbb{R}^n . \tag{70}$$

However, using (68) and (69), this is possible if and only if $w_k = \bar{w}_k$. Therefore, if the weight of the node $k \in \mathrm{pa}(i)$ changes, Assumption 1 is satisfied for the node pair (i, k).

Given the known result that perfect identifiability is impossible for linear Gaussian models given soft interventions (Squires et al., 2023), the purpose of Assumption 2 is to get more insight into the extent of identifiability guarantees under soft interventions. Intuitively, the mentioned impossibility results for linear Gaussian models are due to the rank deficiency of score differences for linear models – specifically, we know that $\mathrm{rank}(\mathbf{R}_{\mathbf{Z}}^m) \leq 2$ for linear Gaussian models. In contrast, for sufficiently nonlinear causal models, $\mathrm{rank}(\mathbf{R}_{\mathbf{Z}}^m)$ can be as high as $\overline{\mathrm{pa}}(I^m)$. Assumption 2 ensures that this upper bound is satisfied with equality for all nodes. This condition holds for the class of sufficiently nonlinear models, such as quadratic causal models. In particular, we show that this condition holds for the two-layer neural networks (NNs) as a function class that can effectively approximate any continuous function. This result is formalized in the next lemma.

Lemma 20 *Consider the additive model in (7) where f_i is a two-layer NN with sigmoid activation function, and weight matrices $\mathbf{W}^i$ and $\bar{\mathbf{W}}^i$ for observational and interventional mechanisms, respectively. If $\max\{\mathrm{rank}(\mathbf{W}^i), \mathrm{rank}(\bar{\mathbf{W}}^i)\} = |\mathrm{pa}(i)|$ for all $i \in [n]$, then Assumption 2 holds.*

We discuss the nonlinearity and the proof of Theorem 20 in Appendix D.2.

6. CRL under General Transformations

In this setting, we consider general transformations without any parametric assumption for the transformation g. In the previous section, we exploited the transformation's linearity and recovered the true encoder's parameters sequentially using *one* intervention per node. For general transformations (parametric or nonparametric), however, we cannot use the same parametric approach and rely on the properties of linear transforms. To rectify these and design the general CRL algorithm, we use more information in the form of *two* interventions per node. Specifically, we will present the steps of leveraging the score functions under two interventions and design an algorithm that identifies the true encoder and recovers the true causal representations. The algorithm is referred to as **G**eneral **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (GSCALE-I), which will be summarized in Algorithm 3. We also present additional results and discuss the role of various inputs and the distinction of our results compared to the existing literature.

Inputs. The inputs of GSCALE-I are the observed data from the observational environment, the data from two interventional environments per node, whether environments are coupled/uncoupled, and a set of valid encoders $\mathcal{H}$. Two sets of interventional environments are denoted by $\mathcal{E} = \{\mathcal{E}^1, \dots, \mathcal{E}^n\}$ and $\tilde{\mathcal{E}} = \{\tilde{\mathcal{E}}^1, \dots, \tilde{\mathcal{E}}^n\}$, as defined in Section 3.3. For these inputs, we compute the score functions $\boldsymbol{s}_{\mathbf{X}}$, $\{\boldsymbol{s}_{\mathbf{X}}^1, \dots, \boldsymbol{s}_{\mathbf{X}}^n\}$ and $\{\tilde{\boldsymbol{s}}_{\mathbf{X}}^1, \dots, \tilde{\boldsymbol{s}}_{\mathbf{X}}^n\}$. Note that we will use Theorem 8 again to ensure access to the latent score differences by using these observed score functions.

Statistical diversity. For being more informative than a single intervention mechanism, we assume that the two intervention mechanisms per node are sufficiently distinct. This is formalized by defining *interventional discrepancy* (Liang et al., 2023) among the causal mechanisms of a latent variable.

Definition 21 (Interventional Discrepancy) *Two intervention mechanisms with pdfs* $p, q : \mathbb{R} \to \mathbb{R}$ *are said to satisfy* interventional discrepancy *if*

$$\frac{\partial}{\partial u} \frac{p(u)}{q(u)} \neq 0 \ , \quad \forall u \in \mathbb{R} \setminus \mathcal{T} \ , \tag{71}$$

where $\mathcal{T}$ is a null set (i.e., has a zero Lebesgue measure).

This condition ensures that the two distributions are sufficiently distinct, formally expressed as the partial derivative of their ratio with respect to the intervened variable is nonzero almost everywhere. For instance, two univariate, distinct Gaussians trivially satisfy this discrepancy condition. As shown by Liang et al. (2023), even when the latent graph $\mathcal{G}$ is known, for identifiability via one intervention per node, it is necessary to have interventional discrepancy between observational distribution p_i and interventional distribution q_i, for all $\mathbf{z}_{\mathrm{pa}(i)} \in \mathbb{R}^{|\mathrm{pa}(i)|}$.

6.1 Rationale of GSCALE-I and Partial Identifiability

Similarly to LSCALE-I, analyzing GSCALE-I involves leveraging score differences. However, in contrast to linear transformations, estimated and true latent score differences are not always related by a constant matrix for general transformations. To circumvent this issue, we rely on Theorem 7(iii), i.e., the score difference of *coupled* interventional environments is one-sparse. As such, the key idea is to find an encoder $h \in \mathcal{H}$ that adheres to this sparsity structure, and we will show that such an encoder ensures component-wise identifiability of the latent variables. To formalize these, corresponding to each valid encoder $h \in \mathcal{H}$, we define the score change matrix $\mathbf{D}_{\mathrm{t}}(h)$ with entries:

$$[\mathbf{D}_{\mathrm{t}}(h)]_{i,m} \triangleq \mathbb{E}\Big[\big| \boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; h) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; h) \big|_i \Big] \ , \tag{72}$$

where expectations are under the measures of latent score functions induced by the probability measure of observational data. We also denote the *true* score change matrix under true encoder g^{-1} by $\mathbf{D}_{\mathrm{t}} \triangleq \mathbf{D}_{\mathrm{t}}(g^{-1})$ with entries

$$[\mathbf{D}_{\mathrm{t}}]_{i,m} \triangleq \mathbb{E}\Big[\big| \boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z}) \big|_i \Big] \ , \quad \forall i, m \in [n] \ . \tag{73}$$

28

For clarity in the exposition of the ideas, we consider coupled environments here, i.e., $I^m = \tilde{I}^m$ for all $m \in [n]$. Then, using (33) in Theorem 7 we have

$$[\mathbf{D}_{\mathrm{t}}]_{i,m} \neq 0 \quad \Longleftrightarrow \quad i = I^m \ . \tag{74}$$

This implies that $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}\}$ is a permutation matrix, $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}\} = \mathbf{P}_{\mathcal{I}}^{\top}$ and

$$\mathbb{1}\{[\mathbf{D}_{\mathrm{t}}]_{:,m}\} = \mathbf{e}_{\ell} \ , \quad \text{where } \ell = I^m \ , \tag{75}$$

and $\mathbf{e}_{\ell}$ denotes the ℓ-th standard basis vector. Subsequently, the key idea for identifying the encoder is that the changes between score estimates $\mathbf{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h)$ and $\tilde{\mathbf{s}}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{z}}; h)$ should match the sparsity structure for those under the true encoder g^{-1}. In the next result, we leverage the sparsity structure in (75) for two interventions on the same node to identify an intervened variable, formalized as follows.

Theorem 22 (General – Node-level Identifiability) *Two hard interventions (with interventional discrepancy) on the same target node $I^m = \tilde{I}^m = \ell$ suffice to recover Z_{ℓ} up to a diffeomorphism. In particular, the following encoder*

$$\hat{h}_m = \arg\min_{h \in \mathcal{H}} \left\| [\mathbf{D}_{\mathrm{t}}(h)]_{:,m} - \mathbf{e}_m \right\|^2 \tag{76}$$

satisfies $\left[\hat{\mathbf{Z}}(\mathbf{X}; \hat{h}_m)\right]_m = \phi_{\ell}(Z_{\ell})$ *for a diffeomorphism* $\phi_{\ell} : \mathbb{R} \to \mathbb{R}$.

We will show shortly that given a complete set of coupled interventions, Theorem 22 will readily imply the identifiability of all variables (in a similar spirit to going from partial to complete identifiability in Section 5). Furthermore, the results for uncoupled interventions will also leverage this key result. Hence, we provide the detailed proof of Theorem 22 as follows.

Proof: We start by providing a direct result of Theorem 9, which will be used in multiple instances in this section. Consider an encoder $h \in \mathcal{H}$ and an invertible matrix $\mathbf{A} \in \mathsf{GL}(n, \mathbb{R})$. Then, using Theorem 9, score functions of $\hat{\mathbf{Z}}(\mathbf{X}; h)$ and $\hat{\mathbf{Z}}(\mathbf{X}; \mathbf{A} \cdot h)$ are related by

$$\mathbf{s}_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}}; \mathbf{A} \cdot h) = \mathbf{A}^{-\top} \cdot \mathbf{s}_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}}; h) \ . \tag{77}$$

Recall that score change matrices are defined via the absolute value of the score differences. Therefore, (77) implies that for a positively scaled permutation matrix $\mathbf{A}$ we have

$$\mathbf{D}_{\mathrm{t}}(\mathbf{A} \cdot h) = \mathbf{A}^{-\top} \cdot \mathbf{D}_{\mathrm{t}}(h) \ , \quad \forall h \in \mathcal{H} \ . \tag{78}$$

Denote the interventional environments by $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ with $I^m = \tilde{I}^m = \ell$. We show that the minimum value of (76) is zero as follows. Let $h^* = \mathbf{D}_{\mathrm{t}}^{\top} \cdot g^{-1}$. Recall that $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}\} = \mathbf{P}_{\mathcal{I}}^{\top}$, and by definition, $\mathbf{D}_{\mathrm{t}}$ has only nonzero entries. Thus, using (78), we have $\mathbf{D}_{\mathrm{t}}(h^*) = \mathbf{D}_{\mathrm{t}}^{-1} \cdot \mathbf{D}_{\mathrm{t}} = \mathbf{I}_{\mathrm{n} \times \mathrm{n}}$, which makes the objective in (76) zero. Next, we show that *any* solution of (76) recovers $\hat{Z}_{\ell}$ up to a diffeomorphism.

Consider encoder $\hat{h}_m \in \mathcal{H}$ that satisfies $[\mathbf{D}_{\mathrm{t}}(\hat{h}_m)]_{:,m} = \mathbf{e}_m$. Let $\hat{\mathbf{Z}}$ denote $\hat{\mathbf{Z}}(\mathbf{X}; \hat{h}_m)$ and $f \triangleq \hat{h}_m \circ g$, so we have $\hat{\mathbf{Z}} = f(\mathbf{Z})$ and $\mathbf{Z} = f^{-1}(\hat{\mathbf{Z}})$. Using score difference transformation

property in (40) and one-sparse property of $[\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})]$ via Theorem 7(iii), we have

$$[\mathbf{D}_{\mathrm{t}}(\hat{h}_m)]_{i,m} = \mathbb{E}\Big[\big|\boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}})\big|_i\Big] \tag{79}$$

$$= \mathbb{E}\Big[\big|\big[J_f^{-\top}(\mathbf{Z})\big]_i \cdot \big[\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\big]\big|\Big] \tag{80}$$

$$= \mathbb{E}\Big[\big[J_f^{-1}(\mathbf{Z})\big]_{\ell,i} \cdot \big[\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\big]_{\ell}\Big] . \tag{81}$$

By definition of $\hat{h}_m$, $[\mathbf{D}_{\mathrm{t}}(\hat{h}_m)]_{i,m} = 0$ for all $i \neq m$. Also, interventional discrepancy between $q_\ell(z_l)$ and $\tilde{q}_\ell(z_l)$ implies that $[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})]_\ell \neq 0$ except for a null set. Hence, if $[J_f^{-1}(\mathbf{z})]_{\ell,i}$ is nonzero over a nonzero-measure set within $\mathbb{R}^n$, then $[\mathbf{D}_{\mathrm{t}}(\hat{h}_m)]_{i,m}$ would not be zero. Therefore, we have $[J_f^{-1}(\mathbf{z})]_{\ell,i} = 0$ except for a null set. Since J_f^{-1} is a continuous function, this implies that

$$[J_f^{-1}(\mathbf{z})]_{\ell,i} = 0 , \quad \forall i \in [n] \setminus m, \ \forall \mathbf{z} \in \mathbb{R}^n, \tag{82}$$

Furthermore, since $J_f^{-1}(\mathbf{z})$ must be invertible, we have $[J_f^{-1}(\mathbf{z})]_{\ell,m} \neq 0$ for all $\mathbf{z} \in \mathbb{R}^n$. Consider $Z_\ell = [f^{-1}(\hat{\mathbf{Z}})]_\ell$. Since $J_f^{-1}(\mathbf{z}) = J_{f^{-1}}(\hat{\mathbf{z}})$, (82) implies that $Z_\ell = \psi(\hat{Z}_m)$ for some diffeomorphism $\psi : \mathbb{R} \to \mathbb{R}$. Thus, $\hat{Z}_m = \psi^{-1}(\hat{Z}_\ell)$ where $\ell = I^m$, which concludes the proof. $\square$

6.2 Complete Identifiability under Coupled Interventions

After achieving the node-level latent variable identifiability from one pair of coupled interventions, we extend the objective function in Theorem 22 for a *complete* set of coupled interventions.

Encoder estimation. In Stage G1 of Algorithm 3, we combine the objectives in (76) for all $m \in [n]$ and form our encoder estimate by solving the following optimization problem

$$\min_{h \in \mathcal{H}} \big\|\mathbf{D}_{\mathrm{t}}(h) - \mathbf{I}_{n \times n}\big\|_{\mathrm{F}}^2 . \tag{83}$$

Recall that the valid encoders set $\mathcal{H}$ consists of invertible encoders over the input space $\mathcal{X}$. Hence, (83) is readily equivalent to solving the following problem for any $\lambda \in \mathbb{R}_+$:

$$\min_h \big\|\mathbf{D}_{\mathrm{t}}(h) - \mathbf{I}_{n \times n}\big\|_{\mathrm{F}}^2 + \lambda \, \mathbb{E}\Big[\big\|h^{-1}(h(\mathbf{X})) - \mathbf{X}\big\|_2^2\Big] . \tag{84}$$

Lemma 23 *Given a complete set of coupled interventions that satisfy interventional discrepancy, output $\hat{h}$ of Algorithm 3 satisfies component-wise latent recovery.*

Proof: Following Theorem 22 for all $m \in [n]$ immediately implies the desired result. First, recall that $h^* = \mathbf{D}_{\mathrm{t}} \cdot g^{-1}$ satisfies $\mathbf{D}_{\mathrm{t}}(h^*) = \mathbf{I}_{n \times n}$ and belongs to $\mathcal{H}$. Therefore, h^* makes the objective (84) zero. Next, consider a solution $\hat{h}$ that makes (84) zero. Since $\hat{h}$ makes the reconstruction loss term zero, i.e., it has a valid inverse $\hat{h}^{-1}$, $\hat{h}$ belongs to the set of valid encoders $\mathcal{H}$. Then, for all $m \in [n]$, we have $[\mathbf{D}_{\mathrm{t}}(\hat{h})]_{:,m} = \mathbf{e}_m$. By Theorem 22, this implies

$$\hat{\mathbf{Z}}(\mathbf{X}; \hat{h}) = \mathbf{P}_{\mathcal{I}} \cdot \phi(\mathbf{Z}) , \tag{85}$$

where $\phi(\mathbf{Z}) = (\phi_1(Z_1), \dots, \phi_n(Z_n))$ is a componentwise diffeomorphism, which concludes the proof. $\square$

Algorithm 3 **G**eneralized **S**core-based **C**ausal **L**atent **E**stimation via Interventions (GSCALE-I)

1: **Input:** $\mathcal{H}$, samples of $\mathbf{X}$ from environment $\mathcal{E}^0$ and coupled environments $\{(\mathcal{E}^m, \tilde{\mathcal{E}}^m) : m \in [n]\}$

2: Compute score functions: $\boldsymbol{s}_{\mathbf{X}}, \boldsymbol{s}_{\mathbf{X}}^m$, and $\tilde{\boldsymbol{s}}_{\mathbf{X}}^m$ for all $m \in [n]$.

3: **Stage G1: Encoder estimation**:

$$\hat{h} \leftarrow \arg\min_h \left\| \mathbf{D}_{\mathrm{t}}(h) - \mathbf{I}_{n \times n} \right\|_{\mathrm{F}}^2 + \lambda\, \mathbb{E}\left[\left\| h^{-1}(h(\mathbf{X})) - \mathbf{X} \right\|_2^2 \right]$$

4: Latent variable estimates: $\hat{\mathbf{Z}} = \hat{h}(\mathbf{X})$.

5: **Stage G2: Latent graph estimation**

6: Construct latent DAG estimate $\hat{\mathcal{G}}$ with parent sets

$$\hat{\mathrm{pa}}(m) \triangleq \left\{ i \neq m : \mathbb{E}\left[\left| \boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}) \right|_i \right] \neq 0 \right\}, \quad \forall m \in [n] .$$

7: **return** $\hat{\mathcal{G}}$, $\hat{h}$, and $\hat{\mathbf{Z}}$

Latent graph estimation. Next, we use the encoder estimate $\hat{h}$ for recovering the latent graph $\mathcal{G}$. In Stage G2 of Algorithm 3, we use the score differences between observational and interventional environments and form $\hat{\mathcal{G}}$ with parent sets

$$\hat{\mathrm{pa}}(m) \triangleq \left\{ i \neq m : \mathbb{E}\left[\left| \boldsymbol{s}_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}}; \hat{h}) - \boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{h}) \right|_i \right] \neq 0 \right\}, \quad \forall m \in [n] . \tag{86}$$

Lemma 24 *Given a complete set of coupled interventions that satisfy interventional discrepancy, Algorithm 3 output $\hat{\mathcal{G}}$ and true latent DAG $\mathcal{G}$ are related through a graph isomorphism.*

Proof: Denote $f = \hat{h} \circ g$, so we have $\hat{\mathbf{Z}} = f(\mathbf{Z})$. Using the score transform between $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ and the structure of J_f^{-1} given in (82), we have

$$\left| \boldsymbol{s}_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}}; \hat{h}) - \boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{h}) \right|_i = \left| \left[J_f^{-1}(\mathbf{Z}) \right]_{i, I^i} \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^m(\mathbf{Z}) \right]_i \right| \tag{87}$$

Since $[J_f^{-1}(\mathbf{Z})]_{i, I^i} \neq 0$ for all $\mathbf{z} \in \mathbb{R}^n$, using Theorem 7(i), we find

$$i \in \hat{\mathrm{pa}}(m) \iff \mathbb{E}\left[\left[\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^m(\mathbf{Z}) \right]_{I^i} \right] \neq 0 \iff I^i \in \mathrm{pa}(I^m), \tag{88}$$

which concludes the proof that $\mathcal{G}$ and $\hat{\mathcal{G}}$ are related through a graph isomorphism. $\square$

Theorem 25 (General – Coupled Environments) *Using observational data and interventional data from two coupled hard environments for which the pair $(q_i, \tilde{q}_i)$ satisfies interventional discrepancy for all $i \in [n]$, suffice to identify (i) the latent DAG $\mathcal{G}$ perfectly and (ii) the latent variables $\mathbf{Z}$ up to componentwise diffeomorphisms specified in Theorem 5. Specifically, Algorithm 3 achieves these identifiability guarantees.*

Proof: Combining Theorem 23 and Theorem 24 gives the theorem statements. □

We emphasize that a key contribution of Theorem 25 beyond the identifiability result is that it provides a well-defined objective function, a minimizer of which provably recovers the latent variables and the graph up to identifiability guarantees. To our knowledge, this is the first such result for interventional CRL with general transformations. In comparison, von Kügelgen et al. (2023) also use two hard interventions per node to prove identifiability, albeit without providing a provably correct algorithm. Furthermore, the result in (von Kügelgen et al., 2023) requires that the estimated latent distribution is faithful to the associated candidate graph for all $h \in \mathcal{H}$. Although a faithfulness assumption does not compromise the identifiability result, it is a strong requirement to verify and poses challenges to devise recovery algorithms. In contrast, we only require observational data, which is generally accessible in practice.

Next, we shed light on the role of observational data. Theorem 23 requires only interventional data for identifying the encoder, whereas Theorem 24 uses observational data. We further tighten this result by showing that for DAG recovery, the observational data becomes unnecessary when we have additive noise models and a weak faithfulness condition holds.

Theorem 26 (No Observational Data) *Using interventional data from two* coupled *hard environments for which the pair* $(q_i, \tilde{q}_i)$ *satisfies interventional discrepancy for all* $i \in [n]$, *suffices to identify the latent variables* $\mathbf{Z}$ *up to component-wise diffeomorphisms. If the latent causal model has additive noise,* $p(\mathbf{Z})$ *is twice differentiable and satisfies the adjacency-faithfulness with respect to* $\mathcal{G}$, *then the latent DAG* $\mathcal{G}$ *is also identifiable.*

Proof sketch: The recovery of latent variables follow from Theorem 23. For the recovery of latent DAG, we leverage Theorem 7(iv) under the additive noise assumption. Specifically, using the score difference between environments $\mathcal{E}^i$ and $\tilde{\mathcal{E}}^j$, we obtain $\overline{\mathrm{pa}}(i, j)$ for all $i, j \in [n], i \neq j$. Finally, we perform at most n conditional independence tests to identify all parent sets $\{\mathrm{pa}(i) : i \in [n]\}$. See Appendix C.1 for the complete proof.

6.3 Complete Identifiability under Uncoupled Interventions

In this setting, we consider the case where the interventional environments corresponding to the same nodes are not specified in pairs. That is, not only is it unknown which node is intervened in an environment, but the learner also does not know which two environments intervene on the same node. Hence, additionally, we need to determine the correct *coupling* between the interventional environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$ as well. However, this is not a straightforward objective since we cannot readily determine the intervention target in an environment. To address this issue, we solve the problems of finding the correct coupling and optimizing the estimated score variations together. Let σ denote the permutation that takes $\tilde{\mathcal{I}} = (\tilde{I}^1, \ldots, \tilde{I}^n)$ to $\mathcal{I} = (I^1, \ldots, I^n)$, i.e., $\sigma(\tilde{I}^m) = I^m$ for all $m \in [n]$. We will modify the objective function in (84) to solve for permutation π as an estimate of σ, in addition to solving for encoder h. To this end, we first modify the score change matrix in (72) and define $\mathbf{D}_t(h; \pi)$ as

$$[\mathbf{D}_t(h; \pi)]_{i,m} \triangleq \mathbb{E}\left[\left| s_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; h) - \tilde{s}_{\hat{\mathbf{Z}}}^{\pi_m}(\hat{\mathbf{Z}}; h) \right|_i \right] , \qquad (89)$$

Unlike the case of coupled interventions, we will also leverage the observational data to identify the encoder. As such, we also define the score change matrices for $\mathcal{E}$ and $\tilde{\mathcal{E}}$ with

respect to $\mathcal{E}$ with entries

$$[\mathbf{D}(h)]_{i,m} \triangleq \mathbb{E}\left[\big|s_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}};h) - s_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}};h)\big|_i\right] \ , \tag{90}$$

$$[\tilde{\mathbf{D}}(h)]_{i,m} \triangleq \mathbb{E}\left[\big|s_{\hat{\mathbf{Z}}}(\hat{\mathbf{Z}};h) - \tilde{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}};h)\big|_i\right] \ . \tag{91}$$

Encoder and coupling estimation. We modify the objective in (84) and solve the following constrained optimization problem for any $\lambda \in \mathbb{R}_+$:

$$\min_{h,\pi} \left\|\mathbf{D}_{\mathrm{t}}(h;\pi) - \mathbf{I}_{n\times n}\right\|_{\mathrm{F}}^2 \ + \ \lambda \, \mathbb{E}\left[\big\|h^{-1}(h(\mathbf{X})) - \mathbf{X}\big\|_2^2\right]$$
$$\text{such that} \quad \mathbb{1}\{\mathbf{D}(h)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h)\} \cdot \mathbf{P}_\pi^\top \tag{92}$$
$$\mathbb{1}\{\mathbf{D}(h)\} \odot \mathbb{1}\{[\mathbf{D}(h)]^\top\} = \mathbf{I}_{n\times n} \ .$$

Note that for the true encoder g^{-1}, both $\mathbf{D}(g^{-1})$ and $\tilde{\mathbf{D}}(g^{-1})$ capture the graph structure via Theorem 7(i), albeit under different permutations. Hence, this equivalence of graph structures for the learned encoder is ensured by the first constraint of $\mathbb{1}\{\mathbf{D}(h)\} = \mathbf{P}_\pi \cdot \mathbb{1}\{\tilde{\mathbf{D}}(h)\}$. Also, the acyclicity of the graphs implied by the score differences is ensured by the second constraint, that is $\mathbf{D}(h)$ does not contain 2-cycles. To prove that solving this problem leads to the identifiability of the latent variables $\mathbf{Z}$, we first show that there exists a global minimizer of this problem.

Lemma 27 (Existence) *Encoder $h^* = \left[\mathbf{D}_{\mathrm{t}}(g^{-1};\sigma)\right]^\top \cdot g^{-1}$ and permutation $\pi^* = \sigma$ minimize the objective in* (92).

Proof sketch: See Appendix C.2 for the complete proof. The main intuition is that $\pi^* = \sigma$ makes the problem more similar to the case of coupled interventions. The proof then follows by using the results of the coupled interventions to show that $\mathbf{D}_{\mathrm{t}}(g^{-1};\sigma)$ is a scaled permutation matrix, using (77) to derive the relationship between the score differences under $\hat{h}^*$ and the true encoder g^{-1}, and using the sparsity structure of $\mathbf{D}(g^{-1})$ to show that the constraints are satisfied.

Next, we show that the objective in (92) can attain its minimum value zero *only* for the correct coupling.

Lemma 28 (Feasibility) *The optimization problem in* (92) *attains its minimum value zero only for the correct coupling $\pi = \sigma$.*

Proof sketch: See Appendix C.3 for the complete proof. The main intuition is that the constraints make it impossible to achieve $\mathbf{D}_{\mathrm{t}}(h;\pi) = \mathbf{I}_{n\times n}$ for an incorrect coupling $\pi \neq \sigma$. We prove it by contradiction. We assume that h^* is a solution, hence, $\mathbf{D}_{\mathrm{t}}(h^*) = \mathbf{I}_{n\times n}$, and $\mathbb{1}\{\mathbf{D}(h)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h)\}$. Then, by scrutinizing the *eldest* node in $\mathcal{G}$ with mismatched interventional environments, we show that $\mathbf{D}(h^*) \odot [\mathbf{D}(h^*)]^\top$ cannot be a diagonal matrix, which contradicts the premise that h^* is a feasible solution of (84).

Finally, we denote the minimizer encoder of (92) by $\hat{h}$ and form the graph estimate similarly to the coupled intervention case via (86). Combining these results, we present our strongest result for general transformations.

Theorem 29 (General – Uncoupled Environments) *Using observational data and interventional data from two uncoupled hard environments for which each pair in $(p_i, q_i, \tilde{q}_i)$ satisfies interventional discrepancy for all $i \in [n]$, suffice to identify (i) the latent DAG $\mathcal{G}$ perfectly and (ii) the latent variables $\mathbf{Z}$ up to component-wise diffeomorphisms.*

Proof: Theorem 28 shows that minimizers of the objective in (92) necessarily have $\pi = \sigma$. Under this correct coupling, any minimizer of the constrained optimization problem in (92) that attains the minimum value zero is also a minimizer of the unconstrained optimization problem in (84). Finally, Theorem 23 and Theorem 27 show that such a shared minimizer exists, that is $h^* = [\mathbf{D}_t(g^{-1})]^\top \cdot g^{-1}$. Hence, by Theorem 25, minimizer $\hat{h}$ of (92) satisfies component-wise latent recovery. Similarly, using $\hat{h}$, proof of the perfect graph recovery follows from Theorem 24. □

Theorem 29 shows that using observational data enables us to resolve any mismatch between the uncoupled environment sets and shows identifiability in the setting of uncoupled environments. This generalizes the identifiability result of von Kügelgen et al. (2023), which requires coupled environments.

Comparison to Theorem 25. We note that Theorem 29 requires slightly stronger interventional discrepancy conditions than Theorem 25. In particular, when environments are coupled, we only need $(q_i, \tilde{q}_i)$ to satisfy the interventional discrepancy. On the other hand, to find the correct coupling while performing CRL, Theorem 29 requires each of the pairs $\{(q_i, \tilde{q}_i), (p_i, q_i), (p_i, \tilde{q}_i)\}$ to satisfy interventional discrepancy.

Remark 30 *For the nonparametric identifiability results, having an oracle that solves the functional optimization problems (84) and (92) is sufficient. Solving these two problems in their most general form requires calculus of variations. These two problems, however, for any desired parameterized family of functions $\mathcal{H}$ (e.g., linear, polynomial, and neural networks), reduce to parametric optimization problems.*

6.4 Intervention Extrapolation without CRL

So far, we have studied learning the latent causal representations using the data from single-node interventional environments. A related research problem is extrapolating to unseen combinations of interventions. Namely, given a set of interventions $\mathcal{I} = \{I^1, \dots, I^k\}$, it is desired to emulate the data from an unseen combination of these interventions, e.g., accessing multi-node interventional data using only the single-node interventional data. This is especially important in domains where interventions can be costly or not viable, e.g., not all combinations of different drugs can be clinically tested. Causal representation learning literature has also taken an interest in this problem. Specifically, Zhang et al. (2023, Theorem 3) show for polynomial transformations that given interventions $\mathcal{I} = \{I^1, \dots, I^k\}$, one can sample from any intervention $I \subseteq \mathcal{I}$ *after* learning the latent representations. We argue that extrapolation to unseen combinations of interventions can be achieved on the observed space *without performing CRL* for the general transformations.

Consider two single-node interventional environments, $\mathcal{E}^1$ and $\mathcal{E}^2$. Without loss of generality, suppose that $I^1 = \{1\}$ and $I^2 = \{2\}$. Also consider the observational environment $\mathcal{E}^0$ with $I^0 = \emptyset$ and the *unseen* double-node interventional environment $\mathcal{E}^m$ with $I^m = \{1, 2\}$.

34

First, using the score function decompositions in (18) and (19), we have

$$s(\mathbf{z}) = \nabla_{\mathbf{z}} \log p_1(z_1 \mid \mathbf{z}_{\mathrm{pa}(1)}) + \nabla_{\mathbf{z}} \log p_2(z_2 \mid \mathbf{z}_{\mathrm{pa}(2)}) + \sum_{k=3}^{n} \nabla_{\mathbf{z}} \log p_k(z_k \mid \mathbf{z}_{\mathrm{pa}(k)}) , \quad (93)$$

$$s^1(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_1(z_1 \mid \mathbf{z}_{\mathrm{pa}(1)}) + \nabla_{\mathbf{z}} \log p_2(z_2 \mid \mathbf{z}_{\mathrm{pa}(2)}) + \sum_{k=3}^{n} \nabla_{\mathbf{z}} \log p_k(z_k \mid \mathbf{z}_{\mathrm{pa}(k)}) , \quad (94)$$

$$s^2(\mathbf{z}) = \nabla_{\mathbf{z}} \log p_1(z_1 \mid \mathbf{z}_{\mathrm{pa}(1)}) + \nabla_{\mathbf{z}} \log q_2(z_2 \mid \mathbf{z}_{\mathrm{pa}(2)}) + \sum_{k=3}^{n} \nabla_{\mathbf{z}} \log p_k(z_k \mid \mathbf{z}_{\mathrm{pa}(k)}) , \quad (95)$$

$$s^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_1(z_1 \mid \mathbf{z}_{\mathrm{pa}(1)}) + \nabla_{\mathbf{z}} \log q_2(z_2 \mid \mathbf{z}_{\mathrm{pa}(2)}) + \sum_{k=3}^{n} \nabla_{\mathbf{z}} \log p_k(z_k \mid \mathbf{z}_{\mathrm{pa}(k)}) . \quad (96)$$

Then, we have

$$\left(s^1(\mathbf{z}) - s(\mathbf{z}) \right) + \left(s^2(\mathbf{z}) - s(\mathbf{z}) \right) = s^m(\mathbf{z}) - s(\mathbf{z}) . \quad (97)$$

Next, recall that Theorem 8 gives us the relationship for going from the latent score differences to observed score differences. Applying it to observed $\mathbf{X}$ in environment pairs $(\mathcal{E}^1, \mathcal{E}^0)$, $(\mathcal{E}^2, \mathcal{E}^0)$, and $(\mathcal{E}^m, \mathcal{E}^0)$, and using (97) we obtain

$$\left(s_{\mathbf{X}}^1(\mathbf{x}) - s_{\mathbf{X}}(\mathbf{x}) \right) + \left(s_{\mathbf{X}}^2(\mathbf{x}) - s_{\mathbf{X}}(\mathbf{x}) \right) = \left[[J_g(\mathbf{z})]^\dagger \right]^\top \cdot \left[\left(s^1(\mathbf{z}) - s(\mathbf{z}) \right) + \left(s^2(\mathbf{z}) - s(\mathbf{z}) \right) \right] \quad (98)$$

$$= \left[[J_g(\mathbf{z})]^\dagger \right]^\top \left(s^m(\mathbf{z}) - s(\mathbf{z}) \right) \quad (99)$$

$$\stackrel{(37)}{=} s_{\mathbf{X}}^m(\mathbf{x}) - s_{\mathbf{X}}(\mathbf{x}) . \quad (100)$$

This means that, given score functions of observed data from environments $\mathcal{E}^0$, $\mathcal{E}^1$, and $\mathcal{E}^2$, we can obtain the score function of the unseen interventional environment with $I^m = \{I^1, I^2\}$, and subsequently generate data from this synthetic environment (using techniques like Langevin sampling (Welling and Teh, 2011) that uses score function as the drift vector field). Note that this process does not require learning the latent causal representations and can be applied to obtain the observed score function $s_{\mathbf{X}}^m$ for any intervention $I^m \subset \mathcal{I}$.

Finally, we note that concurrent work by Jain et al. (2024) makes a similar observation for the extrapolation of two single-node interventions without performing CRL. In particular, (Jain et al., 2024) focuses on detecting pairwise interactions between two biological perturbations (i.e., interventions), investigates the conditions under which two perturbations are *separable*, e.g., distinct single-node interventions, and designs statistical tests for the separability of two perturbations. If the two perturbations are confirmed to be separable, our results indicate that we can use their score functions to generate synthetic environments that emulate combined perturbation effects.

7. Empirical Studies

In this section, we provide empirical assessments of our theoretical guarantees. Specifically, we empirically evaluate the performance of the LSCALE-I (Section 7.1) and GSCALE-I

(Section 7.2) algorithms for recovering the latent causal variables $\mathbf{Z}$ and the latent DAG $\mathcal{G}$ on synthetic data. In Section 7.3, we compare the performance of LSCALE-I to that of the existing algorithms in the closely related literature on both synthetic and biological data. Next, we also apply GSCALE-I on image data to demonstrate the potential of our approach in realistic high-dimensional datasets (Section 7.5). Finally, we note that any desired score estimator can be modularly incorporated into our algorithms. In Section 7.6, we assess our performance's sensitivity to the estimators' quality. [7] Additional results and further implementation details are deferred to Appendix E.

Evaluation metrics. The objectives are recovering the graph $\mathcal{G}$ and the latent variables $\mathbf{Z}$. We use the following metrics to evaluate the accuracy of LSCALE-I and GSCALE-I in recovering these (depending on the specifics of the transformations and interventions, we will also have more specific metrics). For each metric, we will report the mean and standard error over multiple runs.

- **Structural Hamming distance:** For assessing the recovery of the latent DAG, we report structural Hamming distance (SHD) between the estimate $\hat{\mathcal{G}}$ and true DAG $\mathcal{G}$. This captures the number of edge operations (add, delete, flip) needed to transform $\hat{\mathcal{G}}$ to $\mathcal{G}$.

- **Mean correlation coefficient:** For the recovery of the latent variables, we use the mean correlation coefficient (MCC), which was introduced in (Khemakhem et al., 2020b) and commonly used as a standard metric in CRL. Specifically, MCC measures the linear correlation between the estimated and ground-truth latent variables. Since the recovery of latent variables is up to permutations, it is reported for the best matching permutation between the components of $\hat{\mathbf{Z}}$ and $\mathbf{Z}$, i.e.,

$$\mathrm{MCC}(\mathbf{Z}, \hat{\mathbf{Z}}) \triangleq \max_{\pi} \frac{1}{n} \sum_{i \in [n]} \mathrm{corr}(Z_i, \hat{Z}_{\pi(i)}) \ . \tag{101}$$

Score functions. LSCALE-I and GSCALE-I algorithms, in their first steps, compute estimates of the score differences in the observational environment. The designs of our algorithms are agnostic to how this is performed, i.e., any reliable method for estimating the score differences can be adopted and incorporated into our algorithms in a modular way. In our experiments, we adopt two score estimators necessary for describing the different aspects of our results.

- **Perfect score oracle for identifiability:** Identifiability, by definition, refers to the possibility of recovering the causal graph and latent variables under all idealized assumptions for the data. Assessing the identifiability guarantees formalized in Theorems 14–29 requires using *perfect* estimates for the score differences. Hence, we adopt a perfect score oracle for evaluating identifiability. Specifically, we use a perfect score oracle that computes the score difference inputs in LSCALE-I and GSCALE-I by leveraging Theorem 8 and using the ground truth score functions $\boldsymbol{s}$, $\boldsymbol{s}^m$ and $\tilde{\boldsymbol{s}}^m$ (see Appendix E.1 for details).

---

7. The codebase for the algorithms and simulations is available at:
   https://github.com/acarturk-e/score-based-crl.

- **Data-driven score estimates:** For evaluating the accuracy of our algorithms in practice, we need real score estimates, which are inevitably noisy. For this purpose, when the pdf $p_{\mathbf{X}}$ has a parametric form and the score function $s_{\mathbf{X}}$ has a closed-form expression, then we can estimate the parameters to form an estimated score function. For instance, when $\mathbf{X}$ follows a linear Gaussian distribution, then we have $s_{\mathbf{X}}(\mathbf{x}) = -\Theta \cdot \mathbf{x}$ in which Θ is the precision matrix of $\mathbf{X}$ and can be estimated from samples of $\mathbf{X}$. In other cases in which $s_{\mathbf{X}}$ does not have a known closed-form, we adopt nonparametric score estimators. In particular, we use sliced score matching with variance reduction (SSM-VR) for score estimation due to its efficiency and accuracy for downstream tasks (Song et al., 2020). We also introduce a classification-based score difference estimation method, inspired by Gutmann and Hyvärinen (2012, Section 2.1). The key observation is that given two distributions p and q, the optimal minimum cross-entropy classifier for distinguishing the samples from two distributions is the log density ratio function $\log p/q$. The difference between the score functions of these distributions, $\nabla \log p - \nabla \log q = \nabla \log p/q$ is exactly the gradient of the learned function, which enables us to estimate score differences using a classifier directly.

7.1 LSCALE-I Algorithm for Linear Transformations

Data generation. To generate $\mathcal{G}$, we use Erdős-Rényi model with density 0.5 and $n \in \{5, 8\}$ nodes, which is generally the size of the latent graphs considered in CRL literature. We consider the observed dimension $d = 100$ and generate 100 latent graphs. For the causal mechanisms, we adopt both linear and nonlinear models:

1. **Linear causal model:** We adopt the linear Gaussian model with

$$Z_i = \mathbf{A}_i \cdot \mathbf{Z} + N_i \ , \qquad \forall i \in [n] \ , \tag{102}$$

   where $\mathbf{A}_i \in \mathbb{R}^{1 \times n}$ are the rows of the weight matrix $\mathbf{A}$ in which $\mathbf{A}_{i,j} \neq 0$ if and only $j \in \mathrm{pa}(i)$. The nonzero edge weights are sampled from $\mathrm{Unif}(\pm[0.5, 1.5])$, and the noise terms are zero-mean Gaussian variables with variances σ_i^2 sampled from $\mathrm{Unif}([0.5, 1.5])$. For node i, a hard intervention is given by $Z_i = \bar{N}_i$ where $\bar{N}_i \sim \mathcal{N}(0, \frac{\sigma_i^2}{4})$, and a soft intervention is given by $Z_i = \frac{\bar{\mathbf{A}}_i}{2} \cdot \mathbf{Z} + N_i$.

2. **Quadratic causal model:** We adopt an additive noise model with

$$Z_i = \sqrt{\mathbf{Z}_{\mathrm{pa}(i)}^\top \cdot \mathbf{Q}_i \cdot \mathbf{Z}_{\mathrm{pa}(i)}} + N_i \ , \tag{103}$$

   where $\{\mathbf{Q}_i : i \in [n]\}$ are positive-definite matrices, and the noise terms are zero-mean Gaussian variables with variances σ_i^2 sampled randomly from $\mathrm{Unif}([0.5, 1.5])$. For node i, a hard intervention is given by $Z_i = \bar{N}_i$ where $\bar{N}_i \sim \mathcal{N}(0, 5 \cdot \sigma_i^2)$ to ensure a non-negligible change. A soft intervention is given by

$$Z_i = \frac{1}{2} \sqrt{\mathbf{Z}_{\mathrm{pa}(i)}^\top \cdot \mathbf{Q}_i \cdot \mathbf{Z}_{\mathrm{pa}(i)}} + \bar{N}_i \ .$$

3. **Multi-layer perceptron (MLP) causal model**: We adopt an additive noise model with $Z_i = f_i(\mathbf{Z}_{\mathrm{pa}(i)}) + N_i$ where f_i is parameterized as a randomly initialized two-layer

Table 3: LSCALE-I for a linear causal model with **one hard** intervention per node (varying n_s)

			perfect scores			noisy scores		
n	d	n_s	MCC	ℓ_scale	$\mathrm{SHD}(\mathcal{G},\hat{\mathcal{G}})$	MCC	ℓ_scale	$\mathrm{SHD}(\mathcal{G},\hat{\mathcal{G}})$
5	100	5000	1.00 ± 0.00	0.02 ± 0.00	0.17 ± 0.04	1.00 ± 0.00	0.06 ± 0.01	0.11 ± 0.03
5	100	10000	1.00 ± 0.00	0.01 ± 0.00	0.09 ± 0.03	1.00 ± 0.00	0.04 ± 0.00	0.09 ± 0.04
5	100	50000	1.00 ± 0.00	0.01 ± 0.00	0.03 ± 0.01	1.00 ± 0.00	0.03 ± 0.00	0.07 ± 0.03
8	100	5000	1.00 ± 0.00	0.03 ± 0.00	0.03 ± 0.02	1.00 ± 0.00	0.08 ± 0.00	0.46 ± 0.08
8	100	10000	1.00 ± 0.00	0.02 ± 0.00	0.06 ± 0.02	1.00 ± 0.00	0.06 ± 0.00	0.18 ± 0.04
8	100	50000	1.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	1.00 ± 0.00	0.03 ± 0.00	0.04 ± 0.02

MLP, with hidden dimension 32 and ReLU activation function. For node i, a hard intervention is given by $Z_i = \bar{N}_i$ where $\mathcal{N}(0, 5 \cdot \sigma_i^2)$, and a soft intervention is given by $Z_i = \frac{1}{2} f_i(\mathbf{Z}_{\mathrm{pa}(i)}) + \bar{N}_i$.

For each graph, we sample n_s independent and identically distributed (i.i.d.) samples of $\mathbf{Z}$ from each environment. We consider $n_\mathrm{s} \in \{5000, 10000, 50000\}$ to investigate the effect of the number of samples on the performance of LSCALE-I. The observed variables $\mathbf{X}$ are generated according to $\mathbf{X} = \mathbf{G} \cdot \mathbf{Z}$, in which $\mathbf{G} \in \mathbb{R}^{d \times n}$ is a randomly sampled full-rank matrix.

7.1.1 HARD INTERVENTIONS

Theorem 16 ensures scaling consistency and perfect DAG recovery under hard interventions for linear transformations. As such, we assess the recovery of the latent DAG by the SHD between the estimate $\hat{\mathcal{G}}$ and true graph $\mathcal{G}$. For latent variable recovery, we report MCC between $\hat{\mathbf{Z}}$ and $\mathbf{Z}$. Note that $\hat{\mathbf{Z}} = (\mathbf{H} \cdot \mathbf{G}) \cdot \mathbf{Z}$ implies that the effective transform recovery can also be measured by the closeness of $\mathbf{H} \cdot \mathbf{G}$ to the identity matrix. Therefore, in addition to MCC, we also report the normalized *effective transform error*, defined as

$$\ell_\mathrm{scale} \triangleq \|\mathbf{H} \cdot \mathbf{G} - \mathbf{I}_{n \times n}\|_2 . \tag{104}$$

Table 3 shows the performance of the LSCALE-I algorithm using perfect scores and noisy scores under hard interventions on linear causal models. The first observation is that we achieve excellent performance in latent variable recovery, as demonstrated by a perfect MCC and nearly zero effective transform error ℓ_scale, even when using noisy score estimates. For graph recovery, SHD between the estimated and true latent graphs reduces to less than 0.1 even when using noisy scores given enough samples, e.g., $n_s = 50000$ in Table 3. Another key observation is that the results remain consistent while the dimension of observed variables increases from $d = 25$ to $d = 200$, as shown in Table 4. This confirms our analysis that the performance of LSCALE-I is agnostic to the dimension of the observations. Hence, we suffice by using $d = 100$ for the rest of the experiments.

Causal models with more complex score functions. Next, we test the performance of LSCALE-I on quadratic and MLP causal models specified earlier. A difference in this setting compared to linear causal models is that we estimate the score functions of the

Table 4: LSCALE-I for a linear causal model with **one hard** intervention per node (varying d)

n	d	n_{s}	perfect scores			noisy scores		
			MCC	ℓ_{scale}	SHD$(\mathcal{G},\hat{\mathcal{G}})$	MCC	ℓ_{scale}	SHD$(\mathcal{G},\hat{\mathcal{G}})$
5	25	10000	1.00 ± 0.00	0.01 ± 0.00	0.28 ± 0.05	1.00 ± 0.00	0.03 ± 0.00	0.03 ± 0.02
5	50	10000	1.00 ± 0.00	0.01 ± 0.00	0.23 ± 0.04	1.00 ± 0.00	0.03 ± 0.00	0.04 ± 0.02
5	100	10000	1.00 ± 0.00	0.01 ± 0.00	0.06 ± 0.02	1.00 ± 0.00	0.04 ± 0.00	0.09 ± 0.03
5	200	10000	1.00 ± 0.00	0.01 ± 0.00	0.07 ± 0.03	1.00 ± 0.00	0.05 ± 0.00	0.16 ± 0.04

Table 5: LSCALE-I for a quadratic causal model with **one hard** intervention per node ($n_{\mathrm{s}} = 50000$).

n	d	perfect scores			noisy scores		
		MCC	ℓ_{scale}	SHD$(\mathcal{G},\hat{\mathcal{G}})$	MCC	ℓ_{scale}	SHD$(\mathcal{G},\hat{\mathcal{G}})$
5	100	1.00 ± 0.00	0.01 ± 0.00	0.03 ± 0.02	0.93 ± 0.01	0.69 ± 0.02	2.62 ± 0.20

observed variables using SSM-VR (Song et al., 2020) as $p_{\mathbf{X}}$ is not amenable to parameter estimation. For quadratic causal models, Table 5 shows that using perfect scores, LSCALE-I performs nearly perfectly. Under noisy scores, an MCC of 0.93 indicates a strong performance for the latent variable recovery. The graph recovery performance suffers more from noisy score estimates, yet it remains reasonable with an approximate SHD of 2.62 (where the expected number of true edges is 5). Table 6 shows a similar trend, where perfect scores ensure perfect performance, whereas noisy score estimates lead to a degradation in graph recovery performance.

Effect of noisy score estimates on graph recovery under noisy scores. Table 5 and Table 6 show that graph recovery performance suffers more than the latent variable recovery under noisy score estimates. This discrepancy is due to the difficulty of applying (59) under noisy scores. Specifically, we observe that the nonzero entries in true score differences (entries of $\boldsymbol{d}^m$) can be small, especially for MLP causal models. As such, under noisy score estimates, it becomes more difficult to successfully threshold the empirical score difference quantities to identify the edges of the latent graph. In Section 7.6, we provide further empirical evaluations on the effect of score estimation quality on the performance of LSCALE-I.

7.1.2 Soft Interventions

To evaluate soft interventions, we first consider linear causal models. In this setting, Theorem 14 ensures the recovery of the latent variables and the latent DAG up to ancestors when using soft interventions. So, we report the SHD between the transitive closure of the estimate $\hat{\mathcal{G}}_{\mathrm{tc}}$ and that of the true graph $\mathcal{G}_{\mathrm{tc}}$. Recall that recovery of latent variables for generic soft interventions is guaranteed up to mixing with parents. To measure the accuracy of $\hat{\mathbf{Z}} = \mathbf{H} \cdot \mathbf{G} \cdot \mathbf{Z}$ with respect to this guarantee, we define $\mathbf{L}_{\mathrm{pa}}$ with entries $[\mathbf{L}_{\mathrm{pa}}]_{i,j} = \mathbb{1}\{j \in \overline{\mathrm{pa}}(i)\}$, and define ℓ_{pa} by

$$\ell_{\mathrm{pa}} \triangleq \|\mathbf{H} \cdot \mathbf{G} \odot (\mathbf{1}_{n \times n} - \mathbf{L}_{\mathrm{pa}})\|_2 \ , \tag{105}$$

which effectively measures the effect of incorrect mixing in estimated latent variables. Table 7 shows that by using perfect scores from as few as $n_{\mathrm{s}} = 5000$ samples, we meet the theoretical

Table 6: LSCALE-I for an MLP causal model with **one hard** intervention per node ($n_{\mathrm{s}} = 50000$).

		perfect scores			noisy scores		
n	d	MCC	ℓ_{scale}	SHD$(\mathcal{G}, \hat{\mathcal{G}})$	MCC	ℓ_{scale}	SHD$(\mathcal{G}, \hat{\mathcal{G}})$
5	100	1.00 ± 0.00	0.03 ± 0.00	0.01 ± 0.01	0.94 ± 0.01	0.62 ± 0.02	4.27 ± 0.20

Table 7: LSCALE-I for a linear causal model with **one soft** intervention per node.

			perfect scores			noisy scores		
n	d	n_{s}	MCC	ℓ_{pa}	SHD$(\mathcal{G}_{\mathrm{tc}}, \hat{\mathcal{G}}_{\mathrm{tc}})$	MCC	ℓ_{pa}	SHD$(\mathcal{G}_{\mathrm{tc}}, \hat{\mathcal{G}}_{\mathrm{tc}})$
5	100	5000	0.98 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.98 ± 0.00	0.04 ± 0.00	0.59 ± 0.11
5	100	10000	0.98 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.00	0.03 ± 0.00	0.36 ± 0.08
5	100	50000	0.98 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.00	0.01 ± 0.00	0.28 ± 0.06
8	100	5000	0.98 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.00	0.07 ± 0.00	3.84 ± 0.36
8	100	10000	0.98 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.00	0.05 ± 0.00	1.23 ± 0.20
8	100	50000	0.98 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.00	0.02 ± 0.00	0.49 ± 0.10

identifiability guarantees, implied by zero values of SHD and incorrect mixing norm ℓ_{pa}. Similar to the case of hard interventions, we observe that increasing the number of samples improves the performance under noisy scores. For instance, given $n_{\mathrm{s}} = 50000$ samples, the average SHD between the transitive closures $\hat{\mathcal{G}}_{\mathrm{tc}}$ and $\mathcal{G}_{\mathrm{tc}}$ is approximately 0.5 for $n = 8$, where graph density being 0.5 implies that the expected number of edges is 14. Also, the latent variable recovery becomes near perfect, indicated by MCC of 0.98 and near-zero incorrect mixing norm ℓ_{pa}.

Next, we consider quadratic causal models. In this case, Assumption 2 is satisfied, and Theorem 18 ensures the perfect recovery of the latent DAG and the recovery of the latent variables up to surrounding variables. Hence, we report SHD between true $\mathcal{G}$ and estimate $\hat{G}$, and the incorrect mixing norm for this setting given by

$$\ell_{\mathrm{sur}} \triangleq \|\mathbf{H} \cdot \mathbf{G} \odot (\mathbf{1}_{n \times n} - \mathbf{L}_{\mathrm{sur}})\|_2 \ , \tag{106}$$

where $\mathbf{L}_{\mathrm{sur}}$ is defined with entries $[\mathbf{L}_{\mathrm{pa}}]_{i,j} = \mathbb{1}\{j \in \overline{\mathrm{sur}}(i)\}$. Table 8 shows that when using perfect scores, LSCALE-I performs nearly perfectly, verifying the results of Theorem 18. Similarly to the case of hard interventions, the performance of LSCALE-I suffers under noisy score estimations, while it remains reasonable, e.g., an SHD of 2.79 with respect to the true graph when using only soft interventions.

7.2 GSCALE-I Algorithm for General Transformations

Next, we focus on nonlinear transformations to showcase the settings for which the existing interventional CRL literature lacks provably correct algorithms and provides only identifiability results under nonlinear transformations.

Choice of nonlinearity. In this section, we consider M-layer perceptrons with tanh activation as our class of nonlinear transformations. Specifically, we consider functions of the

Table 8: LSCALE-I for a quadratic causal model with **one soft** intervention per node ($n_{\mathrm{s}} = 50000$).

		perfect scores			noisy scores		
n	d	MCC	ℓ_{sur}	$\mathrm{SHD}(\mathcal{G}, \hat{\mathcal{G}})$	MCC	ℓ_{sur}	$\mathrm{SHD}(\mathcal{G}, \hat{\mathcal{G}})$
5	100	0.87 ± 0.01	0.34 ± 0.04	0.53 ± 0.09	0.60 ± 0.01	0.51 ± 0.03	2.79 ± 0.18

form

$$\mathbf{X} = g(\mathbf{Z}) = (\tanh \circ \mathbf{A}^M \circ \cdots \tanh \circ \mathbf{A}^1)(\mathbf{Z}) \ , \tag{107}$$

in which tanh is applied element-wise, and parameters $\{\mathbf{A}^1, \ldots, \mathbf{A}^M\}$ are compatible, randomly sampled *full column rank* matrices. We consider two settings of increasing complexity: 1-layer MLP, and 3-layer MLP (for nonparametric transformations, see experiments on images in Section 7.5).

7.2.1 SINGLE-LAYER MLP

First, consider the single-layer perceptron as the nonlinear transformation g, given by

$$\mathbf{X} = g(\mathbf{Z}) = \tanh(\mathbf{G} \cdot \mathbf{Z}) \ , \tag{108}$$

where $\mathbf{G} \in \mathbb{R}^{d \times n}$. Because it is a relatively simple nonlinear transformation, this setting enables us to evaluate the performance of GSCALE-I more effectively. Specifically, leveraging (108), we parameterize valid encoders h with parameter $\mathbf{H} \in \mathbb{R}^{n \times d}$, which gives

$$\hat{\mathbf{Z}}(\mathbf{X}; h) = h(\mathbf{X}) = \mathbf{H} \cdot \mathrm{arctanh}(\mathbf{X}) \ , \tag{109}$$

$$\hat{\mathbf{X}} = h^{-1}(\hat{\mathbf{Z}}(\mathbf{X}; h)) = \tanh\left(\mathbf{H}^\dagger \cdot \hat{\mathbf{Z}}(\mathbf{X}; h)\right) \ , \tag{110}$$

$$\hat{\mathbf{Z}} = \mathbf{H} \cdot \mathbf{G} \cdot \mathbf{Z} \ . \tag{111}$$

This is equivalent to reducing the original nonparametric functional estimation problem to a finite-dimensional parameter estimation problem. This enables stable training and large-scale testing. Furthermore, this parametrization enables us to compute the ground truth score functions of the observed variables, allowing us to directly assess the effect of score estimation errors on GSCALE-I performance.

Data generation. In this setting, we focus on quadratic causal models. To generate $\mathcal{G}$ we use the Erdős-Rényi model with density 0.5 and $n \in \{5, 8\}$ nodes. For observational causal mechanisms, we use Equation (103) and the corresponding parameterization. For the two hard interventions on node i, Z_i is set to $N_{q,i} \sim \mathcal{N}(0, \sigma_{q,i}^2)$ and $N_{\tilde{q},i} \sim \mathcal{N}(0, \sigma_{\tilde{q},i}^2)$, where we set $\sigma_{q,i}^2 = \sigma_i^2/4$ and $\sigma_{\tilde{q},i}^2 = 4 \cdot \sigma_i^2$. Similarly to LSCALE-I experiments, we consider a target dimension of $d = 100$ and generate 20 latent graphs, with $n_{\mathrm{s}} = 30000$ samples per environment for each graph.

Results. Recall that for general transformations, we can guarantee $\hat{Z}_i = \phi_i(Z_i)$ for a diffeomorphism ϕ_i. However, $\mathrm{MCC}(Z_i, \phi_i(Z_i))$ can largely deviate from 1 for a nonlinear ϕ_i. For instance, for $Z_i \sim \mathcal{N}(0, 1)$ and $\phi_i(Z_i) = Z_i^3 + 0.1 Z_i$, we have $\mathrm{MCC}(Z_i, \phi_i(Z_i)) \approx 0.786$.

Table 9: GSCALE-I for a quadratic causal model with **two coupled hard** interventions per node. Noisy scores are obtained using SSM-VR with $n_{\text{score}} = 30000$ samples.

			expected num.	perfect scores		noisy scores	
n	d	n_{s}	edges in $\mathcal{G}$	MCC	SHD$(\mathcal{G}, \hat{\mathcal{G}})$	MCC	SHD$(\mathcal{G}, \hat{\mathcal{G}})$
5	100	200	5	1.00 ± 0.00	0.00 ± 0.00	0.85 ± 0.02	4.50 ± 0.38
8	100	500	14	0.95 ± 0.01	1.50 ± 0.27	0.75 ± 0.02	12.9 ± 0.44

On the other hand, when we use the parameterization in (109), the only element-wise diffeomorphism between $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ is an element-wise scaling. Therefore, MCC remains a perfectly informative metric. Table 9 shows that we can almost perfectly recover the latent variables (indicated by MCC of 1) latent DAG for $n = 5$ nodes by using perfect scores. Furthermore, increasing the latent dimension to $n = 8$ has no significant impact on performance.

The effect of the quality of score estimation. When using noisy scores, the performance of GSCALE-I degrades significantly. For instance, MCC goes down to approximately 0.85 for $n = 5$ and 0.75 for $n = 8$. This degradation is similar to what happens when using LSCALE-I on quadratic causal models in Table 8. It is noteworthy that the transition from perfect to noisy scores is remarkably smoother when using LSCALE-I on linear causal models (Table 3 and Table 7). This discrepancy is attributed to the distinct score estimation procedures adopted in the two experimental settings. Specifically, linear Gaussian latent models allow us to directly estimate the parameters of the closed-form score function $s_{\mathbf{X}}$. However, when using a quadratic latent model, we rely on a nonparametric score estimation via SSM-VR (Song et al., 2020). This comparison between two experiment settings and results underscores that the performance gap between the theoretical guarantees and practical results can be significantly mitigated through advances in general score estimation techniques. In Section 7.6, we provide a further empirical evaluation of how the quality of the score estimation affects the final performance.

7.2.2 MULTI-LAYER NONLINEARITY

Next, we consider general MLPs by setting the number of layers to 3, which is in line with the MLP depths considered in CRL literature (Liang et al., 2023). In this setting, we do not use any knowledge of the transformation parametrization and aim to learn a generic NN-based encoder-decoder pair h, h^{-1} parameterized by $\theta_{\text{enc}}, \theta_{\text{dec}}$. In essence, this enables us to test our algorithm's performance on a standard procedure of function approximation via neural networks.

Data generation. In this setting, we focus on linear causal models. To generate $\mathcal{G}$, we use the Erdős-Rényi model with density 0.5 and $n = 5$ nodes. For observational causal mechanisms, we use Equation (102) and its associated parameterization. For the two hard interventions on node i, Z_i is set to $N_{q,i} \sim \mathcal{N}(0, \sigma_{q,i}^2)$ and $N_{\tilde{q},i} \sim \mathcal{N}(0, \sigma_{\tilde{q},i}^2)$, where we set $\sigma_{q,i}^2 = \sigma_i^2/4$ and $\sigma_{\tilde{q},i}^2 = \sigma_i^2 \cdot 4$. Due to the increased computational complexity of training,

Table 10: MCC across different runs in MLP transform + MLP causal model setting

Runs					Mean (std. error)
0.64	0.54	0.43	0.58	0.54	0.54 ± 0.03

we consider target dimension value $d = 5$, and we generate 5 latent graphs and $n_s = 10000$ samples per graph.

Results. GSCALE-I algorithm ensures latent variables recovery up to element-wise transformations. As a proxy, we report its linear analogue, i.e., MCC. We summarize the results of these experiments in Table 10.

In this setting, we obtain a smaller average MCC rate (0.54) compared to the 1-layer nonlinearity experiments in Table 9. Performance degradation can be attributed to two factors: First, making the transformation more complex and making it more difficult to construct *any* kind of autoencoder. In contrast to the parametric single-layer MLP setting in the previous section, fitting a generic NN has considerably higher sample complexity. Second, our methodology requires the estimation of score functions with relatively good accuracy. However, complex distributions can push forward distributions into very complex and, similarly, difficult-to-estimate landscapes, which can limit the accuracy of our algorithms.

7.3 Comparison with the Existing CRL Studies

In this section, we compare the performance of LSCALE-I with those of the approaches designed for comparable settings in (Squires et al., 2023; Zhang et al., 2023).

7.3.1 Comparison with linear causal models on synthetic data

A closely related study in the linear transformation setting is (Squires et al., 2023). As discussed in Section 5, (Squires et al., 2023) presents identifiability guarantees and provides algorithms for linear transformations with linear causal models. For a fair comparison, we perform comparisons for linear Gaussian SEMs under the parameterization detailed in Section 7.1. Experiments with synthetic linear causal models in (Squires et al., 2023) mostly focus on recovering the parameters of the true encoder $\mathbf{G}^\dagger$. Hence, in Table 11, we report performance comparisons in terms of the latent variable recovery metric ℓ_{scale}. Table 11 shows that LSCALE-I achieves close to perfect accuracy with as few as $n_s = 5000$ samples, whereas the algorithm of Squires et al. (2023) requires significantly more samples (more than $n_s = 50,000$) to achieve a reasonable performance.

7.3.2 Comparison with DiscrepancyVAE on synthetic data

In this section, we compare LSCALE-I with the DiscrepancyVAE algorithm of (Zhang et al., 2023) on the setting described in Section 7.1. We consider 10 runs of the quadratic causal model setting with soft interventions, latent dimension $n = 5$, observed dimension $d = 100$, and 50000 samples per environment. We observe that the average SHD of DiscrepancyVAE on this dataset is 3.1 ± 0.7, which is comparable to LSCALE-I performance of 2.79 ± 0.18 reported in Table 8.

Table 11: Comparison of LSCALE-I with the algorithm of Squires et al. (2023) for a linear causal model with **one hard** intervention per node.

n	d	n_{s}	LSCALE-I		(Squires et al., 2023)	
			mean ℓ_{scale}	fraction of $\ell_{\mathrm{scale}} < 0.1$	mean ℓ_{scale}	fraction of $\ell_{\mathrm{scale}} < 0.1$
5	100	5000	0.06	0.92	0.66	0.67
5	100	10000	0.04	0.96	0.51	0.74
5	100	50000	0.03	0.94	0.24	0.88
8	100	5000	0.08	0.86	1.28	0.36
8	100	10000	0.06	0.95	1.09	0.45
8	100	50000	0.03	0.96	0.68	0.66

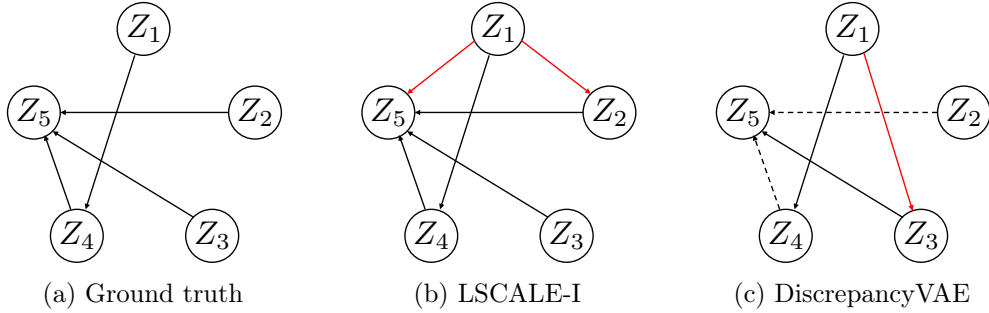(a) Ground truth   (b) LSCALE-I   (c) DiscrepancyVAE

Figure 3: Latent graph recovery comparison. Extra edges are marked in red, and dashed lines mark the missing edges.

To provide a visual comparison, we consider a single run and plot the ground truth latent graph (Figure 3a), the estimate generated by LSCALE-I (Figure 3b), and the estimate generated by DiscrepancyVAE (Figure 3c). The differences between the latter and the ground truth graph are shown in these figures by red (added) and dashed (deleted) edges. We refer further comparison between these methods to Appendix E.3.

7.3.3 Comparison with DiscrepancyVAE on biological data

In this section, we apply our algorithms to the Perturb-seq dataset of Norman et al. (2019), which is also used in Zhang et al. (2023). Following the same pre-processing steps, there are 8,907 unperturbed and 99,590 perturbed cells. Each cell (sample) is denoted by a sparse vector of dimension 5000, which represents the expression levels of select genes. The perturbed cells are generated by CRISPR activation (Gilbert et al., 2014) on one or two target genes out of 105 genes. Hence, this is modeled by a 5000-dimensional observable variable $\mathbf{X}$ and a 105-dimensional latent variable $\mathbf{Z}$ representing the perturbed genes. In this dataset, the ground truth latent graph and variables are unknown, with the exception of limited gene interactions that have been experimentally verified (Norman et al., 2019). For graph estimation, instead of working with a latent dimension of 105, Zhang et al. (2023) first learns groups of perturbation targets subject to a regularity constraint and then learns a latent graph among these groups. Since the methodology of LSCALE-I does not naturally

Table 12: Representative target genes from super nodes.

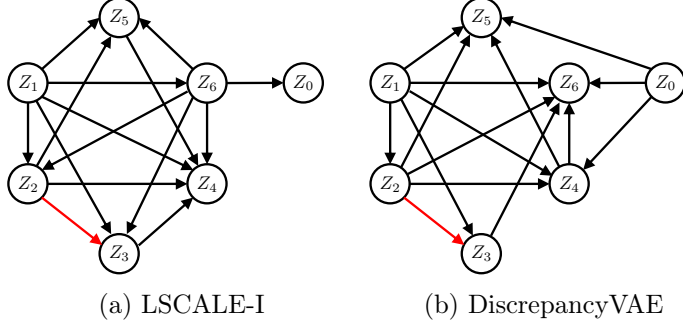| Super node index	Representative gene(s)
0	OSR2
1	TBX2
2	DUSP9
3	MAPK1, ETS2
4	COL2A1
5	SET
6	KLF1

(a) LSCALE-I   (b) DiscrepancyVAE

Figure 4: Graph estimate for Perturb-seq dataset. The edge from DUSP9 to MAPK1/ETS2 is marked in red.

extend to learning such groups of latent nodes, we use the perturbation groupings reported by Zhang et al. (2023) as our super-nodes and run LSCALE-I to learn a latent graph over them. We list representative perturbation targets from each super-node in Table 12.

The most important observation is that LSCALE-I and DiscrepancyVAE both estimate a causal edge from DUSP9 to the MAPK1/ETS2 group, which is a gene activation relation that is demonstrated experimentally (Norman et al., 2019). Due to the lack of ground truth for other graph edges, comparing the algorithms for recovering other edges is not informative. Nevertheless, we provide the estimated latent graph from LSCALE-I (Figure 4a) and DiscrepancyVAE (Figure 4b) for completeness. This verifies that our score-based methodology can recover causal genomics relations.

7.4 Intervention Extrapolation on Biological Data

In Section 6.4, we have demonstrated that it is possible to sample from unseen combinations of given interventions via only learning the score functions of observed variables. We verify this on the Perturb-seq data discussed in Section 7.3 through the following steps:

1. Collect observed variables' data from observational distribution $p_{\mathbf{X}}$, two single-node interventions $p_{\mathbf{X}}^1$ and $p_{\mathbf{X}}^2$, and double-node intervention $p_{\mathbf{x}}^{\{1,2\}}$.

2. Compute score functions $\mathbf{s}_{\mathbf{X}}$, $\mathbf{s}_{\mathbf{X}}^1$, and $\mathbf{s}_{\mathbf{X}}^2$.

3. Perform score function extrapolation for double-node intervention via (100): $\tilde{\mathbf{s}}_{\mathbf{X}}^{\{1,2\}} = \mathbf{s}_{\mathbf{X}}^1 + \mathbf{s}_{\mathbf{X}}^2 - \mathbf{s}_{\mathbf{X}}$.

4. Single-node setting: Generate samples from $\mathbf{s}_{\mathbf{X}}^1$ and $\mathbf{s}_{\mathbf{X}}^2$ using Langevin dynamics (Welling and Teh, 2011), compare them to the original samples from $p_{\mathbf{X}}^1$ and $p_{\mathbf{X}}^2$.

5. Double-node intervention extrapolation: Generate samples from $\tilde{\mathbf{s}}_{\mathbf{X}}^{\{1,2\}}$ using Langevin dynamics, compare them to the original samples from $p_{\mathbf{X}}^{\{1,2\}}$.

We note that Zhang et al. (2023) also report experiments on the same dataset for the single-node data generation and double-node intervention extrapolation settings. However, as discussed in Section 6.4, they perform these tasks using the VAE model for CRL, whereas

our approach shows that learning the latent causal representations is unnecessary. Similarly to Zhang et al. (2023), we use the maximum mean discrepancy (MMD) (Gretton et al., 2012) to measure performance quantitatively. In Table 13, we show that our approach performs at least as well as DiscrepancyVAE on the Perturb-seq dataset. Finally, in Appendix E.4, we illustrate the simulated interventional distributions via UMAP clustering, which visually supports the performance implied by Table 13. These results validate the theoretical analysis that we do not need to learn latent variables to perform intervention extrapolation.

Table 13: MMD evaluation for single-node and double-node intervention extrapolation of our score-based approach and DiscrepancyVAE (Zhang et al., 2023) on Perturb-seq dataset.

	Single-node	Double-node
Score-based	0.057 ± 0.013	0.208 ± 0.036
DiscrepancyVAE	≥ 0.15	≥ 0.2

7.5 Experiments on Image Data

In this section, we perform experiments on CRL where the general transformation is image rendering, a highly nonlinear transformation, by applying the GSCALE-I algorithm to synthetic image data.

Image data generation. For image-based experiments, we follow the setup of the closely related studies in (Ahuja et al., 2023; Buchholz et al., 2023). Specifically, we consider images of the form in Figure 6, which are generated as follows. The pairs of latent variables (Z_{2i-1}, Z_{2i}) describe the coordinates of the i-th ball's center in a $64 \times 64 \times 3$ RGB image. We use two and three balls in our experiments, which corresponds to $n \in \{4, 6\}$ latent variables. We sample the latent graph $\mathcal{G}$ from Erdős–Rényi model for n nodes and set the expected number of edges to $2n$. Given this graph $\mathcal{G}$, we adopt a *truncated* linear Gaussian latent causal model. The details of the linear model are the same as the Section 7.1, with the addition of a second set of interventions represented by the causal mechanisms $\tilde{q}_i$, by setting $Z_i = \tilde{N}_i$, where $\tilde{N}_i \sim \mathcal{N}(0, \frac{\sigma_i^2}{4})$. Latent variables $\mathbf{Z}$ are sampled from the distribution defined by this linear model, but samples with any coordinate absolute value greater than 1 are discarded to truncate the distribution in the box $[-1, 1]^n$. For each latent sample $\mathbf{Z}$, we render an image of the balls centered at the specified coordinates (shifted and scaled to fit entirely within rendered image boundaries) with a radius of 8 pixels and use these as our observed data $\mathbf{X}$. Balls are color-coded, filled with three different colors, and the background is white. For both two and three balls, we generate 5 graphs and 10000 samples from each graph under each environment.

Candidate encoder and training. For learning the latent variables, we construct a two-step autoencoder, depicted in Figure 5. Specifically, in step one, we train an autoencoder on the observational image dataset with a bottleneck dimension of 64 using only a reconstruction objective. In step two, we train another autoencoder on the 64-dimensional output of the first encoder with the bottleneck dimension of n (the latent dimension) using both reconstruction loss and the score-based loss described earlier in Section 7.2. Architecture details and training
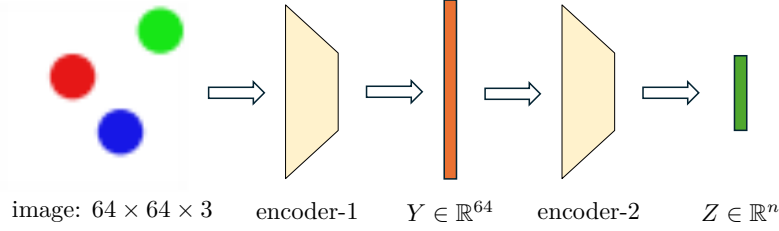
Figure 5: Learning the latent variables from images. Autoencoder-1 is trained to minimize only the reconstruction loss. Autoencoder-2 is trained to minimize the score-based loss in (83) while using the reconstruction loss as regularization.
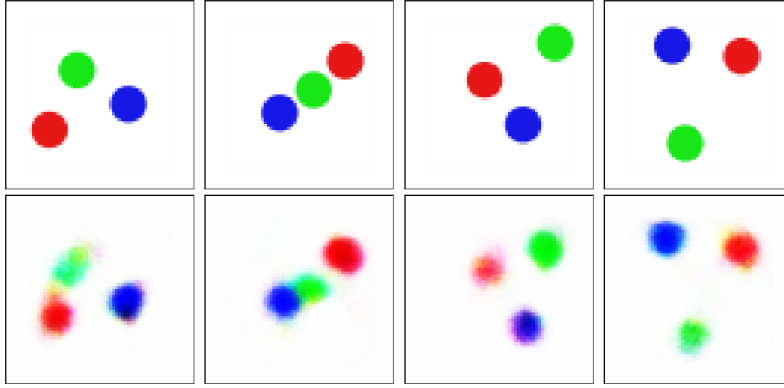

Figure 6: Sample images (top row) versus their reconstructions (bottom row).

schedule are given in Table 17. Finally, we repeat the training of the second autoencoder 5 times for each dataset.

Score difference estimation. The main loss function for the second autoencoder of our setup is $\|\mathbf{D}_t(h) - \mathbf{I}_{n \times n}\|_{1,1}$, which is equivalent to the score-based loss in (83). For computing $\mathbf{D}_t(h)$ for an encoder h, we need to compute the score differences of image datasets. To this end, we adopt a binary classifier-based log density ratio (LDR) estimator and use the gradients of these learned LDRs as our score difference functions, as described earlier. Specifically, we parameterize the LDR at any image through a CNN-based model, details of which are given in Table 18. The output of this model is used to compute class probabilities, which are used to minimize cross-entropy to train the model. We train one LDR model for each pair of hard interventions on the same node and one for the observational-interventional environment pairs for one set of hard interventions.

Results. GSCALE-I ensures perfect graph recovery and latent variables recovery up to permutation and element-wise scaling. Similarly to related work (Ahuja et al., 2023; Buchholz et al., 2023), we report MCC between $\hat{\mathbf{Z}}$ and $\mathbf{Z}$ as a metric of latent variables recovery, shown in Table 14. During training, we observed that MCC highly correlates with the reconstruction performance, i.e., when reconstruction errors in both autoencoders converge to small values, latent variable recovery is also successful. We note that the reconstruction does not have

47

Table 14: MCC comparison in image experiments (over 5 runs).

Algorithm	SCM	# balls	# int. / node	int. type	mean (std. error)
GSCALE-I	linear	2	2	hard	0.80 ± 0.03
GSCALE-I	linear	3	2	hard	0.76 ± 0.08
GSCALE-I	nonlinear	2	2	hard	0.93 ± 0.02
GSCALE-I	linear	2	1	hard	0.79 ± 0.03
GSCALE-I	nonlinear	2	1	hard	0.92 ± 0.02
Ahuja et al. (2023)	linear	2	1	do	0.13 ± 0.03
Ahuja et al. (2023)	linear	2	3	do	0.73 ± 0.03
Ahuja et al. (2023)	linear	2	5	do	0.83 ± 0.03
Buchholz et al. (2023)	linear	2	1	hard	0.87 ± 0.03
Buchholz et al. (2023)	linear	5	1	hard	0.94 ± 0.01

to be perfect. For instance, the samples in Figure 6 are from a run with 3 balls with an MCC of 0.98. Conversely, for the runs with non-convergent reconstruction loss in either of the autoencoder training steps, latent variable recovery also struggled. Hence, as long as we can ensure reasonable reconstruction performance when we are minimizing the score-based loss, our algorithm manages to disentangle the true causal variables embedded in the latent domain. We report the results from related work in Table 14 and discuss them as follows. Overall, we achieve a better performance than that of Ahuja et al. (2023), and slightly weaker than that of Buchholz et al. (2023) for linear SCMs. This can perhaps be explained by the versatility of our approach, which works for general SCMs, whereas the algorithm of Buchholz et al. (2023) is designed to work exclusively for linear SCMs, both theoretically and due to its loss function design. We demonstrate in Table 14 that our algorithm performs even better when the latent SCM is nonlinear (specifically, quadratic) than when it is linear, thereby verifying its versatility. Finally, we also test our algorithm when using one intervention per node, using the score differences between interventional and observational environments instead of two interventional environments. While two interventions per node are required for theoretical guarantees, our experiments show that using one intervention per node generally leads to competitive performance, as illustrated in Table 14.

Gap between the theory and practice in CRL. We highlight that the nature of the results presented in this paper is theoretical. We have established unknown identifiability guarantees in several settings. Furthermore, we have designed algorithms by defining a differentiable loss function whose global optima achieve identifiability for general CRL and established that the algorithms generate provably correct representations. The experiments presented on various synthesized and real datasets demonstrate the potential of the score-based framework for complex nonlinear transforms, such as image rendering. We achieve this goal by showing reasonable performance in latent variable recovery. That being said, the considered image dataset and other synthetic datasets are still relatively simple compared to real-world problems, and a gap remains between theoretical guarantees and practical applications. In other words, the theory developed is only a *necessary* step for guaranteed practical performance. From our perspective, closing this gap and obtaining *sufficient*

conditions for practical performance requires improvements in two aspects. First, using better score difference estimators (which we empirically investigate in Section 7.6) would enable a more effective application of our score-based CRL framework. Second, designing more suitable architectures for leveraging the invariance properties of causal models can help scale up CRL applications.

7.6 Sensitivity to Score Estimation Noise

As discussed earlier, score estimators can be modularly incorporated into our algorithm. For cases where $p_{\mathbf{X}}$ is not amenable to parameter estimation, we can use the noisy score estimates generated by SSM-VR as achievable baselines, as presented in this section. In this subsection, we evaluate the potential performance improvement that can be achieved as the score estimates become more accurate. To this end, we test the LSCALE-I algorithm under varying levels of score estimation noise. Specifically, we run the LSCALE-I algorithm using the scores generated by the following model:

$$\hat{s}_{\mathbf{X}}(\mathbf{x}; \sigma^2) = s_{\mathbf{X}}(\mathbf{x}) \cdot \left(1 + \Xi\right) , \quad \text{where} \quad \Xi \sim \mathcal{N}(0, \sigma^2 \cdot \mathbf{I}_{d \times d}) . \tag{112}$$

We consider hard interventions on graph size of $n = 5$, set $d = 25$, and vary the value of σ^2 within $[10^{-4}, 10^{-2}]$. We repeat the experiments 100 times with $n_{\mathrm{s}} = 10000$ samples from each environment. We plot mean normalized $\mathbf{Z}$ error, defined as

$$\ell_{\mathrm{norm}}(\mathbf{Z}, \hat{\mathbf{Z}}) \triangleq \frac{\|\mathbf{Z} - \hat{\mathbf{Z}}\|_2}{\|\mathbf{Z}\|_2} . \tag{113}$$

and mean SHD in Figure 7 with respect to the signal-to-noise ratio (SNR). It is clear that when the score estimation error is small, indicated by a high SNR value, the LSCALE-I algorithm demonstrates a strong performance in recovering both latent causal variables and the latent graph. We also note that the baseline results with noisy scores computed via SSM-VR in Table 5 yield similar success at graph recovery at the SNR of approximately 25 dB and latent recovery at the SRM of approximately 8 dB. The curves also confirm our observations in Section 7.1 and Section 7.2 that graph recovery is more sensitive to score estimation errors. The trend of the curves in Figure 7 indicates that our algorithm would greatly benefit from a better score estimator.

8. Discussion and Concluding Remarks

In this paper, we have proposed a score function-based CRL framework that uses stochastic interventions to learn latent causal representations and the latent causal graph underlying them. In this framework, by uncovering novel connections between score functions and CRL, we have established identifiability results for linear and general transformations without restricting the latent causal models, and designed LSCALE-I and GSCALE-I algorithms that achieve these identifiability guarantees. There are several exciting directions for future work.

As discussed in Section 3.4 after defining the identifiability objectives, and emphasized when presenting the corresponding results through the paper, our results are tight *given a complete set of atomic interventions.* A natural direction for future work is relaxing the atomic intervention requirement. In this aspect, the paper (Varıcı et al., 2024) extends the

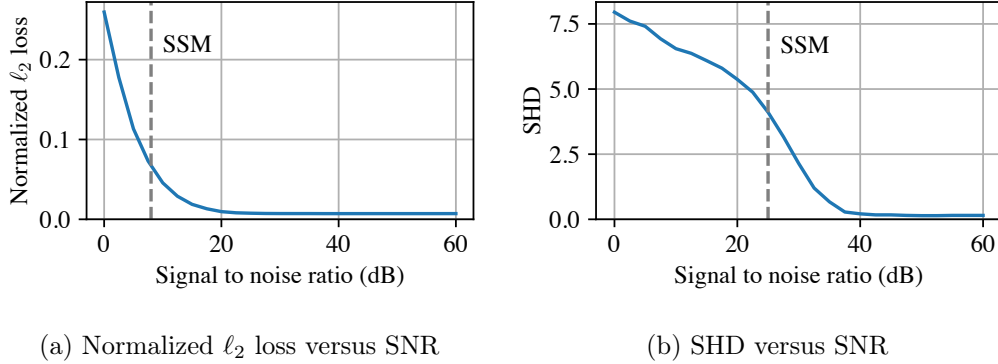(a) Normalized ℓ_2 loss versus SNR (b) SHD versus SNR

Figure 7: The performance of LSCALE-I under noisy scores with varying SNR for $n = 5$ and $d = 25$. The dashed vertical lines correspond to SNR values that correspond to the performance attained by SSM-VR.

score-based framework for linear transformations into the setting of unknown multi-node interventions by using combinations of multi-node score differences. Given a sufficiently diverse multi-node intervention set, similar identifiability results to those in Section 5 are shown. That being said, establishing the *necessary* conditions for multi-node interventions and the case of general transformations remains an open problem. Next, we note that our LSCALE-I algorithm for *linear transformations* is agnostic to the intervention type and latent causal model, meaning that it can be used for both soft and hard interventions and different causal models. Designing similar universal CRL algorithms that can handle general transformations under different sizes and types of interventional environments is the ultimate goal for studying the identifiability of CRL from interventions. A missing component of existing CRL literature is the finite-sample analysis. Probabilistic identifiability results for a given number of interventional data samples can be useful, especially in applications where performing interventions is costly. In this direction, the recent paper (Acartürk et al., 2024) establishes finite sample guarantees for CRL under linear transformations. Extending such analysis to general transformations can provide insights for bridging the gap between theory and practice in CRL.

Acknowledgments

Appendix

Table of Contents

Table 15: Notation.

	$[n]$	:	$\{1, \ldots, n\}$
	$\mathbb{1}$	:	indicator function
ground-truth	$\mathcal{G}$	:	latent causal graph over $\mathbf{Z}$
variables	$\mathcal{G}_{\text{tc}}$	:	transitive closure of $\mathcal{G}$
	$\mathcal{G}_{\text{tr}}$	:	transitive reduction of $\mathcal{G}$
	$\text{pa}(i)$	:	parents of node i in $\mathcal{G}$
	$\text{ch}(i)$	:	children of node i in $\mathcal{G}$
	$\text{an}(i)$	:	ancestors of node i in $\mathcal{G}$
	$\text{de}(i)$	:	descendants of node i in $\mathcal{G}$
	$\mathbf{X}$	:	$[X_1, \ldots, X_d]^\top$ observed random variables
	$\mathbf{Z}$	:	$[Z_1, \ldots, Z_n]^\top$ latent random variables
	$\mathbf{Z}_{\text{pa}(i)}$	:	vector formed by Z_i for all $i \in \text{pa}(i)$
	g	:	true decoder
	h	:	a valid encoder
	$\mathcal{H}$	:	the set of valid encoders h
	$\mathbf{G}$	:	true linear decoder
	$\mathbf{H}$	:	a valid linear encoder
intervention	$\mathcal{E}^0$	:	observational environment
notations	$\mathcal{E}$	:	$(\mathcal{E}^1, \ldots, \mathcal{E}^n)$ first interventional environments
	$\tilde{\mathcal{E}}$	:	$(\tilde{\mathcal{E}}^1, \ldots, \tilde{\mathcal{E}}^n)$ second interventional environments
	I^m	:	the intervened node(s) in $\mathcal{E}^m$
	$\tilde{I}^m$	:	the intervened node(s) in $\tilde{\mathcal{E}}^m$
	$\mathcal{I}$	:	the set of intervened nodes $(I^1, \ldots, I^n)$
	$\tilde{\mathcal{I}}$	:	the set of intervened nodes $(\tilde{I}^1, \ldots, \tilde{I}^n)$
statistical	$\hat{\mathbf{Z}}(\mathbf{X})$	:	generic estimator of $\mathbf{Z}$ given $\mathbf{X}$
models	$\hat{\mathbf{Z}}(\mathbf{X}; h)$	:	an auxiliary estimator of $\mathbf{Z}$ given $\mathbf{X}$ and encoder h
	$\hat{\mathcal{G}}$	:	estimate of $\mathcal{G}$
	$p, p^m, \tilde{p}^m$	:	pdfs of $\mathbf{Z}$ in $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$
	$p_{\mathbf{X}}, p_{\mathbf{X}}^m, \tilde{p}_{\mathbf{X}}^m$	:	pdfs of $\mathbf{X}$ in $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$
	$\boldsymbol{s}, \boldsymbol{s}^m, \tilde{\boldsymbol{s}}^m$	:	score functions of $\mathbf{Z}$ in $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$
	$\boldsymbol{s}_{\mathbf{X}}, \boldsymbol{s}_{\mathbf{X}}^m, \tilde{\boldsymbol{s}}_{\mathbf{X}}^m$	:	score functions of $\mathbf{X}$ in $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$
	$\boldsymbol{s}_{\hat{\mathbf{Z}}}, \boldsymbol{s}_{\hat{\mathbf{Z}}}^m, \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^m$	:	score functions of $\hat{\mathbf{Z}}$ in $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$ for encoder h
	$\hat{\text{pa}}(i)$	:	parents of node i in $\hat{\mathcal{G}}$
matrix	$\mathbf{A}^\dagger$	:	Pseudo-inverse of matrix $\mathbf{A}$
notations	$\mathbf{A}_i$	:	row i of matrix $\mathbf{A}$
	$\mathbf{A}_{i,j}$	:	entry of matrix $\mathbf{A}$ at row i and column j
	$\mathbf{P}_\pi$	:	Permutation matrix associated with permutation π of $[n]$
	$\mathbf{D}, \tilde{\mathbf{D}}, \mathbf{D}_{\text{t}}$	:	True score change matrices
	$\mathbf{D}(h), \tilde{\mathbf{D}}(h), \mathbf{D}_{\text{t}}(h)$	:	Score change matrices under encoder h

Appendix A. Proofs of Score Function Properties and Transformations

We start by providing the following facts that will be used repeatedly in the proofs.

Proposition 31 *Consider two continuous functions $f, g : \mathbb{R}^n \to \mathbb{R}$ with full support. Then, for any $\alpha > 0$,*

$$\exists \mathbf{z} \in \mathbb{R}^n \ \ f(\mathbf{z}) \neq g(\mathbf{z}) \quad \Longleftrightarrow \quad \mathbb{E}\Big[\big|f(\mathbf{Z}) - g(\mathbf{Z})\big|^\alpha\Big] \neq 0 \ . \tag{114}$$

Specifically, for $\alpha = 1$, we have

$$\exists \mathbf{z} \in \mathbb{R}^n \ \ f(\mathbf{z}) \neq g(\mathbf{z}) \quad \Longleftrightarrow \quad \mathbb{E}\Big[\big|f(\mathbf{Z}) - g(\mathbf{Z})\big|\Big] \neq 0 \ . \tag{115}$$

Proof: If there exists $\mathbf{z} \in \mathbb{R}^n$ such that $f(\mathbf{z}) \neq g(\mathbf{z})$, then $f(\mathbf{z}) - g(\mathbf{z})$ is nonzero over a nonzero-measure set due to continuity. Then, $\mathbb{E}\big[|f(\mathbf{Z}) - g(\mathbf{Z})|^\alpha\big] \neq 0$ since p (pdf of $\mathbf{Z}$) has full support. On the other direction, if $f(\mathbf{z}) = g(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^n$, then $\mathbb{E}\big[|f(\mathbf{Z}) - g(\mathbf{Z})|^\alpha\big] = 0$. This means that $\mathbb{E}\big[|f(\mathbf{Z}) - g(\mathbf{Z})|^\alpha\big] \neq 0$ implies that there exists $\mathbf{z} \in \mathbb{R}^n$ such that $f(\mathbf{z}) \neq g(\mathbf{z})$.

A.1 Proof of Lemma 7

Our score-based methodology builds on the changes in score functions under interventions. For proving Theorem 7, we start by showing that $\mathbb{E}\Big[\big|\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^m(\mathbf{Z})|_i\big|\Big] \neq 0 \implies i \in \overline{\mathrm{pa}}(I^m)$, which holds true regardless of the causal model and the intervention type.

Proof of $\mathbb{E}\Big[\big|\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^m(\mathbf{Z})|_i\big|\Big] \neq 0 \implies i \in \overline{\mathrm{pa}}(I^m)$: Let ℓ denote the node intervened in $\mathcal{E}^m$, i.e., $I^m = \ell$. Recalling (18) and (19), the latent scores $\boldsymbol{s}(\mathbf{z})$ and $\boldsymbol{s}^m(\mathbf{z})$ are decomposed as

$$\boldsymbol{s}(\mathbf{z}) = \nabla_{\mathbf{z}} \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) + \sum_{i \neq \ell} \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \ , \tag{116}$$

$$\text{and} \quad \boldsymbol{s}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) + \sum_{i \neq \ell} \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \ . \tag{117}$$

Hence, $\boldsymbol{s}(\mathbf{z})$ and $\boldsymbol{s}^m(\mathbf{z})$ differ in only the causal mechanism of node ℓ. Next, we check the derivatives of $p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})$ and $q_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})$ in their i-th coordinates. Note that these two depend on $\mathbf{Z}$ only through $\{Z_j : j \in \overline{\mathrm{pa}}(\ell)\}$. Therefore, if $i \notin \overline{\mathrm{pa}}(\ell)$,

$$\frac{\partial}{\partial z_i} \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) = \frac{\partial}{\partial z_i} \log q_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) = 0 \ , \tag{118}$$

which indicates that if $i \notin \overline{\mathrm{pa}}(\ell)$, then $[\boldsymbol{s}(\mathbf{z})]_i = [\boldsymbol{s}^m(\mathbf{z})]_i$ for all $\mathbf{z}$. This, equivalently, means that if $\mathbb{E}\big[|\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^m(\mathbf{Z})|_i\big] \neq 0$, then $i \in \overline{\mathrm{pa}}(\ell)$. $\qquad \square$

For the reverse direction, we will use the following intermediate result which formalizes the weakest possible requirement for a meaningful intervention and shows that it is a property of (i) hard interventions under any causal model, and (ii) additive noise model under either soft or hard interventions.

Lemma 32 (Interventional Regularity) *Causal mechanisms p_i and q_i of node i are said to satisfy* interventional regularity *if*

$$\exists \mathbf{z} \in \mathbb{R}^n \ \ such\ that \ \ \frac{\partial}{\partial z_k} \frac{q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} \neq 0 \ , \qquad \forall k \in \mathrm{pa}(i) \ . \tag{119}$$

Then, p_i and q_i satisfy interventional regularity if at least one of the following conditions is true:

1. *The intervention is hard, i.e., $q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = q_i(z_i)$.*

2. *The causal model is an additive noise model in which the pdfs of the noise variables are analytic.*

Proof: See Appendix A.3.

Case (i) and Case (ii). $\mathbb{E}\!\left[\left|s(\mathbf{Z}) - s^m(\mathbf{Z})\right|_i\right] \neq 0 \implies i \in \overline{\mathrm{pa}}(I^m)$ is already shown above. For the reverse direction, we give the proof for soft interventions on additive noise models, Case (ii). We will use interventional regularity since Theorem 32 shows that it is satisfied for additive noise models. The proof for hard interventions, Case (i), follows from similar arguments since interventional regularity is also satisfied for hard interventions by Theorem 32.

Proof of $i \in \overline{\mathrm{pa}}(I^m) \implies \mathbb{E}\!\left[\left|s(\mathbf{Z}) - s^m(\mathbf{Z})\right|_i\right] \neq 0$: Note that the two score functions s and s^m are equal in their coordinate $i \in \overline{\mathrm{pa}}(\ell)$ only if

$$0 = \frac{\partial \log q_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_i} - \frac{\partial \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_i} = \frac{\partial}{\partial z_i} \log \frac{q_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(i)})}{p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})} \ . \tag{120}$$

However, (120) contradicts with interventional regularity. Therefore, if $i \in \overline{\mathrm{pa}}(\ell)$, $[s(\mathbf{z})_i]$ and $[s^m(\mathbf{z})]_i$ are not identical and by Theorem 31, $\mathbb{E}\!\left[\left|s(\mathbf{Z}) - s^m(\mathbf{Z})\right|_i\right] \neq 0$.

Case (iii) Coupled environments. Suppose that $I^m = \tilde{I}^m = \ell$. Following (16), we have

$$s^m(\mathbf{z}) = \nabla_\mathbf{z} \log q_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_\mathbf{z} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \ , \tag{121}$$

$$\text{and} \quad \tilde{s}^m(\mathbf{z}) = \nabla_\mathbf{z} \log \tilde{q}_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_\mathbf{z} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \ . \tag{122}$$

Then, subtracting (122) from (121) and looking at i-th coordinate, we have

$$\left[s^m(\mathbf{z}) - \tilde{s}^m(\mathbf{z})\right]_i = \frac{\partial \log q_\ell(z_\ell)}{\partial z_i} - \frac{\partial \log \tilde{q}_\ell(z_\ell)}{\partial z_i} \ . \tag{123}$$

If $i \neq \ell$, the right-hand side is zero and we have $\left[s^m(\mathbf{z}) - \tilde{s}^m(\mathbf{z})\right]_i = 0$ for all $\mathbf{z}$. On the other hand, if $i = \ell$, since $q_\ell(z_\ell)$ and $\tilde{q}_\ell(z_\ell)$ are distinct, there exists $\mathbf{z} \in \mathbb{R}^n$ such that $q_\ell(z_\ell) \neq \tilde{q}_\ell(z_\ell)$. Subsequently, by Theorem 31, we have $\mathbb{E}\!\left[\left|s^m(\mathbf{Z}) - \tilde{s}^m(\mathbf{Z})\right|_i\right] \neq 0$.

Case (iv) Uncoupled environments. Suppose that $I^m = \ell$ and $\tilde{I}^m = j$, and $\ell \neq j$. Following (16), we have

$$\boldsymbol{s}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_\ell(z_\ell) + \nabla_{\mathbf{z}} \log p_j(z_j \mid \mathbf{z}_{\mathrm{pa}(j)}) + \sum_{k \in [n] \setminus \{\ell, j\}} \nabla_{\mathbf{z}} \log p_k(z_k \mid \mathbf{z}_{\mathrm{pa}(k)}) \;, \quad (124)$$

$$\text{and} \quad \tilde{\boldsymbol{s}}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_j(z_j) + \nabla_{\mathbf{z}} \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) + \sum_{k \in [n] \setminus \{\ell, j\}} \nabla_{\mathbf{z}} \log p_k(z_k \mid \mathbf{z}_{\mathrm{pa}(k)}) \;. \quad (125)$$

Then, subtracting (125) from (124) we have

$$\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_\ell(z_\ell) + \nabla_{\mathbf{z}} \log p_j(z_j \mid \mathbf{z}_{\mathrm{pa}(j)}) - \nabla_{\mathbf{z}} \log q_j(z_j) - \nabla_{\mathbf{z}} \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) \;. \quad (126)$$

Scrutinizing the i-th coordinate, we have

$$\left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z}) \right]_i = \frac{\partial \log q_\ell(z_\ell)}{\partial z_i} + \frac{\partial \log p_j(z_j \mid \mathbf{z}_{\mathrm{pa}(j)})}{\partial z_i} - \frac{\partial \log q_j(z_j)}{\partial z_i} - \frac{\partial \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_i} \;. \quad (127)$$

Proof of $\mathbb{E}\left[\left| \boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\right|_i\right] \neq 0 \implies i \in \overline{\mathrm{pa}}(\ell, j)$: Suppose that $i \notin \overline{\mathrm{pa}}(\ell, j)$. Then, none of the terms in the RHS of (127) is a function of z_i. Therefore, all the terms in the RHS of (127) are zero, and we have $\left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_i = 0$ for all $\mathbf{z}$. By Theorem 31, $\mathbb{E}\left[\left| \boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\right|_i\right] = 0$. This, equivalently, means that if $\mathbb{E}\left[\left| \boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\right|_i\right] \neq 0$, then $i \in \overline{\mathrm{pa}}(\ell, j)$.

Proof of $\mathbb{E}\left[\left| \boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})\right|_i\right] \neq 0 \impliedby i \in \overline{\mathrm{pa}}(\ell, j)$: We prove it by contradiction. Assume that $\left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_i = 0$ for all $\mathbf{z}$. Without loss of generality, let $\ell \notin \overline{\mathrm{pa}}(j)$.

1. **If** $i = \ell$. In this case, (127) is simplified to

$$0 = \left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_\ell = \frac{\partial \log q_\ell(z_\ell)}{\partial z_\ell} - \frac{\partial \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_\ell} \;. \quad (128)$$

If ℓ is a root node, i.e., $\mathrm{pa}(\ell) = \emptyset$, (128) implies that $(\log q_\ell)'(z_\ell) = (\log p_\ell)'(z_\ell)$ for all z_ℓ. Integrating, we get $p_\ell(z_\ell) = \alpha q_\ell(z_\ell)$ for some constant α. Since both p_ℓ and q_ℓ are pdfs, they both integrate to one, implying $\alpha = 1$ and $p_\ell(z_\ell) = q_\ell(z_\ell)$, which contradicts the premise that observational and interventional mechanisms are distinct. If ℓ is not a root node, consider some $k \in \mathrm{pa}(\ell)$. Then, taking the derivative of (128) with respect to z_k, we have

$$0 = \frac{\partial^2 \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_\ell \partial z_k} \;. \quad (129)$$

Recall the equation $Z_\ell = f_\ell(\mathbf{Z}_{\mathrm{pa}(\ell)}) + N_\ell$ for additive noise models specified in (7). Denote the pdf of the noise term N_ℓ by p_N. Then, the conditional pdf $p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})$ is given by $p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)}) = p_N(z_\ell - f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)}))$. Denoting the score function of p_N by r_p,

$$r_p(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log p_N(u) \;, \quad (130)$$

we have

$$\frac{\partial \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_\ell} = \frac{\partial \log p_N(z_\ell - f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)}))}{\partial z_\ell} = r_p(z_\ell - f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)})) \ . \tag{131}$$

Substituting this into (129), we obtain

$$0 = \frac{\partial r_p(z_\ell - f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)}))}{\partial z_k} = -\frac{\partial f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_k} \cdot r_p'(z_\ell - f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)})) \ , \quad \forall \mathbf{z} \in \mathbb{R}^n \ . \tag{132}$$

Since k is a parent of ℓ, there exists a fixed $\mathbf{Z}_{\mathrm{pa}(\ell)} = \mathbf{z}^*_{\mathrm{pa}(\ell)}$ realization for which $\partial f_\ell(\mathbf{z}^*_{\mathrm{pa}(\ell)})/\partial z_k$ is nonzero. Otherwise, $f_\ell(\mathbf{z}_{\mathrm{pa}(\ell)})$ would not be sensitive to z_k which is contradictory to k being a parent of ℓ. Note that Z_ℓ can vary freely after fixing $\mathbf{Z}_{\mathrm{pa}(\ell)}$. Therefore, for (132) to hold, the derivative of r_p must always be zero. However, the score function of a valid pdf with full support cannot be constant. Therefore, $\left[\boldsymbol{s}^m(\mathbf{z})_i - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_i$ is not always zero, and we have $\mathbb{E}\left[|\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})|_i\right] \neq 0$.

2. **If $i \neq \ell$.** In this case, (127) is simplified to

$$0 = \left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_i = \frac{\partial \log p_j(z_j \mid \mathbf{z}_{\mathrm{pa}(j)})}{\partial z_i} - \frac{\partial \log q_j(z_j)}{\partial z_i} - \frac{\partial \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_i} \ . \tag{133}$$

We investigate case by case and reach a contradiction for each case. First, suppose that $i \notin \mathrm{pa}(\ell)$. Then, we have $i \in \overline{\mathrm{pa}}(j)$, and (133) becomes

$$0 = \left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_i = \frac{\partial \log p_j(z_j \mid \mathbf{z}_{\mathrm{pa}(j)})}{\partial z_i} - \frac{\partial \log q_j(z_j)}{\partial z_i} \ . \tag{134}$$

If $i = j$, the impossibility of (134) directly follows from the impossibility of (128). The remaining case is $i \in \mathrm{pa}(j)$. In this case, taking the derivative of the right-hand side of (134) with respect to z_j, we obtain

$$0 = \frac{\partial^2 \log p_j(z_j \mid \mathbf{z}_{\mathrm{pa}(j)})}{\partial z_i \partial z_j} \ , \tag{135}$$

which is a realization of (129) for $i \in \mathrm{pa}(j)$ and j in place of $k \in \mathrm{pa}(\ell)$ and ℓ, which we proved to be impossible in $i = \ell$ case. Therefore, $i \notin \mathrm{pa}(\ell)$ is not viable. Finally, suppose that $i \in \mathrm{pa}(\ell)$. Then, taking the derivative of the right-hand side of (133) with respect to z_ℓ, we obtain

$$0 = \frac{\partial^2 \log p_\ell(z_\ell \mid \mathbf{z}_{\mathrm{pa}(\ell)})}{\partial z_i \partial z_\ell} \ , \tag{136}$$

which is again a realization of (129) for $k = i$, which we proved to be impossible.

Hence, we showed that $\left[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^m(\mathbf{z})\right]_i$ cannot be zero for all $\mathbf{z}$ values. Then, by Theorem 31 we have $\mathbb{E}\left[|\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^m(\mathbf{Z})|_i\right] \neq 0$, and the proof is concluded.

A.2 Proof of Lemma 8

Let us recall the setting. Consider random vectors $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}^r$ and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^s$ that are related through $\mathbf{Y}_1 = f(\mathbf{W}_1)$ and $\mathbf{Y}_2 = f(\mathbf{W}_2)$ such that $r \geq s$, probability measures of $\mathbf{W}_1, \mathbf{W}_2$ are absolutely continuous with respect to the s-dimensional Lebesgue measure and $f : \mathbb{R}^s \to \mathbb{R}^r$ is an injective and continuously differentiable function.

In this setting, the realizations of $\mathbf{W}_1$ and $\mathbf{Y}_1$, and that of $\mathbf{W}_2$ and $\mathbf{Y}_2$, are related through $\mathbf{y} = f(\mathbf{w})$. Since f is injective and continuously differentiable, volume element $d\mathbf{w}$ in $\mathbb{R}^s$ gets mapped to $\left|\det\left([J_f(\mathbf{w})]^\top \cdot J_f(\mathbf{w})\right)\right|^{1/2} d\mathbf{w}$ on $\mathrm{im}(f)$. Since $\mathbf{W}_1$ has density $p_{\mathbf{W}_1}$ absolutely continuous with respect to the s-dimensional Lebesgue measure, using the area formula (Boothby, 2003), we can define a density for $\mathbf{Y}_1$, denoted by $p_{\mathbf{Y}_1}$, supported only on manifold $\mathcal{M} \triangleq \mathrm{im}(f)$ which is absolutely continuous with respect to the s-dimensional Hausdorff measure:

$$p_{\mathbf{Y}_1}(\mathbf{y}) = p_{\mathbf{W}_1}(\mathbf{w}) \cdot \left|\det\left([J_f(\mathbf{w})]^\top \cdot J_f(\mathbf{w})\right)\right|^{-1/2} , \quad \text{where} \quad \mathbf{y} = f(\mathbf{w}) . \tag{137}$$

Densities $p_{\mathbf{Y}_2}$ and $p_{\mathbf{W}_2}$ of $\mathbf{Y}_2$ and $\mathbf{W}_2$ are related similarly. Subsequently, score functions of $\{\mathbf{W}_1, \mathbf{W}_2\}$ and $\{\mathbf{Y}_1, \mathbf{Y}_2\}$ are specified similarly to (9) and (14), respectively. Denote the Jacobian matrix of f at point $\mathbf{w} \in \mathbb{R}^s$ by $J_f(\mathbf{w})$, which is an $r \times s$ matrix with entries given by

$$[J_f(\mathbf{w})]_{i,j} = \frac{\partial [f(\mathbf{w})]_i}{\partial w_j} = \frac{\partial y_i}{\partial w_j} , \quad \forall i \in [r] , j \in [s] . \tag{138}$$

Next, consider a function $\phi \colon \mathcal{M} \to \mathbb{R}$. Since the domain of ϕ is a manifold, its differential, denoted by $D\phi$, is defined according to (12). By noting $\mathbf{y} = f(\mathbf{w})$, we can also differentiate ϕ with respect to $\mathbf{w} \in \mathbb{R}^s$ as (Simon, 2014, p. 57)

$$\nabla_w \phi(\mathbf{y}) = \nabla_w (\phi \circ f)(\mathbf{w}) = [J_f(\mathbf{w})]^\top \cdot D\phi(\mathbf{y}) . \tag{139}$$

Next, given the identities in (137) and (139), we find the relationship between score functions of $\mathbf{W}_1$ and $\mathbf{Y}_1$ as follows.

$$s_{\mathbf{W}_1}(\mathbf{w}) = \nabla_{\mathbf{w}} \log p_{\mathbf{W}_1}(\mathbf{w}) \tag{140}$$

$$\stackrel{(137)}{=} \nabla_{\mathbf{w}} \log p_{\mathbf{Y}_1}(\mathbf{y}) + \nabla_w \log \left|\det\left([J_f(\mathbf{w})]^\top \cdot J_f(\mathbf{w})\right)\right|^{1/2} \tag{141}$$

$$\stackrel{(139)}{=} [J_f(\mathbf{w})]^\top \cdot D \log p_{\mathbf{Y}_1}(\mathbf{y}) + \nabla_{\mathbf{w}} \log \left|\det\left([J_f(\mathbf{w})]^\top \cdot J_f(\mathbf{w})\right)\right|^{1/2} \tag{142}$$

$$= [J_f(\mathbf{w})]^\top \cdot s_{\mathbf{Y}_1}(\mathbf{y}) + \nabla_w \log \left|\det\left([J_f(\mathbf{w})]^\top \cdot J_f(\mathbf{w})\right)\right|^{1/2} . \tag{143}$$

Following the similar steps that led to (143) for $\mathbf{W}_2$ and $\mathbf{Y}_2$, we obtain

$$s_{\mathbf{W}_2}(\mathbf{w}) = [J_f(\mathbf{w})]^\top \cdot s_{\mathbf{Y}_2}(\mathbf{y}) + \nabla_{\mathbf{w}} \log \left|\det\left([J_f(\mathbf{w})]^\top \cdot J_f(\mathbf{w})\right)\right|^{1/2} . \tag{144}$$

Subtracting (144) from (143), we obtain the desired result

$$s_{\mathbf{W}_1}(\mathbf{w}) - s_{\mathbf{W}_2}(\mathbf{w}) = [J_f(\mathbf{w})]^\top \cdot [s_{\mathbf{Y}_1}(\mathbf{y}) - s_{\mathbf{Y}_2}(\mathbf{y})] . \tag{145}$$

Proof of the reverse direction. Multiplying (145) from left with $\left[[J_f(\mathbf{w})]^\dagger\right]^\top$, we obtain

$$\left[[J_f(\mathbf{w})]^\dagger\right]^\top \cdot \left[\boldsymbol{s}_{\mathbf{W}_1}(\mathbf{w}) - \boldsymbol{s}_{\mathbf{W}_2}(\mathbf{w})\right] = \left[[J_f(\mathbf{w})]^\dagger\right]^\top \cdot [J_f(\mathbf{w})]^\top \cdot \left[\boldsymbol{s}_{\mathbf{Y}_1}(\mathbf{y}) - \boldsymbol{s}_{\mathbf{Y}_2}(\mathbf{y})\right] . \quad (146)$$

Note that

$$\left[[J_f(\mathbf{w})]^\dagger\right]^\top \cdot [J_f(\mathbf{w})]^\top = J_f(\mathbf{w}) \cdot [J_f(\mathbf{w})]^\dagger . \quad (147)$$

By properties of the Moore-Penrose inverse, for any matrix $\mathbf{A}$, we have $\mathbf{A} \cdot \mathbf{A}^\dagger \cdot \mathbf{A} = \mathbf{A}$. This means that $\mathbf{A} \cdot \mathbf{A}^\dagger$ acts as a left identity for vectors in the column space of $\mathbf{A}$. By definition, $\boldsymbol{s}_{\mathbf{Y}_1}$ and $\boldsymbol{s}_{\mathbf{Y}_2}$ have values in $T_{\mathbf{w}}\mathrm{im}(f)$, the tangent space of the image manifold f at point $\mathbf{w}$. This space is equal to the column space of the matrix $J_f(\mathbf{w})$. Therefore, $J_f(\mathbf{w}) \cdot [J_f(\mathbf{w})]^\dagger$ acts as a left identity for $\boldsymbol{s}_{\mathbf{Y}_1}(\mathbf{y})$ and $\boldsymbol{s}_{\mathbf{Y}_2}(\mathbf{y})$, and we have

$$J_f(\mathbf{w}) \cdot [J_f(\mathbf{w})]^\dagger \cdot \left[\boldsymbol{s}_{\mathbf{Y}_1}(\mathbf{y}) - \boldsymbol{s}_{\mathbf{Y}_2}(\mathbf{y})\right] = \boldsymbol{s}_{\mathbf{Y}_1}(\mathbf{y}) - \boldsymbol{s}_{\mathbf{Y}_2}(\mathbf{y}) . \quad (148)$$

Substituting (147) and (148) into (146) completes the proof.

Proof of Theorem 9 For a given linear transform $\mathbf{F}$, we have $J_f(\mathbf{w}) = \mathbf{F}$, which is independent of $\mathbf{w}$. Then, in the proof of Theorem 8, (143) reduces to $\boldsymbol{s}_{\mathbf{W}}(\mathbf{w}) = \mathbf{F}^\top \cdot \boldsymbol{s}_{\mathbf{Y}}(\mathbf{y})$. Finally, we note that the score difference of $\mathbf{Y}$ can be similarly written in terms of the score difference of $\mathbf{W}$.

A.3 Proof of Lemma 32

We will prove that causal mechanisms p_i and q_i satisfy interventional regularity for (i) hard interventions and (ii) additive noise models. To this end, we first define

$$\psi(z_i, \mathbf{z}_{\mathrm{pa}(i)}) \triangleq \frac{q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} . \quad (149)$$

We start by showing that $\psi(z_i, \mathbf{z}_{\mathrm{pa}(i)})$ varies with z_i. We prove it by contradiction. Assume the contrary, i.e., let $\psi(z_i, \mathbf{z}_{\mathrm{pa}(i)}) = \psi(\mathbf{z}_{\mathrm{pa}(i)})$. By rearranging (149) we have

$$q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = \psi(\mathbf{z}_{\mathrm{pa}(i)})\, p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) . \quad (150)$$

Fix a realization of $\mathbf{z}^*_{\mathrm{pa}(i)}$ and integrate both sides of (150) with respect to z_i. Since both p_i and q_i are pdfs, we have

$$1 = \int_{\mathbb{R}} q_i(z_i \mid \mathbf{z}^*_{\mathrm{pa}(i)})\, \mathrm{d}z_i = \int_{\mathbb{R}} \psi(\mathbf{z}^*_{\mathrm{pa}(i)})\, p_i(z_i \mid \mathbf{z}^*_{\mathrm{pa}(i)})\mathrm{d}z_i\mathrm{d}z_i \quad (151)$$

$$= \psi(\mathbf{z}^*_{\mathrm{pa}(i)}) \int_{\mathbb{R}} p_i(z_i \mid \mathbf{z}^*_{\mathrm{pa}(i)})\mathrm{d}z_i \quad (152)$$

$$= \psi(\mathbf{z}^*_{\mathrm{pa}(i)}) . \quad (153)$$

This identity implies that $p_i(z_i \mid \mathbf{z}^*_{\mathrm{pa}(i)}) = q_i(z_i \mid \mathbf{z}^*_{\mathrm{pa}(i)})$ for any arbitrary realization $\mathbf{z}^*_{\mathrm{pa}(i)}$. This contradicts the premise that observational and interventional distributions are distinct. As a result, to check if a model satisfies interventional regularity for node i, it suffices to

investigate whether the function ψ is not invariant with respect to z_k for $k \in \mathrm{pa}(i)$. To this end, from (149) we know that $\psi(z_i, \mathbf{z}_{\mathrm{pa}(i)})$ varies with z_k if and only if

$$\frac{\partial}{\partial z_i} \log \psi(z_i, \mathbf{z}_{\mathrm{pa}(i)}) = \frac{\partial q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_i} \cdot \frac{1}{q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} - \frac{\partial p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_i} \cdot \frac{1}{p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} \neq 0 \ . \tag{154}$$

Next, we investigate the sufficient conditions listed in Theorem 32.

A.3.1 Hard Interventions

Under hard interventions, note that for any $k \in \mathrm{pa}(i)$,

$$\frac{\partial}{\partial z_k} p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \neq 0 \ , \quad \text{and} \quad \frac{\partial}{\partial z_k} q_i(z_i) = 0 \ . \tag{155}$$

Then, it follows directly from (154) that

$$\frac{\partial}{\partial z_k} \log \psi(z_i, \mathbf{z}_{\mathrm{pa}(i)}) = \underbrace{\frac{\partial q_i(z_i)}{\partial z_k}}_{=0} \cdot \frac{1}{q_i(z_i)} - \underbrace{\frac{\partial p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}}_{\neq 0} \cdot \frac{1}{p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} \neq 0 \ . \tag{156}$$

Thus, hard interventions on any latent causal model satisfy interventional regularity.

A.3.2 Additive Noise Models

The additive noise model for node i is given by

$$Z_i = f_i(\mathbf{Z}_{\mathrm{pa}(i)}) + N_i \ , \tag{157}$$

as specified in (7). When node i is soft intervened, Z_i is generated according to

$$Z_i = \bar{f}_i(\mathbf{Z}_{\mathrm{pa}(i)}) + \bar{N}_i \ , \tag{158}$$

in which $\bar{f}_i$ and $\bar{N}_i$ specify the interventional mechanism for node i. Then, denoting the pdfs of N_i and $\bar{N}_i$ by p_N and q_N, respectively, (157) and (158) imply that

$$p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = p_N(z_i - f_i(\mathbf{z}_{\mathrm{pa}(i)})) \ , \quad \text{and} \quad q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = q_N(z_i - \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})) \ . \tag{159}$$

Denote the score functions associated with p_N and q_N by

$$r_p(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log p_N(u) = \frac{p'_N(u)}{p_N(u)} \ , \quad \text{and} \quad r_q(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log q_N(u) = \frac{q'_N(u)}{q_N(u)} \ . \tag{160}$$

We will prove that (154) holds by contradiction. Assume the contrary and let

$$\frac{\partial q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot \frac{1}{q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} = \frac{\partial p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot \frac{1}{p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})} \ . \tag{161}$$

From (159) and (160), for the numerators in (161) we have,

$$\frac{\partial p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} = -\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot p'_N\big(z_i - f_i(\mathbf{z}_{\mathrm{pa}(i)})\big) , \tag{162}$$

$$\text{and} \quad \frac{\partial q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} = -\frac{\partial \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot q'_N\big(z_i - \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})\big) . \tag{163}$$

Hence, the identity in (161) can be written as

$$\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot r_p\big(z_i - f_i(\mathbf{z}_{\mathrm{pa}(i)})\big) = \frac{\partial \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot r_q\big(z_i - \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})\big) . \tag{164}$$

Define n_i and $\bar{n}_i$ as the realizations of N_i and $\bar{N}_i$ when $Z_i = z_i$ and $\mathbf{Z}_{\mathrm{pa}(i)} = \mathbf{z}_{\mathrm{pa}(i)}$. By defining $\delta(\mathbf{z}_{\mathrm{pa}(i)}) \triangleq f_i(\mathbf{z}_{\mathrm{pa}(i)}) - \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})$, we have $\bar{n}_i = n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)})$. Then, (164) can rewritten as

$$\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot r_p(n_i) = \frac{\partial \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot r_q(n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)})) . \tag{165}$$

Note that $\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}$ is a nonzero continuous function. Hence, there exists an interval $\Phi \subseteq \mathbb{R}^{|\mathrm{pa}(i)|}$ over which $\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \neq 0$. Likewise, $r_p(n_i)$ cannot be constantly zero over all possible intervals $\Omega \subseteq \mathbb{R}$. This is because otherwise, it would have to necessarily be a constant zero function (since it is analytic), which is an invalid score function. Hence, there exists an open interval $\Omega \subseteq \mathbb{R}$ over which $r_p(n_i)$ is nonzero for all $n_i \in \Omega$. Then, we can rearrange (165) as

$$\frac{r_p(n_i)}{r_q(n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)}))} = \frac{\frac{\partial \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}}{\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}} , \quad \forall(n_i, \mathbf{z}_{\mathrm{pa}(i)}) \in \Omega \times \Phi . \tag{166}$$

Note that the RHS of (166) is not a function of n_i. Then, taking the derivative of both sides with respect to n_i, we get

$$\frac{r'_p(n_i)}{r_p(n_i)} = \frac{r'_q(n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)}))}{r_q(n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)}))} . \tag{167}$$

In the next step, we show that δ is not a constant function. We prove this by contradiction. Suppose that $\delta(\mathbf{z}_{\mathrm{pa}(i)}) = \delta^*$ is a constant function. Then, the gradients of f_i and $\bar{f}_i$ are equal. From (166), this implies that

$$r_p(n_i) = r_q(n_i + \delta^*) , \quad \forall n_i \in \Omega . \tag{168}$$

Since $r_p(n_i)$ and $r_q(n_i + \delta^*)$ are analytic functions that agree on an open interval of $\mathbb{R}$, they are equal for all $n_i \in \mathbb{R}$. This implies that $p_N(n_i) = \eta \cdot q_N(n_i + \delta^*)$ for some constant $\eta \in \mathbb{R}$. Since p_N and q_N are pdfs, $\eta = 1$ is the only choice that maintains p_N and q_N are pdfs. Therefore, $p_N(n_i) = q_N(n_i + \delta^*)$. However, using (159), this implies that $p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$, which contradicts the premise that an intervention changes the causal mechanism of target node i. Therefore, δ is a continuous, non-constant function, and its image over $\mathbf{z}_{\mathrm{pa}(i)} \in \Phi$ includes an open interval $\Theta \subseteq \mathbb{R}$. With this result in mind, we return to (167). Consider a

fixed realization $n_i = n_i^*$ and denote the value of the left-hand side (LHS) for n_i^* by α. By defining $u \triangleq \delta(\mathbf{z}_{\mathrm{pa}(i)})$, we get

$$\alpha = \frac{r'_q(n_i^* + u)}{r_q(n_i^* + u)} \ , \qquad \forall u \in \Theta \ . \tag{169}$$

This is only possible if r_q is an exponential function, i.e., $r_q(u) = k_1 \exp(\alpha u)$ over interval $u \in \Theta$. Since r_q is an analytic function, it is, therefore, exponential over entire $\mathbb{R}$. Then, the associated pdf must have the form $q_N(u) = k_2 \exp((k_1/\alpha) \exp(\alpha u))$. However, the integral of this function over the entire domain $\mathbb{R}$ diverges. Hence, it is not a valid pdf, rendering a contradiction. Hence, the additive noise model satisfies interventional regularity.

Appendix B. Proofs of the Results for Linear Transformations

B.1 Auxiliary Results

First, we provide a linear algebraic property, which will be used in the proofs.

Lemma 33 *Consider the latent causal graph $\mathcal{G}$, and a matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$.*

1. *Let $\mathbf{L}_{\mathrm{pa}}$ and $\mathbf{L}_{\mathrm{an}}$ be binary matrices that denote the parental and ancestral relationships in $\mathcal{G}$, respectively, i.e.,*

$$[\mathbf{L}_{\mathrm{pa}}]_{i,j} \triangleq \begin{cases} 1 \ , & \text{if} \quad j \in \overline{\mathrm{pa}}(i) \\ 0 \ , & \text{otherwise} \end{cases} \ , \quad and \quad [\mathbf{L}_{\mathrm{an}}]_{i,j} \triangleq \begin{cases} 1 \ , & \text{if} \quad j \in \overline{\mathrm{an}}(i) \\ 0 \ , & \text{otherwise} \end{cases} \ . \tag{170}$$

   Then, if $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}\} \preccurlyeq \mathbf{L}_{\mathrm{pa}}$, we also have $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}^{-1}\} \preccurlyeq \mathbf{L}_{\mathrm{an}}$.

2. *Let $\mathbf{L}_{\mathrm{sur}}$ be a binary matrix that denotes the surrounding relationships in $\mathcal{G}$, i.e.,*

$$[\mathbf{L}_{\mathrm{sur}}]_{i,j} \triangleq \begin{cases} 1 \ , & \text{if} \quad \overline{\mathrm{ch}}(i) \subseteq \overline{\mathrm{ch}}(j) \ , \\ 0 \ , & \text{otherwise} \end{cases} \ . \tag{171}$$

   Then, if $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}\} \preccurlyeq \mathbf{L}_{\mathrm{sur}}$, we also have $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}^{-1}\} \preccurlyeq \mathbf{L}_{\mathrm{sur}}$.

Proof: First, note that $\mathbf{L}_{\mathrm{sur}} \preccurlyeq \mathbf{L}_{\mathrm{pa}} \preccurlyeq \mathbf{L}_{\mathrm{an}}$. Hence, we start with considering a generic lower triangular matrix $\mathbf{L}$ such that $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}\} \preccurlyeq \mathbf{L}_{\mathrm{pa}}$. Matrix $\mathbf{L}$ can be decomposed as

$$\mathbf{L} = \mathbf{E} \cdot (\mathbf{I}_{n \times n} + \mathbf{\Lambda}) \ , \tag{172}$$

in which $\mathbf{E}$ is a diagonal matrix and $\mathbf{\Lambda}$ is a strictly lower triangular matrix that satisfies $\mathbb{1}\{\mathbf{\Lambda}\} \preccurlyeq \mathbf{L}_{\mathrm{pa}}$. Subsequently, since $\mathbf{E}$ is a diagonal matrix, we have

$$\mathbf{L}^{-1} = (\mathbf{I}_{n \times n} + \mathbf{\Lambda})^{-1} \cdot \mathbf{E}^{-1} \ , \quad \text{and} \quad \mathbb{1}\{\mathbf{L}^{-1}\} = \mathbb{1}\{(\mathbf{I}_{n \times n} + \mathbf{\Lambda})^{-1}\} \ . \tag{173}$$

Note that $\mathbf{\Lambda}$ is a strictly lower triangular $n \times n$ matrix, which implies that $\mathbf{\Lambda}^n$ is a zero matrix. Therefore, the inverse of $(\mathbf{I}_{n \times n} + \mathbf{\Lambda})$ can be expanded as

$$(\mathbf{I}_{n \times n} + \mathbf{\Lambda})^{-1} = \mathbf{I}_{n \times n} - \mathbf{\Lambda} + \mathbf{\Lambda}^2 - \cdots + (-1)^{n-1} \mathbf{\Lambda}^{n-1} \ . \tag{174}$$

If $\left[\mathbf{I}_{n \times n} + \mathbf{\Lambda}\right]_{i,j}^{-1} \neq 0$ for some $i, j \in [n]$ with $j < i$, by (174) we have $[\mathbf{\Lambda}^k]_{i,j} \neq 0$ for some $k \in [n-1]$. Expanding matrix $\mathbf{\Lambda}^k$ yields that $[\mathbf{\Lambda}^k]_{i,j}$ is equal to the sum of the products with k terms, i.e.,

$$[\mathbf{\Lambda}^k]_{i,j} = \sum_{i > a_1 > \cdots > a_{k-1} > j} [\mathbf{\Lambda}]_{i,a_1}[\mathbf{\Lambda}]_{a_1,a_2} \ldots [\mathbf{\Lambda}]_{a_{k-1},j} . \tag{175}$$

Therefore, if $[\mathbf{\Lambda}^k]_{i,j} \neq 0$, there exists a sequence of entries $\left([\mathbf{\Lambda}]_{i,a_1}, [\mathbf{\Lambda}]_{a_1,a_2}, \ldots, [\mathbf{\Lambda}]_{a_{k-1},j}\right)$ in which all terms are nonzero. Next, we prove the two cases as follows.

Case 1: Since $\mathbb{1}\{\mathbf{\Lambda}\} \preccurlyeq \mathbf{L}_{\mathrm{pa}}$, we have $[\mathbf{L}_{\mathrm{pa}}]_{i,a_1} = [\mathbf{L}_{\mathrm{pa}}]_{a_1,a_2} = \cdots = [\mathbf{L}_{\mathrm{pa}}]_{a_{k-1},j} = 1$. By the definition of $\mathbf{L}_{\mathrm{pa}}$, this means that there is a path $j \to a_{k-1} \to \cdots \to a_1 \to i$ in $\mathcal{G}$, which implies that $j \in \mathrm{an}(i)$ and $[\mathbf{L}_{\mathrm{an}}]_{i,j} = 1$. Therefore, we conclude that

$$\left[(\mathbf{I}_{n \times n} + \mathbf{\Lambda})^{-1}\right]_{i,j} \neq 0 \quad \Longrightarrow \quad [\mathbf{L}_{\mathrm{an}}]_{i,j} = 1 . \tag{176}$$

Hence, $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}^{-1}\} \preccurlyeq \mathbf{L}_{\mathrm{an}}$.

Case 2: Since $\mathbb{1}\{\mathbf{\Lambda}\} \preccurlyeq \mathbf{L}_{\mathrm{sur}}$, we have $[\mathbf{L}_{\mathrm{sur}}]_{i,a_1} = [\mathbf{L}_{\mathrm{sur}}]_{a_1,a_2} = \cdots = [\mathbf{L}_{\mathrm{sur}}]_{a_{k-1},j} = 1$. By the definition of $\mathbf{L}_{\mathrm{sur}}$, this means that

$$\overline{\mathrm{ch}}(i) \subseteq \mathrm{ch}(a_1) \subset \cdots \subset \mathrm{ch}(a_{k-1}) \subset \mathrm{ch}(j) , \tag{177}$$

which implies $[\mathbf{L}_{\mathrm{sur}}]_{i,j} = 1$ since $\overline{\mathrm{ch}}(i) \subseteq \overline{\mathrm{ch}}(j)$. Therefore, we conclude that if

$$\left[(\mathbf{I}_{n \times n} + \mathbf{\Lambda})^{-1}\right]_{i,j} \neq 0 \quad \Longrightarrow \quad [\mathbf{L}_{\mathrm{sur}}]_{i,j} = 1 . \tag{178}$$

Hence, $\mathbf{I}_{n \times n} \preccurlyeq \mathbb{1}\{\mathbf{L}^{-1}\} \preccurlyeq \mathbf{L}_{\mathrm{sur}}$.

B.2 Proof of Theorem 16

Proof of the scaling consistency. First, by Theorem 14, encoder estimate $\hat{\mathbf{H}}$ at the end of Stage L1 of Algorithm 1 satisfies

$$\hat{\mathbf{H}} \cdot \mathbf{G} = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{L} , \tag{179}$$

where $\mathbf{P}_{\mathcal{I}}$ is the permutation matrix of the intervention order $(I^1, \ldots, I^n)$, $\mathbf{L}$ has nonzero diagonal entries and satisfies $\mathbf{L}_{i,j} = 0$ for all $j \notin \overline{\mathrm{pa}}(i)$. We will show that the output of Stage L3 of Algorithm 1 for hard interventions satisfies that

$$\mathbb{1}\left([\hat{\mathbf{H}} \cdot \mathbf{G}]_m\right) = \mathbf{e}_{I^m}^\top , \quad \forall m \in [n] , \tag{180}$$

which will imply scaling consistency as

$$\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}}) = \hat{\mathbf{H}} \cdot \mathbf{X} = \hat{\mathbf{H}} \cdot \mathbf{G} \cdot \mathbf{Z} = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\mathrm{s}} \cdot \mathbf{Z} , \tag{181}$$

where $\mathbf{C}_{\mathrm{s}}$ is a constant diagonal matrix with nonzero diagonal entries. We prove this as follows.

First, consider the topological order π of $\hat{\mathcal{G}}$. By Theorem 14, transitive closures of $\hat{\mathcal{G}}$ and $\mathcal{G}$ are the same under relabeling of the nodes with permutation $(1, \ldots, n) \to (I^1, \ldots, I^n)$.

Then, since $(1, \ldots, n)$ is assumed to be topological order of $\mathcal{G}$, $\rho \triangleq (I^{\pi_1}, \ldots, I^{\pi_n})$ is also a topological order of $\mathcal{G}$. According to this notation, in the π_t-th environment, node ρ_t is intervened.

Next, since ρ_1 has no parent in $\mathcal{G}$, by Theorem 14 we already know that $\hat{Z}_{\pi_1} = b_1 \cdot Z_{\rho_1}$ for some constant $b_1 \in \mathbb{R}_+$. Consider $m = \pi_k$ step of the algorithm, and assume that for all $t \in [k-1]$, the updated encoder rows satisfy

$$\hat{Z}_{\pi_t} = \hat{\mathbf{H}}_{\pi_t} \cdot \mathbf{G} \cdot \mathbf{Z} = b_t \cdot Z_{\rho_t} , \tag{182}$$

for some constant $b_t \in \mathbb{R}_+$. We will show that the update in step $m = \pi_k$ will ensure that $\hat{Z}_{\pi_k}$ also satisfies (182). Before the update, by Theorem 14 we know that $\hat{Z}_{\pi_k}$ is a linear function of Z_{ρ_k} and $\mathbf{Z}_{\mathrm{pa}(\rho_k)}$. Then, using (182), we have

$$\hat{Z}_{\pi_k} = c_0 \cdot Z_{\rho_k} + \mathbf{c} \cdot \hat{\mathbf{Z}}_{\hat{\mathrm{pa}}(\pi_k)} , \tag{183}$$

for $|\hat{\mathrm{pa}}(\pi_k)|$-dimensional vector $\mathbf{c}$ and nonzero constant $c_0 \in \mathbb{R}_+$. We will use Theorem 15 in environment $\mathcal{E}^{\pi_k}$. For brevity, using $m = \pi_k$ and letting $\mathbf{V} \triangleq \hat{\mathbf{Z}}^m_{\hat{\mathrm{pa}}(\pi_k)}$,

$$\hat{Z}^m_m = c_0 \cdot Z^m_{\rho_k} + \mathbf{c} \cdot \mathbf{V} . \tag{184}$$

Since node ρ_k is intervened in environment $\mathcal{E}^m$, by Theorem 15, we know that $Z^m_{\rho_k} \perp\!\!\!\perp \mathbf{V}$, which implies that

$$\mathrm{Cov}(Z^m_{\rho_k}, \mathbf{V}) = \mathbf{0}_{1 \times |\mathbf{V}|} . \tag{185}$$

Consider a $|\mathbf{V}|$-dimensional row vector $\mathbf{u}$ and let

$$Y = \left(\hat{\mathbf{H}}_m - \mathbf{u} \cdot \hat{\mathbf{H}}_{\hat{\mathrm{pa}}(m)} \right) \cdot \mathbf{X}^m = \hat{Z}^m_m - \mathbf{u} \cdot \mathbf{V} . \tag{186}$$

Note that $\mathbf{u} = \mathbf{c}$ would yield that $Y = c_0 \cdot Z^m_{\rho_k}$, and subsequently $Y \perp\!\!\!\perp \mathbf{V}$. On the other hand, $Y \perp\!\!\!\perp \mathbf{V}$ implies that

$$\mathbf{0}_{1 \times |\mathbf{V}|} = \mathrm{Cov}(Y, \mathbf{V}) = \mathrm{Cov}(\hat{Z}^m_m - \mathbf{u} \cdot \mathbf{V}, \mathbf{V}) , \tag{187}$$

which has a unique solution

$$\mathbf{u} = \mathrm{Cov}(\hat{Z}^m_m, \mathbf{V}) \cdot \left(\mathrm{Cov}(\mathbf{V}) \right)^{-1} . \tag{188}$$

i.e., linear minimum mean square error (LMMSE) estimator. Finally, $\mathrm{Cov}(\mathbf{V})$ is invertible since the causal relationships among the entries in $\hat{Z}^m_{\hat{\mathrm{pa}}(\pi_k)}$ are not deterministic. Using this row vector $\mathbf{u}$, we update

$$\hat{\mathbf{H}}_m \leftarrow \mathbf{H}_m - \mathbf{u} \cdot \mathbf{H}_{\hat{\mathrm{pa}}(\pi_k)} , \tag{189}$$

and achieve $\hat{Z}_m = c_0 \cdot Z_{\rho_k}$. Therefore, (182) holds for $t = k$ as well, and by induction we obtain

$$\hat{\mathbf{Z}} = \hat{\mathbf{H}} \cdot \mathbf{G} \cdot \mathbf{Z} = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\mathrm{s}} \cdot \mathbf{Z} , \tag{190}$$

for a diagonal matrix $\mathbf{C}_{\mathrm{s}}$ with nonzero diagonal entries. $\qquad \square$

Proof of the perfect DAG recovery. We show that the graph construction process in (59) achieves perfect DAG recovery as follows. First, using Theorem 8 and (190), we have

$$\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) = (\mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\mathrm{s}})^{-\top} \cdot \boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{Z}) = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\mathrm{s}}^{-\top} \cdot \boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{Z}) \ . \tag{191}$$

Subsequently, using the fact that $\mathbf{C}_{\mathrm{s}}$ is a diagonal matrix, for all $i, m \in [n]$ we obtain

$$\left[\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}})\right]_i \neq 0 \quad \Longleftrightarrow \quad \left[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{Z})\right]_{I^i} \neq 0 \ . \tag{192}$$

Then, Theorem 7 and (59), we have

$$I^i \in \overline{\mathrm{pa}}(I^m) \quad \Longleftrightarrow \quad \mathbb{E}\left[\left|\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{Z})\right|\right]_{I^i} \neq 0 \quad \Longleftrightarrow \quad \mathbb{E}\left[\left|\boldsymbol{d}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}})\right|_i\right] \quad \Longleftrightarrow \quad i \in \hat{\overline{\mathrm{pa}}}(m) \ . \tag{193}$$

This concludes the proof that $\hat{\mathcal{G}}$ and $\mathcal{G}$ are related through a graph isomorphism by permutation $\mathcal{I}$, which denotes the intervention order. $\qquad\square$

B.3 Proof of Theorem 18

Latent graph estimation. We continue from the intuitions provided in Section 5.4 and prove the perfect recovery of the latent graph as follows. First, recall that output $\hat{\mathcal{G}}$ of Algorithm 1 for soft interventions satisfies that transitive closures of $\hat{\mathcal{G}}$ and $\mathcal{G}$ are the same under permutation $(I^1, \ldots, I^n)$. Since $(1, \ldots, n)$ is assumed to be a topological order of $\mathcal{G}$, $\rho \triangleq (I^{\pi_1}, \ldots, I^{\pi_n})$ is also a topological order of $\mathcal{G}$. Then, for any $k = \pi_u$ and $t = \pi_v \in \{\pi_1, \ldots, \pi_{u-1}\}$, (64) becomes

$$\dim\left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^t) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^k)\right) = |\mathrm{pa}(\rho_u) \cap \mathrm{pa}(\rho_v)| + \mathbb{1}\{\rho_v \in \mathrm{pa}(\rho_u)\} \ . \tag{194}$$

By induction, we will prove that the parent sets generated by Algorithm 2 satisfy

$$\pi_u \in \hat{\mathrm{pa}}(\pi_v) \quad \Longleftrightarrow \quad \rho_u \in \mathrm{pa}(\rho_v) \ . \tag{195}$$

At the base case, consider $k = \pi_2$. The only possible parent is $t = \pi_1$. Since ρ_1 is a root node in $\mathcal{G}$, (194) implies

$$\dim\left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^{\pi_1}) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^{\pi_2})\right) = \mathbb{1}\{\rho_1 \in \mathrm{pa}(\rho_2)\} \ . \tag{196}$$

Since $\hat{\mathrm{pa}}(\pi_1) = \hat{\mathrm{pa}}(\pi_2) = \emptyset$ at this step, the algorithm adds π_1 to $\hat{\mathrm{pa}}(\pi_2)$ if and only if $\rho_1 \in \mathrm{pa}(\rho_2)$, and (195) is satisfied for $k = \pi_2$. Next, **as the induction step**, assume that the algorithm output $\hat{\mathrm{pa}}(k)$ satisfies (195) for $k \in \{\pi_1, \ldots, \pi_{u-1}\}$. We prove that $k = \pi_u$ also satisfies the condition by induction again. At the base case, consider $t = \pi_1$. Similar to the previous base case, (194) implies

$$\dim\left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^{\pi_1}) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^{\pi_u})\right) = \mathbb{1}\{\rho_1 \in \mathrm{pa}(\rho_u)\} \ , \tag{197}$$

and since $\hat{\mathrm{pa}}(\pi_u)$ does not contain π_1 yet, the algorithm adds π_1 to $\hat{\mathrm{pa}}(\pi_u)$ if and only if $\rho_1 \in \mathrm{pa}(\rho_u)$. Next, as the induction step, assume that for $\pi_j \in \{\pi_1, \ldots, \pi_{v-1}\}$ where $v < u$, the algorithm correctly identified the parents, i.e., $\pi_j \in \hat{\mathrm{pa}}(\pi_u)$ if and only if $\rho_j \in \mathrm{pa}(\rho_u)$. Consider $k = \pi_u$ and $t = \pi_v$, for which (194) implies

$$\dim\left(\mathrm{col}(\mathbf{R}_{\mathbf{X}}^{\pi_v}) \cap \mathrm{col}(\mathbf{R}_{\mathbf{X}}^{\pi_u})\right) = |\mathrm{pa}(\rho_u) \cap \mathrm{pa}(\rho_v)| + \mathbb{1}\{\rho_v \in \mathrm{pa}(\rho_u)\} \ . \tag{198}$$

By the induction hypothesis, we know that $|\hat{\mathrm{pa}}(\pi_u) \cap \hat{\mathrm{pa}}(\pi_v)| = |\mathrm{pa}(\rho_u) \cap \mathrm{pa}(\rho_v)|$. Therefore, the algorithm adds π_v to $\hat{\mathrm{pa}}(\pi_v)$ if and only if $\rho_v \in \mathrm{pa}(\rho_u)$. Therefore, by the inner induction, $\hat{\mathrm{pa}}(k)$ satisfies (195). Then, by the outer induction, (195) is satisfied for all parent set estimates, which concludes the proof of perfect graph recovery. $\qquad\square$

Markov property of consistency up to mixing with surrounding parents. Let $\hat{\mathbf{Z}}$ denote $\hat{\mathbf{Z}}(\mathbf{X}; \hat{\mathbf{H}})$. We further investigate the properties of consistency up to mixing with surrounding parents to prove that $\hat{\mathbf{Z}}$ is Markov with respect to $\hat{\mathcal{G}}$. Using the results of the encoder estimation step, we have

$$\hat{\mathbf{Z}} = \mathbf{P}_{\mathcal{I}} \cdot \mathbf{C}_{\text{sur}} \cdot \mathbf{Z} \ , \tag{199}$$

where $\mathbf{C}_{\text{sur}}$ has nonzero diagonal entries and satisfies $[\mathbf{C}_{\text{sur}}]_{i,j} = 0$ for all $j \notin \overline{\text{sur}}(i)$. Recall the following SCM specified in (6)

$$Z_i = f_i(\mathbf{Z}_{\text{pa}(i)}, N_i) \ . \tag{200}$$

Let τ be the permutation that maps $\{1, \ldots, n\}$ to $\mathcal{I}$, i.e., $I^{\tau_i} = i$ for all $i \in [n]$. Since $\text{sur}(i) \subseteq \text{pa}(i)$, (199) implies

$$\hat{Z}_{\tau_i} = [\mathbf{C}_{\text{sur}}]_{i,i} \cdot Z_i + \sum_{j \in \text{sur}(i)} [\mathbf{C}_{\text{sur}}]_{i,j} \cdot Z_j \tag{201}$$

$$= [\mathbf{C}_{\text{sur}}]_{i,i} \cdot f_i(\mathbf{Z}_{\text{pa}(i)}, N_i) + \sum_{j \in \text{sur}(i)} [\mathbf{C}_{\text{sur}}]_{i,j} \cdot Z_j \ , \tag{202}$$

which is a function of $\mathbf{Z}_{\text{pa}(i)}$ and N_i. To prove Markov property, we need to specify Z_{τ_i} in terms of $\hat{\mathbf{Z}}_{\hat{\text{pa}}(\tau_i)}$. To this end, it suffices to show that for any $k \in \text{pa}(i)$, Z_k is a function of $\hat{\mathbf{Z}}_{\hat{\text{pa}}(\tau_i)}$. Using (199), we have

$$\mathbf{Z} = \mathbf{C}_{\text{sur}}^{-1} \cdot \mathbf{P}_{\tau} \cdot \hat{\mathbf{Z}} \ . \tag{203}$$

By Theorem 33, we know that $[\mathbf{C}_{\text{sur}}^{-1}]_{i,j} = 0$ if $j \notin \text{sur}(i)$ for distinct i and j. Subsequently, as counterpart of (201), we have

$$Z_k = [\mathbf{C}_{\text{sur}}^{-1}]_{k,k} \cdot \hat{Z}_{\tau_k} + \sum_{j \in \text{sur}(k)} [\mathbf{C}_{\text{sur}}^{-1}]_{k,j} \cdot \hat{Z}_{\tau_j} \ . \tag{204}$$

Since $k \in \text{pa}(i)$, we have $\tau_k \in \text{pa}(\tau_i)$ and $\hat{Z}_{\tau_k}$ is in $\hat{\mathbf{Z}}_{\hat{\text{pa}}(\tau_i)}$. Note that if $j \in \text{sur}(k)$, j is also in $\text{pa}(i)$. Therefore, every term in the RHS of (204) belongs to $\hat{\mathbf{Z}}_{\hat{\text{pa}}(\tau_i)}$ and Z_k is a function of $\hat{\mathbf{Z}}_{\hat{\text{pa}}(\tau_i)}$. Then, using (202), we know that $\hat{Z}_{\tau_i}$ is a function of only $\hat{\mathbf{Z}}_{\hat{\text{pa}}(\tau_i)}$ and N_i, which concludes the proof that $\hat{\mathbf{Z}}$ is Markov with respect to $\hat{\mathcal{G}}$.

Appendix C. Proofs of the Results for General Transformations

The proof of Theorem 25 is already given in Section 6.2. In this section, we first prove Theorem 26, that is, identifiability for coupled interventions without observational environment. Then, we prove the results in Section 6.3 related to the uncoupled interventions.

C.1 Proof of Theorem 26

We will show that if $p(\mathbf{Z})$ is adjacency-faithful to $\mathcal{G}$ and the latent causal model is an additive noise model, then we can recover $\mathcal{G}$ without having access to observational environment $\mathcal{E}^0$.

Let τ be the permutation that maps $\{1, \ldots, n\}$ to $\mathcal{I}$, i.e., $I^{\tau_i} = i$ for all $i \in [n]$ and $\mathbf{P}_\tau$ to denote the permutation matrix that corresponds to τ, i.e.,

$$[\mathbf{P}_\tau]_{i,m} = \begin{cases} 1 , & m = \rho_i , \\ 0 , & \text{else} . \end{cases} \tag{205}$$

By Theorem 7(iv), true latent score changes across $\{\mathcal{E}^{\tau_i}, \tilde{\mathcal{E}}^{\tau_j}\}$ gives us $\overline{\mathrm{pa}}(i, j)$ for $i \neq j$. First, we use the perfect latent recovery result to show that Theorem 7(iv) also applies to estimated latent score changes. Denote $f = \hat{h} \circ g$. Using score transform between $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ in (40), and recalling $\mathbb{1}\{J_f^{-1}\} = \mathbf{P}_\tau$, we have

$$\left[\boldsymbol{s}_{\hat{\mathbf{Z}}}^{\tau_i}(\hat{\mathbf{z}}; \hat{h}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^{\tau_j}(\hat{\mathbf{z}}; \hat{h})\right]_{\tau_k} = \left[J_f^{-\top}(\mathbf{z})\right]_{\tau_k} \cdot \left[\boldsymbol{s}^{\tau_i}(\mathbf{z}) - \tilde{\boldsymbol{s}}^{\tau_j}(\mathbf{z})\right] \tag{206}$$

$$= \left[J_f^{-\top}(\mathbf{z})\right]_{\tau_k, k} \cdot \left[\boldsymbol{s}^{\tau_i}(\mathbf{z}) - \tilde{\boldsymbol{s}}^{\tau_j}(\mathbf{z})\right]_k . \tag{207}$$

Using $\left[J_f^{-\top}(\mathbf{z})\right]_{\tau_k, k} \neq 0$ for all $\mathbf{z} \in \mathbb{R}^n$, we have

$$\mathbb{E}\left[\left|\boldsymbol{s}_{\hat{\mathbf{Z}}}^{\tau_i}(\hat{\mathbf{Z}}; \hat{h}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^{\tau_j}(\hat{\mathbf{Z}}; \hat{h})\big|_{\tau_k}\right] \neq 0 \quad \Longleftrightarrow \quad \mathbb{E}\left[\left|\boldsymbol{s}^{\tau_i}(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\tau_j}(\mathbf{Z})\big|_{\tau_k}\right] \neq 0 . \tag{208}$$

Hence, by Theorem 7(iv),

$$\mathbb{E}\left[\left|\boldsymbol{s}_{\hat{\mathbf{Z}}}^{\tau_i}(\hat{\mathbf{Z}}; \hat{h}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^{\tau_j}(\hat{\mathbf{Z}}; \hat{h})\big|_{\tau_k}\right] \neq 0 \quad \Longleftrightarrow \quad k \in \overline{\mathrm{pa}}(i, j) . \tag{209}$$

Let us denote the graph $\mathcal{G}_\tau$ that is related to $\mathcal{G}$ by permutation τ, i.e., $i \in \mathrm{pa}(j)$ if and only if $\tau_i \in \mathrm{pa}_\tau(\tau_j)$ for which $\mathrm{pa}_\tau(\tau_j)$ denotes the parents of node τ_j in $\mathcal{G}_\tau$. Using (209), we have

$$\mathbb{E}\left[\left|\boldsymbol{s}_{\hat{\mathbf{Z}}}^{\tau_i}(\hat{\mathbf{Z}}; \hat{h}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^{\tau_j}(\hat{\mathbf{Z}}; \hat{h})\big|_{\tau_k}\right] \neq 0 \quad \Longleftrightarrow \quad \tau_k \in \overline{\mathrm{pa}}_\tau(\tau_i, \tau_j) . \tag{210}$$

In the rest of the proof, we will show how to obtain $\{\mathrm{pa}_\tau(i) : i \in [n]\}$ using $\{\overline{\mathrm{pa}}_\tau(i, j) : i, j \in [n], i \neq j\}$. Since $\mathcal{G}_\tau$ is a graph isomorphism of $\mathcal{G}$, this problem is equivalent to obtaining $\{\mathrm{pa}(i) : i \in [n]\}$ using $\{\overline{\mathrm{pa}}(i, j) : i, j \in [n], i \neq j\}$. Note that $\hat{Z}_i$ (which corresponds to node i in $\mathcal{G}_\tau$) is intervened in environments $\mathcal{E}^i$ and $\tilde{\mathcal{E}}^i$. We denote the set of root nodes by

$$\mathcal{K} \triangleq \{i \in [n] : \mathrm{pa}(i) = \emptyset\} , \tag{211}$$

and also define

$$\mathcal{B}_i \triangleq \bigcap_{j \neq i} \overline{\mathrm{pa}}(i, j) , \quad \forall i \in [n] , \quad \text{and} \quad \mathcal{B} \triangleq \{i : |\mathcal{B}_i| = 1\} . \tag{212}$$

Note that $\overline{\mathrm{pa}}(i) \subseteq \mathcal{B}_i$. Hence, $|\mathcal{B}_i| = 1$ implies that i is a root node. We investigate the graph recovery in three cases.

1. $|\mathcal{B}| \geq 3$: For any node $i \in [n]$, we have

$$\overline{\mathrm{pa}}(i) \subseteq \mathcal{B}_i \subseteq \bigcap_{j \in \mathcal{K} \setminus \{i\}} \overline{\mathrm{pa}}(i, j) = \overline{\mathrm{pa}}(i) \cup \left\{\bigcap_{j \in \mathcal{K} \setminus \{i\}} \{j\}\right\} = \overline{\mathrm{pa}}(i) . \tag{213}$$

Note that the last equality is due to $\bigcap_{j \in \mathcal{K} \setminus \{i\}} \{j\} = \emptyset$ since there are at least two root nodes excluding i. Then, $\mathcal{B}_i = \overline{\mathrm{pa}}(i)$ for all $i \in [n]$ and we are done.

2. $|\mathcal{B}| = 2$: The two nodes in $\mathcal{B}$ are root nodes. If there were at least three root nodes, we would have at least three nodes in $\mathcal{B}$. Hence, the two nodes in $\mathcal{B}$ are the only root nodes. Subsequently, every $i \notin \mathcal{B}$ is also not in $\mathcal{K}$ and we have

$$\overline{\mathrm{pa}}(i) \subseteq \mathcal{B}_i \subseteq \bigcap_{j \in \mathcal{K}} \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(i) \cup \left\{ \bigcap_{j \in \mathcal{K}} \{j\} \right\} = \overline{\mathrm{pa}}(i) \ . \tag{214}$$

Hence, $\mathcal{B}_i = \overline{\mathrm{pa}}(i)$ for every non-root node i and we already have the two root nodes in $\mathcal{B}$, which completes the graph recovery.

3. $|\mathcal{B}| \leq 1$: First, consider all (i,j) pairs such that $|\overline{\mathrm{pa}}(i,j)| = 2$. For such an (i,j) pair, at least one of the nodes is a root node; otherwise $\overline{\mathrm{pa}}(i,j)$ would contain a third node. Using these pairs, we identify all root nodes as follows. Note that a hard intervention on node i makes Z_i independent of all of its non-descendants, and all conditional independence relations are preserved under componentwise diffeomorphisms such as f. Then, using the adjacency-faithfulness assumption, we infer that

- if $\hat{Z}_i \perp\!\!\!\perp \hat{Z}_j$ in $\mathcal{E}^i$ and $\hat{Z}_i \perp\!\!\!\perp \hat{Z}_j$ in $\tilde{\mathcal{E}}^j$, then both i and j are root nodes.
- if $\hat{Z}_i \not\perp\!\!\!\perp \hat{Z}_j$ in $\mathcal{E}^i$, then $i \to j$ and i is a root node.
- if $\hat{Z}_i \not\perp\!\!\!\perp \hat{Z}_j$ in $\tilde{\mathcal{E}}^j$, then $j \to i$ and j is a root node.

This implies that we can determine whether i and j nodes are root nodes by using at most two independence tests. Hence, we identify all root nodes by using at most n independence tests. We also know that there are at most two root nodes. If we have two root nodes, then $\mathcal{B}_i = \overline{\mathrm{pa}}(i)$ for all non-root nodes, and the graph is recovered. If we have only one root node i, then for any $j \neq i$ we have

$$\overline{\mathrm{pa}}(j) \subseteq \mathcal{B}_j \subseteq \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(j) \cup \{i\} \ . \tag{215}$$

Finally, if $\hat{Z}_j \perp\!\!\!\perp \hat{Z}_i \mid \{\hat{Z}_\ell : \ell \in \mathcal{B}_j \setminus \{i\}\}$ in $\tilde{\mathcal{E}}^j$, we have $i \notin \overline{\mathrm{pa}}(j)$ due to adjacency-faithfulness. Otherwise, we conclude that $i \in \overline{\mathrm{pa}}(j)$. Hence, an additional $(n-1)$ conditional independence tests ensure the recovery of all $\overline{\mathrm{pa}}(j)$ sets, and the graph recovery is complete.

C.2 Proof of Lemma 27

Let us start by scrutinizing the constraints. For the true encoder g^{-1}, Theorem 7 gives us

$$[\mathbf{D}(g^{-1})]_{i,m} \neq 0 \iff i \in \overline{\mathrm{pa}}(I^m) \ , \tag{216}$$

$$\text{and} \quad [\tilde{\mathbf{D}}(g^{-1})]_{i,m} \neq 0 \iff i \in \overline{\mathrm{pa}}(\tilde{I}^m) \ . \tag{217}$$

Therefore, using $\tilde{\mathcal{I}} = \sigma \circ \mathcal{I}$, we have $\mathbb{1}\{\tilde{\mathbf{D}}(g^{-1})\} = \mathbb{1}\{\mathbf{D}(g^{-1})\} \cdot \mathbf{P}_\sigma$. Note that acyclicity of graph $\mathcal{G}$ also implies that $\mathbb{1}\{\mathbf{D}(g^{-1})\} \odot \mathbb{1}\{\mathbf{D}(g^{-1})^\top\} = \mathbf{I}_{n \times n}$.

Next, note that $\mathbf{D}_t(g^{-1}; \sigma)$ is equal to the true score change matrix $\mathbf{D}_t(g^{-1})$ defined for the coupled interventions, which satisfies $\mathbb{1}\{\mathbf{D}_t(g^{-1}; \sigma)\} = \mathbf{P}_{\mathcal{I}}^\top$. Let $h^* = [\mathbf{D}_t(g^{-1}; \sigma)]^\top \cdot g^{-1}$.

Since $\mathbf{D}_t(g^{-1}; \sigma)$ is a scaled permutation matrix, using (77) we obtain

$$\mathbf{D}_t(h^*; \sigma) = [\mathbf{D}_t(g^{-1}; \sigma)]^{-1} \cdot \mathbf{D}_t(g^{-1}; \sigma) = \mathbf{I}_{n \times n} \ , \tag{218}$$

$$\mathbf{D}(h^*) = [\mathbf{D}_t(g^{-1}; \sigma)]^{-1} \cdot \mathbf{D}(g^{-1}) \tag{219}$$

$$\tilde{\mathbf{D}}(h^*) = [\mathbf{D}_t(g^{-1}; \sigma)]^{-1} \cdot \tilde{\mathbf{D}}(g^{-1}) \ . \tag{220}$$

Thus, using $\mathbb{1}\{\tilde{\mathbf{D}}(g^{-1})\} = \mathbb{1}\{\mathbf{D}(g^{-1})\} \cdot \mathbf{P}_\sigma$, we have

$$\mathbb{1}\{\tilde{\mathbf{D}}(h^*)\} = \mathbb{1}\{\mathbf{D}(h^*)\} \cdot \mathbf{P}_\sigma \ , \tag{221}$$

which satisfies the first constraint. Also, using $\mathbb{1}\{\mathbf{D}(h)\} = \mathbf{P}_\sigma \cdot \mathbb{1}\{\tilde{\mathbf{D}}(h)\}$,

$$\mathbb{1}\{\mathbf{D}(h^*)\} = \mathbf{P}_{\mathcal{I}} \cdot \mathbb{1}\{\mathbf{D}(g^{-1})\} \ . \tag{222}$$

Therefore, we have $\mathbb{1}\{\mathbf{D}(h^*)\} \odot \mathbb{1}\{\mathbf{D}(h^*)^\top\} = \mathbf{I}_{n \times n}$. Finally, $(h^*)^{-1} \triangleq g \cdot \left[\mathbf{D}_t(g^{-1}; \sigma)\right]^{-\top}$ satisfies $(h^*)^{-1} \circ h(\mathbf{X}) = \mathbf{X}$. Hence, (h^*, σ) minimizes the objective in (92) at value zero. $\square$

C.3 Proof of Lemma 28

We will prove the desired result by contradiction. Suppose that (h^*, π) is a solution to the optimization problem specified in (92) and let $\pi \neq \sigma$. Denote $f = h^* \circ g$, so we have $\hat{\mathbf{Z}} = f(\mathbf{Z})$. If $\pi_m = \sigma_m$ for some $m \in [n]$, this means $I^m = \tilde{I}^{\pi_m} = \ell$ for some node $\ell \in [n]$. We follow the same approach in the proof of Theorem 22. Specifically, using score difference transformation property in (40) and one-sparse property of $[\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\pi_m}(\mathbf{Z})]$ via Theorem 7(iii), we have

$$[\mathbf{D}_t(h^*)]_{i,m} = \mathbb{E}\left[\left|\boldsymbol{s}_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}) - \tilde{\boldsymbol{s}}_{\hat{\mathbf{Z}}}^{\pi_m}(\hat{\mathbf{Z}})\right|_i\right] \tag{223}$$

$$= \mathbb{E}\left[\left|\left[J_f^{-\top}(\mathbf{Z})\right]_i \cdot \left[\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\pi_m}(\mathbf{Z})\right]\right|\right] \tag{224}$$

$$= \mathbb{E}\left[\left[J_f^{-1}(\mathbf{Z})\right]_{\ell,i} \cdot \left[\boldsymbol{s}^m(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\pi_m}(\mathbf{Z})\right]_\ell\right] \ . \tag{225}$$

By definition of h^*, $[\mathbf{D}_t(h^*)]_{i,m} = 0$ for all $i \neq m$. Also, interventional discrepancy between $q_\ell(z_\ell)$ and $\tilde{q}_\ell(z_\ell)$ implies that $[\boldsymbol{s}^m(\mathbf{z}) - \tilde{\boldsymbol{s}}^{\pi_m}(\mathbf{z})]_\ell \neq 0$ except for a null set. Then, $[\mathbf{D}_t(h^*)]_{i,m} = 0$ implies that $[J_f^{-1}(\mathbf{z})]_{\ell,i} = 0$ except for a null set. Since J_f^{-1} is a continuous function, this implies that $[J_f^{-1}(\mathbf{z})]_{\ell,i} = 0$ for all $\mathbf{z} \in \mathbb{R}^n$. Furthermore, since J_f^{-1} is invertible for all $\mathbf{z}$, none of its columns can be a zero vector. Hence, for all $\mathbf{z} \in \mathbb{R}^n$, $[J_f^{-1}(\mathbf{z})]_{\ell,m} \neq 0$. To summarize, if $\pi_m = \sigma_m$, then

$$\forall \mathbf{z} \in \mathbb{R}^n \ \left[J_f^{-1}(\mathbf{z})\right]_{I^m,i} \neq 0 \quad \Longleftrightarrow \quad i = m \ . \tag{226}$$

Next, consider the set of *mismatched nodes*

$$\mathcal{A} \triangleq \{I^m : I^m \neq \tilde{I}^{\pi_m}\} \ . \tag{227}$$

Let $I^a \in \mathcal{A}$ be a non-descendant of all the other nodes in $\mathcal{A}$. There exist nodes $I^b, I^c \in \mathcal{A}$, not necessarily distinct, such that

$$I^a = \tilde{I}^{\pi_b} \ , \quad \text{and} \quad I^c = \tilde{I}^{\pi_a} \ . \tag{228}$$

In four steps, we will show that $[\mathbf{D}(h^*)]_{a,b} \neq 0$ and $[\mathbf{D}(h^*)]_{b,a} \neq 0$, which violates the constraint $\mathbb{1}\{\mathbf{D}(h^*)\} \odot \mathbb{1}\{\mathbf{D}^\top(h^*)\} = \mathbf{I}_{n \times n}$ and will conclude the proof by contradiction. Before giving the steps, we provide the following argument which we repeatedly use in the rest of the proof. For any continuous function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbb{E}\Big[\big|f(\mathbf{Z})\big|\Big] \neq 0 \iff \mathbb{E}\Big[\big|f(\mathbf{Z}) \cdot \big[s(\mathbf{Z}) - s^a(\mathbf{Z})\big]_{I^a}\big|\Big] \neq 0 , \tag{229}$$

$$\text{and} \quad \mathbb{E}\Big[\big|f(\mathbf{Z})\big|\Big] \neq 0 \iff \mathbb{E}\Big[\big|f(\mathbf{Z}) \cdot \big[s(\mathbf{Z}) - \tilde{s}^{\pi_b}(\mathbf{Z})\big]_{I^a}\big|\Big] \neq 0 . \tag{230}$$

First, suppose that $\mathbb{E}\big[|f(\mathbf{Z})|\big] \neq 0$. Then, there exists an open set $\Psi \subseteq \mathbb{R}^n$ for which $f(\mathbf{z}) \neq 0$ for all $z \in \Psi$. Due to the interventional discrepancy condition, there exists an open set within Ψ for which $[s(\mathbf{Z}) - s^a(\mathbf{Z})]_{I^a} \neq 0$. This implies that

$$\mathbb{E}\Big[\big|f(\mathbf{Z}) \cdot \big[s(\mathbf{Z}) - s^a(\mathbf{Z})\big]_{I^a}\big|\Big] \neq 0 \tag{231}$$

For the other direction, suppose that $\mathbb{E}\Big[\big|f(\mathbf{Z}) \cdot \big[s(\mathbf{Z}) - s^a(\mathbf{Z})\big]_{I^a}\big|\Big] \neq 0$, which implies that there exists an open set Ψ for which both $f(\mathbf{z})$ and $[s(\mathbf{z}) - s^a(\mathbf{z})]_{I^a}$ are nonzero. Then, $\mathbb{E}\big[|f(\mathbf{Z})|\big] \neq 0$, and we have (229). Similarly, due to $\tilde{I}^{\pi_b} = I^a$ and interventional discrepancy, we obtain (230).

Step 1: Show that $\mathbb{E}\Big[\big|[J_f^{-1}(\mathbf{Z})]_{I^a,a}\big|\Big] \neq 0$. First, using (40) and Theorem 7(i), we have

$$\big[\mathbf{D}(h^*)\big]_{a,a} = \mathbb{E}\Big[\big|[J_f^{-\top}(\mathbf{Z})]_a \cdot [s(\mathbf{Z}) - s^a(\mathbf{Z})]\big|\Big] \tag{232}$$

$$= \mathbb{E}\Big[\big|\sum_{I^j \in \overline{\mathrm{pa}}(I^a)} [J_f^{-1}(\mathbf{Z})]_{I^j,a} \cdot [s(\mathbf{Z}) - s^a(\mathbf{Z})]_{I^j}\big|\Big] . \tag{233}$$

Note that $\overline{\mathrm{pa}}(I^a) \cap \mathcal{A} = \{I^a\}$ since I^a is non-descendant of the other nodes in $\mathcal{A}$. Consider $I^j \in \mathrm{pa}(I^a)$, which implies that $I^j \notin \mathcal{A}$ and $I^j = \tilde{I}^{\pi_j}$. By (226), we have $[J_f^{-1}(\mathbf{Z})]_{I^j,a} = 0$. Then, (233) becomes

$$\big[\mathbf{D}(h^*)\big]_{a,a} = \mathbb{E}\Big[\big|[J_f^{-1}(\mathbf{Z})]_{I^a,a} \cdot [s(\mathbf{Z}) - s^a(\mathbf{Z})]_{I^a}\big|\Big] \neq 0 , \tag{234}$$

since diagonal entries of $\mathbf{D}(h^*)$ are nonzero due to the last constraint in (92). Then, (229) implies that $\mathbb{E}\Big[\big|[J_f^{-1}(\mathbf{Z})]_{I^a,a}\big|\Big] \neq 0$.

Step 2: Show that $\big[\tilde{\mathbf{D}}(h^*)\big]_{a,b} \neq 0$. Next, we use $I^a = \tilde{I}^{\pi_b}$ and Theorem 7(i) to obtain

$$\big[\tilde{\mathbf{D}}(h^*)\big]_{a,b} = \mathbb{E}\Big[\big|[J_f^{-\top}(\mathbf{Z})]_a \cdot [s(\mathbf{Z}) - \tilde{s}^{\pi_b}(\mathbf{Z})]\big|\Big] \tag{235}$$

$$= \mathbb{E}\Big[\big|\sum_{I^j \in \overline{\mathrm{pa}}(I^a)} [J_f^{-1}(\mathbf{Z})]_{I^j,a} \cdot [s(\mathbf{Z}) - \tilde{s}^{\pi_b}(\mathbf{Z})]_{I^j}\big|\Big] \tag{236}$$

$$= \mathbb{E}\Big[\big|[J_f^{-1}(\mathbf{Z})]_{I^a,a} \cdot [s(\mathbf{Z}) - \tilde{s}^{\pi_b}(\mathbf{Z})]_{I^a}\big|\Big] . \tag{237}$$

Using (230) and Step 1 result, we have $\big[\tilde{\mathbf{D}}(h^*)\big]_{a,b} \neq 0$.

Step 3: Show that $\mathbb{E}\left[\left|[J_f^{-1}(\mathbf{Z})]_{I^a,b}\right|\right] \neq 0.$ Using (40) and Theorem 7(i), we have

$$\left[\tilde{\mathbf{D}}(h^*)\right]_{b,b} = \mathbb{E}\left[\left|\left[J_f^{-1}(\mathbf{Z})\right]_b \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\pi_b}(\mathbf{Z})\right]\right|\right] \tag{238}$$

$$= \mathbb{E}\left[\left|\sum_{I^j \in \overline{\mathrm{pa}}(I^a)} \left[J_f^{-1}(\mathbf{Z})\right]_{I^j,b} \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\pi_b}(\mathbf{Z})\right]_{I^j}\right|\right] \tag{239}$$

$$= \mathbb{E}\left[\left|[J_f^{-1}(\mathbf{Z})]_{I^a,b} \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \tilde{\boldsymbol{s}}^{\rho_c}(\mathbf{Z})\right]_{I^a}\right|\right] . \tag{240}$$

Since $\mathbb{1}\{\mathbf{D}(h^*)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h^*)\}$, the diagonal entry $\left[\tilde{\mathbf{D}}(h^*)\right]_{b,b}$ is nonzero. Then, using (230) we have $\mathbb{E}\left[\left|[J_f^{-1}(\mathbf{Z})]_{I^a,b}\right|\right] \neq 0.$

Step 4: Show that $\left[\mathbf{D}(h^*)\right]_{b,a} \neq 0.$ Next, we use $I^a = \tilde{I}^{\pi_b}$ and Theorem 7(i) to obtain

$$\left[\mathbf{D}(h^*)\right]_{b,a} = \mathbb{E}\left[\left|\left[J_f^{-\top}(\mathbf{Z})\right]_b \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^a(\mathbf{Z})\right]\right|\right] \tag{241}$$

$$= \mathbb{E}\left[\left|\sum_{I^j \in \overline{\mathrm{pa}}(I^a)} \left[J_f^{-1}(\mathbf{Z})\right]_{I^j,b} \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^a(\mathbf{Z})\right]_{I^j}\right|\right] \tag{242}$$

$$= \mathbb{E}\left[\left|\left[J_f^{-1}(\mathbf{Z})\right]_{I^a,b} \cdot \left[\boldsymbol{s}(\mathbf{Z}) - \boldsymbol{s}^a(\mathbf{Z})\right]_{I^a}\right|\right] . \tag{243}$$

Using (229) and Step 3 result, we have $[\mathbf{D}(h^*)]_{b,a} \neq 0.$

Finally, using the constraint $\mathbb{1}\{\mathbf{D}(h^*)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h^*)\}$, Step 2 implies that $[\mathbf{D}(h^*)]_{a,b} \neq 0.$ Combining with Step 4 result, we have $[\mathbf{D}(h^*) \odot \mathbf{D}^\top(h^*)]_{a,b} \neq 0$, which violates the last constraint in (92). Therefore, if (h^*, π) is a minimizer of (92), then π must be the correct coupling σ.

Appendix D. Analysis of the Assumptions

D.1 Analysis of Assumption 1

In this subsection, we prove Theorem 19 statement, i.e., Assumption 1 is satisfied for additive noise models under hard interventions, as follows. Consider a hard interventional environment $\mathcal{E}^m$ and let $I^m = i$ be the intervened node. Recall (29), which becomes

$$[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_k = \left[\boldsymbol{s}(\mathbf{z}) - \boldsymbol{s}^m(\mathbf{z})\right]_k = \frac{\partial}{\partial z_k} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) , \tag{244}$$

for the hard intervention on node i and parent node $k \in \mathrm{pa}(i)$. Then, the ratio in Assumption 1 is given by

$$\frac{[\nabla_{\mathbf{z}} \log p(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})]_k}{[\nabla_{\mathbf{z}} \log p(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})]_i - [\nabla_{\mathbf{z}} \log q(z_i)]_i} = \frac{[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_k}{[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_i} . \tag{245}$$

We will prove the desired result by contradiction. Assume that there exists a nonzero constant $c \in \mathbb{R}$ such that $[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_k = c \cdot [\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_i$ for all $\mathbf{z} \in \mathbb{R}^n.$

Additive noise model. The additive noise model for node i is given by $Z_i = f_i(\mathbf{Z}_{\mathrm{pa}(i)}) + N_i$ as specified in (7). When node i is hard intervened, Z_i is generated according to $Z_i = \bar{N}_i$ in which $\bar{N}_i$ denotes the exogenous noise term for the interventional model. Then, denoting the pdfs of N_i and $\bar{N}_i$ by p_N and q_N, respectively, we have

$$p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = p_N(z_i - f_i(\mathbf{z}_{\mathrm{pa}(i)})) , \quad \text{and} \quad q_i(z_i) = q_N(z_i) . \tag{246}$$

Denote the score functions associated with p_N and q_N by

$$r_p(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log p_N(u) = \frac{p'_N(u)}{p_N(u)} , \quad \text{and} \quad r_q(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log q_N(u) = \frac{q'_N(u)}{q_N(u)} . \tag{247}$$

Define n_i and $\bar{n}_i$ as the realizations of N_i and $\bar{N}_i$ when $Z_i = z_i$ and $\mathbf{Z}_{\mathrm{pa}(i)} = \mathbf{z}_{\mathrm{pa}(i)}$. Then, we have $\bar{n}_i = n_i + f_i(\mathbf{z}_{\mathrm{pa}(i)})$. Using (246) and (247), we can express (245) as

$$\frac{[\boldsymbol{d}^m_{\mathbf{Z}}(\mathbf{z})]_k}{[\boldsymbol{d}^m_{\mathbf{Z}}(\mathbf{z})]_i} = \frac{-\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} \cdot r_p(n_i)}{r_p(n_i) - r_q(n_i + f_i(\mathbf{z}_{\mathrm{pa}(i)}))} . \tag{248}$$

Then, $[\boldsymbol{d}^m_{\mathbf{Z}}(\mathbf{z})]_k = c \cdot [\boldsymbol{d}^m_{\mathbf{Z}}(\mathbf{z})]_i$ for all $\mathbf{z} \in \mathbb{R}^n$ becomes

$$c \cdot \frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} + 1 = \frac{r_q(n_i + f_i(\mathbf{z}_{\mathrm{pa}(i)}))}{r_p(n_i)} . \tag{249}$$

We scrutinize (249) in two cases.

Case 1: $\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}$ **is constant.** In this case, let $\eta = \frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}$ for some $\eta \in \mathbb{R}$ for all $\mathbf{z}_{\mathrm{pa}(i)} \in \mathbb{R}^{|\mathrm{pa}(i)|}$. Then, the RHS of (249) is also constant for all $\mathbf{z}_{\mathrm{pa}(i)} \in \mathbb{R}^{|\mathrm{pa}(i)|}$ and $n_i \in \mathbb{R}$. Fix a realization $n_i = n_i^*$ and note that $f_i(\mathbf{z}_{\mathrm{pa}(i)})$ needs to be constant, denoted by δ^*, for all $\mathbf{z}_{\mathrm{pa}(i)} \in \mathbb{R}^{|\mathrm{pa}(i)|}$. However, this implies that $\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k} = 0$, and we have $r_q(n_i + \delta^*) = r_p(n_i)$ for all $n_i \in \mathbb{R}$. This implies that $p_N(n_i) = \eta^* q_N(n_i + \delta^*)$ for some constant η^*. Since p_N and q_N are pdfs, the only choice is $\eta^* = 1$ and $p_N(n_i) = q_N(n_i + \delta^*)$, then $p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = q_i(z_i)$, which contradicts the premise that an intervention changes the causal mechanism of the target node i.

Case 2: $\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_k}$ **is not constant.** In this case, note that LHS of (249) is not a function of n_i. Then, taking the derivative of both sides with respect to n_i and rearranging, we obtain

$$\frac{r'_p(n_i)}{r_p(n_i)} = \frac{r'_q(n_i + f_i(\mathbf{z}_{\mathrm{pa}(i)}))}{r_q(n_i + f_i(\mathbf{z}_{\mathrm{pa}(i)}))} , \quad \forall (n_i, \mathbf{z}_{\mathrm{pa}(i)}) \in \mathbb{R} \times \mathbb{R}^{|\mathrm{pa}(i)|} . \tag{250}$$

Next, consider a fixed realization $n_i = n_i^*$, and denote the value of LHS by α. Since $f_i(\mathbf{z}_{\mathrm{pa}(i)})$ is continuous and not constant, its image contains an open interval $\Theta \subseteq \mathbb{R}$. Denoting $u \triangleq f_i(\mathbf{z}_{\mathrm{pa}(i)})$, we have

$$\alpha = \frac{r'_q(n_i^* + u)}{r_q(n_i^* + u)} , \quad \forall u \in \Theta . \tag{251}$$

The only solution to this equality is that $r_q(n_i^* + u)$ is an exponential function, $r_q(u) = k_1 \exp(\alpha u)$ over interval $u \in \Theta$. Since r_q is an analytic function that equals to an exponential

function over an interval, it is exponential over entire $\mathbb{R}$. This implies that the pdf $q_N(u)$ is of the form $q_N(u) = k_2 \exp((k_1/\alpha) \exp(\alpha u))$. However, this cannot be a valid pdf since its integral over $\mathbb{R}$ diverges. Then, $[\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_k = c \cdot [\boldsymbol{d}_{\mathbf{Z}}^m(\mathbf{z})]_i$ cannot be true, which concludes the proof.

D.2 Analysis of Assumption 2

In this section, we establish the necessary and sufficient conditions under which Assumption 2 holds for additive models. Furthermore, we show that a large class of nonlinear models in the latent space satisfy these conditions, including the two-layer neural networks. We have focused on such NNs since they effectively approximate continuous functions (Cybenko, 1989). Readily, the necessary and sufficient conditions can be investigated for other choices of nonlinear functions.

D.2.1 INTERPRETING ASSUMPTION 2

The implication of Assumption 2 is that the effect of an intervention is not lost in any linear combination of the varying coordinates of the scores. To formalize this, for each node $i \in [n]$, we define

$$\mathcal{C}_i \triangleq \{\mathbf{c} \in \mathbb{R}^n \ : \ \exists j \in \overline{\mathrm{pa}}(i) \ \text{such that} \ c_j \neq 0\} , \tag{252}$$

Let $I^m = i$. Note that $\mathbf{R}_{\mathbf{Z}}^m$ is positive semi-definite and $[\mathbf{R}_{\mathbf{Z}}^m]_{j,k} = 0$ for all $(j,k) \notin \overline{\mathrm{pa}}(i) \times \overline{\mathrm{pa}}(i)$. Therefore, for any $\mathbf{c} \in \mathcal{C}_i$,

$$\mathbf{c}^\top \cdot \mathbf{R}_{\mathbf{Z}}^m \cdot \mathbf{c} = \mathbf{E}\big[(\mathbf{c}^\top \cdot \boldsymbol{d}_{\mathbf{Z}}^m) \cdot (\mathbf{c}^\top \cdot \boldsymbol{d}_{\mathbf{Z}}^m)^\top\big] \neq 0 , \tag{253}$$

since $\mathbf{R}_{\mathbf{Z}}^m$ has rank $|\overline{\mathrm{pa}}(i)|$. The reverse direction also holds true, i.e., if $\mathbf{R}_{\mathbf{Z}}^m$ has rank less than $|\overline{\mathrm{pa}}(i)|$, then there exists $\mathbf{c}\mathcal{C}_i$ which makes $\mathbf{c}^\top \cdot \mathbf{R}_{\mathbf{Z}}^m \cdot \mathbf{c} = 0$. Therefore, Assumption 2 is equivalent to the following statement:

$$\mathbf{E}\big[|\mathbf{c}^\top \cdot \boldsymbol{d}_{\mathbf{Z}}^m|\big] \neq 0 , \quad \forall \mathbf{c} \in \mathcal{C}_i . \tag{254}$$

Necessary and sufficient conditions. Consider the additive noise model for node i

$$Z_i = f_i(\mathbf{Z}_{\mathrm{pa}(i)}) + N_i , \tag{255}$$

as specified in (7). When node i is soft intervened, Z_i is generated according to

$$Z_i = \bar{f}_i(\mathbf{Z}_{\mathrm{pa}(i)}) + \bar{N}_i , \tag{256}$$

in which $\bar{f}_i$ and $\bar{N}_i$ specify the interventional mechanism for node i. The following lemma characterizes the necessary and sufficient conditions under which (254) (equivalently, Assumption 2) is satisfied. In this subsection, we use φ as the shorthand for $\mathbf{z}_{\mathrm{pa}(i)}$.

Lemma 34 *For each node $i \in [n]$ consider the following two set of equations for $c \in \mathbb{R}^n$:*

$$\begin{cases} c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = 0 \\ \\ c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) = 0 \end{cases} , \qquad \forall \varphi \in \mathbb{R}^{|\mathrm{pa}(i)|} . \tag{257}$$

Assumption 2 holds if and only if the only all solutions $\mathbf{c}$ to (257) satisfy $\mathbf{c} \notin \mathcal{C}_i$, or based on (252), equivalently $c_j = 0$ for all $j \in \overline{\mathrm{pa}}(i)$.

Proof: See Appendix D.2.2.

To provide some intuition about the conditions in Theorem 34, we consider a node $i \in [n]$ and discuss the conditions in the context of a few examples. Note that by sweeping $\varphi \in \mathbb{R}^{|\mathrm{pa}(i)|}$ we generate a continuum of linear equations of the form:

$$c_i - \mathbf{c}^\top \cdot \nabla_\mathbf{z} f_i(\varphi) = 0 , \qquad \text{and} \qquad c_i - \mathbf{c}^\top \cdot \nabla_\mathbf{z} \bar{f}_i(\varphi) = 0 . \qquad (258)$$

Note that for all $j \notin \mathrm{pa}(i)$ we have $[\nabla_\mathbf{z} f_i(\varphi)]_j = [\nabla_\mathbf{z} \bar{f}_i(\varphi)]_j = 0$. Hence, in finding the solutions to (258) only the coordinates $\{j \in \overline{\mathrm{pa}}(i)\}$ of $\mathbf{c}$ are relevant. Let us define

$$u(\varphi) \triangleq \nabla_\varphi f_i(\varphi) , \quad \text{and} \quad \bar{u}(\varphi) \triangleq \nabla_\varphi \bar{f}_i(\varphi) , \qquad (259)$$

which are the gradients of f_i and f_i by considering only the coordinates of $\mathbf{z}$ in $\{j \in \mathrm{pa}(i)\}$. Accordingly, we also define $\mathbf{b}$ by concatenating only the coordinates of $\mathbf{c}$ with their indices in $\{j \in \mathrm{pa}(i)\}$. Next, consider w distinct choices of φ and denote them by $\{\varphi^t \in \mathbb{R}^{|\mathrm{pa}(i)|} : t \in [w]\}$. By concatenating the two equations in (259) specialized to these realizations, we get the following linear system with $2w$ equations and $|\mathrm{pa}(i)| + 1$ unknown variables.

$$\underbrace{\begin{bmatrix} [u(\varphi^1)]^\top & -1 \\ [\bar{u}(\varphi^1)]^\top & -1 \\ \vdots & \vdots \\ [u(\varphi^w)]^\top & -1 \\ [\bar{u}(\varphi^w)]^\top & -1 \end{bmatrix}}_{\triangleq \mathbf{V} \in \mathbb{R}^{2w \times (|\mathrm{pa}(i)|+1)}} \begin{bmatrix} \mathbf{b} \\ c_i \end{bmatrix} = \mathbf{0}_{2w} . \qquad (260)$$

When $\mathbf{V}$ is full-rank, i.e., $\mathrm{rank}(\mathbf{V}) = |\mathrm{pa}(i)| + 1$, the system has only the trivial solutions $c_i = 0$ and $\mathbf{b} = \mathbf{0}$. Then, we make the following observations.

1. If f_i and $\bar{f}_i$ are linear functions, the vector spaces generated by u and $\bar{u}$ have dimensions 1. Subsequently, we always have $\mathrm{rank}(\mathbf{V}) \leq 2$, rendering an underdetermined system when $|\mathrm{pa}(i)| \geq 2$. Hence, when the maximum degree of $\mathcal{G}$ is at least 2, a linear causal model does not satisfy Assumption 2.

2. If f_i and $\bar{f}_i$ are quadratic with full-rank matrices, i.e., $f_i(\varphi) = \varphi^\top \mathbf{A} \varphi$ and $\bar{f}_i(\varphi) = \varphi^\top \bar{\mathbf{A}} \varphi$ where $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\bar{\mathbf{A}}) = |\mathrm{pa}(i)|$, there is a choice of $w \geq |\mathrm{pa}(i)| + 1$ and realizations $\{\varphi^t \in \mathbb{R}^{|\mathrm{pa}(i)|} : t \in [w]\}$ for which $\mathrm{rank}(\mathbf{V}) = |\mathrm{pa}(i)| + 1$ and the system in (258) admits only the trivial solutions $a_i = 0$ and $\mathbf{b} = \mathbf{0}$. Hence, quadratic causal models satisfy Assumption 2.

3. If f_i and $\bar{f}_i$ are two-layer NNs with a sufficiently large number of hidden neurons, they also render a fully determined system, and as a result, they satisfy Assumption 2.

We investigate the last example in detail as follows. Assume that f_i and $\bar{f}_i$ are two-layer NNs with $|\mathrm{pa}(i)|$ inputs, w_i and $\bar{w}_i$ hidden nodes, respectively, and with sigmoid activation functions. Denote the weight matrices between input and hidden layers in f_i and $\bar{f}_i$ by $\mathbf{W}^i \in \mathbb{R}^{w_i \times |\mathrm{pa}(i)|}$ and $\bar{\mathbf{W}}^i \in \mathbb{R}^{\bar{w}_i \times |\mathrm{pa}(i)|}$, respectively. Furthermore, define $\boldsymbol{\nu} \in \mathbb{R}^{w_i}$ and

$\bar{\boldsymbol{\nu}} \in \mathbb{R}^{w_i}$ as the weights between the hidden layer and output in f_i and $\bar{f}_i$, respectively. Finally, define ν_0 and $\bar{\nu}_0$ as the bias terms. Hence, we have

$$f_i(\varphi) = \boldsymbol{\nu}^\top \cdot \sigma(\mathbf{W}^i \cdot \varphi) + \nu_0 = \sum_{j=1}^{w_i} \nu_j \cdot \sigma(\mathbf{W}_j^i \cdot \varphi) + \nu_0 \ , \tag{261}$$

$$\text{and} \quad \bar{f}_i(\varphi) = \bar{\boldsymbol{\nu}}^\top \cdot \sigma(\bar{\mathbf{W}}^i \cdot \varphi) + \nu_0 = \sum_{j=1}^{\bar{w}_i} \bar{\nu}_j \cdot \sigma(\bar{\mathbf{W}}_j^i \cdot \varphi) + \bar{\nu}_0 \ , \tag{262}$$

in which activation function σ is applied element-wise.

Theorem 20   Consider NNs f_i and $\bar{f}_i$ specified in (261) and (262). If for all $i \in [n]$ we have

$$\max\{\text{rank}(\mathbf{W}^i) \, , \ \text{rank}(\bar{\mathbf{W}}^i)\} = |\text{pa}(i)| \ , \tag{263}$$

then Assumption 2 holds.

Proof: See Appendix D.2.3.

D.2.2   PROOF OF LEMMA 34

We show that for node $I^m = i$, the condition in (254)

$$\mathbb{E}\Big[\big|\mathbf{c}^\top \cdot \boldsymbol{d}_{\mathbf{Z}}^m\big|\Big] \neq 0 \ , \ \ \forall c \in \mathcal{C}_i \ , \tag{264}$$

holds if and only if the following continuum of equations admit their solutions $\mathbf{c}$ in $\mathbb{R}^n \setminus \mathcal{C}_i$.

$$\begin{cases} c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = 0 \\[2mm] c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) = 0 \end{cases} , \qquad \forall \varphi \in \mathbb{R}^{|\text{pa}(i)|} \ , \tag{265}$$

in which shorthand φ is used for $\mathbf{z}_{\text{pa}(i)}$. We first note that the condition in (264) can be equivalently stated using (18) and (19) as

$$\mathbb{E}\Big[\big|\mathbf{c}^\top \cdot \big[\nabla_{\mathbf{z}} \log p_i(Z_i \mid \mathbf{Z}_{\text{pa}(i)}) - \nabla_{\mathbf{z}} \log q_i(Z_i \mid \mathbf{Z}_{\text{pa}(i)})\big]\big|\Big] \neq 0 \ , \quad \forall \mathbf{c} \in \mathcal{C}_i \ . \tag{266}$$

Also, using Theorem 31, (266) is equivalent to

$$\forall \mathbf{c} \in \mathcal{C}_i, \ \ \exists \mathbf{z}: \ \ \ \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) \neq \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \log q_i(z_i \mid \mathbf{z}_{\text{pa}(i)}) \ . \tag{267}$$

The additive noise model for node i is given by

$$Z_i = f_i(\mathbf{Z}_{\text{pa}(i)}) + N_i \ , \tag{268}$$

as specified in (7). When node i is soft intervened, Z_i is generated according to

$$Z_i = \bar{f}_i(\mathbf{Z}_{\text{pa}(i)}) + \bar{N}_i \ , \tag{269}$$

in which $\bar{f}_i$ and $\bar{N}_i$ specify the interventional mechanism for node i. Then, denoting the pdfs of N_i and $\bar{N}_i$ by p_N and q_N, respectively, (268) and (269) imply that

$$p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = p_N\big(z_i - f_i(\mathbf{z}_{\mathrm{pa}(i)})\big) , \quad \text{and} \quad q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = q_N\big(z_i - \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})\big) . \tag{270}$$

Denote the score functions associated with p_N and q_N by

$$r_p(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log p_N(u) = \frac{p'_N(u)}{p_N(u)} , \quad \text{and} \quad r_q(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log q_N(u) = \frac{q'_N(u)}{q_N(u)} . \tag{271}$$

Define n_i and $\bar{n}_i$ as the realizations of N_i and $\bar{N}_i$ when $Z_i = z_i$ and $\mathbf{Z}_{\mathrm{pa}(i)} = \mathbf{z}_{\mathrm{pa}(i)}$. By defining $\delta(\mathbf{z}_{\mathrm{pa}(i)}) \triangleq f_i(\mathbf{z}_{\mathrm{pa}(i)}) - \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})$, we have $\bar{n}_i = n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)})$. Using (271) and (270), we can express the relevant entries of $\nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$ and $\nabla_{\mathbf{z}} \log q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$ as

$$\big[\nabla_{\mathbf{z}} \log p(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})\big]_j = \begin{cases} r_p(n_i) , & j = i , \\ -\frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_j} r_p(n_i) , & j \in \mathrm{pa}(i) , \\ 0 , & j \notin \overline{\mathrm{pa}}(i) , \end{cases} \tag{272}$$

$$\text{and} \quad \big[\nabla_{\mathbf{z}} \log q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})\big]_j = \begin{cases} r_q(\bar{n}_i) , & j = i , \\ -\frac{\partial \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_j} r_q(\bar{n}_i) , & j \in \mathrm{pa}(i) , \\ 0 , & j \notin \overline{\mathrm{pa}}(i) . \end{cases} \tag{273}$$

By substituting (272)–(273) in (267) and rearranging the terms, the statement in (267) becomes equivalent to following statement. For all $\mathbf{c} \in \mathcal{C}_i$, there exist $n_i \in \mathbb{R}$ and $\mathbf{z}_{\mathrm{pa}(i)} \in \mathbb{R}^{|\mathrm{pa}(i)|}$ such that

$$r_p(n_i) \cdot \left(c_i - \sum_{j \in \mathrm{pa}(i)} c_j \cdot \frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_j}\right) \neq r_q(n_i + \delta(\mathbf{z}_{\mathrm{pa}(i)})) \cdot \left(c_i - \sum_{j \in \mathrm{pa}(i)} c_j \cdot \frac{\partial \bar{f}_i(\mathbf{z}_{\mathrm{pa}(i)})}{\partial z_j}\right) , \tag{274}$$

which by using the shorthand φ for $\mathbf{z}_{\mathrm{pa}(i)}$ can be compactly presented as follows. For all $\mathbf{c} \in \mathcal{C}_i$, there exists $n_i \in \mathbb{R}$ and $\varphi \in \mathbb{R}^{|\mathrm{pa}(i)|}$ such that

$$r_p(n_i) \cdot \left[c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi)\right] \neq r_q\big(n_i + \delta(\varphi)\big) \cdot \left[c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi)\right] . \tag{275}$$

Hence, Assumption 2 is equivalent to the statement in (275), which we use for the rest of the proof.

Sufficient condition. We show that if (265) admit solutions $\mathbf{c}$ only in $\mathbb{R}^n \setminus \mathcal{C}_i$, then the statement in (275) holds. By contradiction, assume that there exists $\mathbf{c}^* \in \mathcal{C}_i$ such that for all $n_i \in \mathbb{R}$ and $\varphi \in \mathbb{R}^{|\mathrm{pa}(i)|}$,

$$r_p(n_i) \cdot \left[c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi)\right] = r_q\big(n_i + \delta(\varphi)\big) \cdot \left[c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi)\right] . \tag{276}$$

We show that $\mathbf{c}^* \in \mathcal{C}_i$ is also a solution to (265), contradicting the premise. In order to show that $\mathbf{c}^* \in \mathcal{C}_i$ is also a solution to (265), suppose, by contradiction, that (276) holds, and there exists $\varphi^* \in \mathbb{R}^{|\mathrm{pa}(i)|}$ corresponding to which

$$(\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi^*) - c_i^* \neq 0 . \tag{277}$$

Note that f_i is a continuously differentiable function and also a function of z_j for all $j \in \mathrm{pa}(i)$. Hence, there exists an open set $\Phi \subseteq \mathbb{R}^{|\mathrm{pa}(i)|}$ for φ for which $(\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi^*) - c_i^*$ is nonzero everywhere in Φ. Likewise, r_p cannot constantly be zero over all possible intervals Ω. This is because otherwise, it would have to necessarily be a constant zero function (since it is analytic), which is an invalid score function. Hence, there exists an open interval $\Omega \subseteq \mathbb{R}$ over which $r_p(n_i)$ is nonzero for all $n_i \in \Omega$. Subsequently, the left-hand side of (276) is nonzero over the Cartesian product $(n_i, \varphi) \in \Omega \times \Phi$. This means that if (276) is true, then both functions on its right-hand side must also be nonzero over $\Omega \times \Phi$. Hence, by rearranging the terms in (276) we have

$$\frac{r_q(n_i + \delta(\varphi))}{r_p(n_i)} = \frac{c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi)}{c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi)} \ , \qquad \forall (n_i, \varphi) \in \Omega \times \Phi \ . \tag{278}$$

In two steps, we show that (278) cannot be valid.

Step 1. First, we show that function δ cannot be a constant over Φ (interval specified above). Suppose the contrary and assume that $\delta(\varphi) = \delta^*$ for all $\varphi \in \Phi$. Hence, the gradient of δ is zero. Using the definition of δ, this implies that

$$\nabla_{\mathbf{z}} f_i(\varphi) = \nabla_{\mathbf{z}} \bar{f}_i(\varphi) \ , \quad \forall \varphi \in \Phi \ . \tag{279}$$

Then, by leveraging (278), we conclude that $r_p(n_i) = r_q(n_i + \delta^*)$ for all $n_i \in \Omega$. Since $r_p(n_i)$ and $r_q(n_i + \delta^*)$ are analytic functions that agree on an open interval of $\mathbb{R}$, they are equal for all $n_i \in \mathbb{R}$ as well. This implies that $p_N(n_i) = \eta \cdot q_N(n_i + \delta^*)$ for some constant η. Since p_N and q_N are pdfs, the only choice is $\eta = 1$ and $p_N(n_i) = q_N(n_i + \delta^*)$, then $p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) = q_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)})$, which contradicts the premise that observational and interventional mechanisms are distinct.

Step 2. Finally, we will show that (278) cannot be true when δ is not a constant function over Φ. Note that the right-hand side of (278) is not a function of n_i. Then, taking the derivative of both sides with respect to n_i and rearranging, we obtain

$$\frac{r_p'(n_i)}{r_p(n_i)} = \frac{r_q'(n_i + \delta(\varphi))}{r_q(n_i + \delta(\varphi))} \ , \quad \forall (n_i, \varphi) \in \Omega \times \Phi \ . \tag{280}$$

Next, consider a fixed realization $n_i = n_i^*$, and denote the value of LHS by α. Since δ is continuous and not constant over Φ, its image contains an open interval $\Theta \subseteq \mathbb{R}$. Denoting $u \triangleq \delta(\varphi)$, we get

$$\alpha = \frac{r_q'(n_i^* + u)}{r_q(n_i^* + u)} \ , \quad \forall u \in \Theta \ . \tag{281}$$

The only solution to this equality is that $r_q(n_i^* + u)$ is an exponential function, $r_q(u) = k_1 \exp(\alpha u)$ over interval $u \in \Theta$. Since r_q is an analytic function that equals to an exponential function over an interval, it is exponential over entire $\mathbb{R}$. This implies that the pdf $q_N(u)$ is of the form $q_N(u) = k_2 \exp((k_1/\alpha) \exp(\alpha u))$. However, this cannot be a valid pdf since its integral over $\mathbb{R}$ diverges. Hence, (278) is not true, and the premise that there exists $\mathbf{c}^* \in \mathcal{C}_i$ and $\varphi \in \mathbb{R}^{|\mathrm{pa}(i)|}$ corresponding to which (277) holds is invalid, concluding that for all

$\varphi \in \mathbb{R}^{|\text{pa}(i)|}$ we have $\mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) = c_i$. Proving the counterpart identity $\mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = c_i$ follows similarly. Therefore, (276) implies

$$
\begin{cases}
c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = 0 \\
\\
c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) = 0
\end{cases}
, \qquad \forall \varphi \in \mathbb{R}^{|\text{pa}(i)|} ,
\tag{282}
$$

which means that we have found a solution to (265) that is not in $\mathbb{R}^n \setminus \mathcal{C}_i$, contradicting the premise.

Necessary condition. Assume that Assumption 2, and equivalently, the statement in (275) holds but (265) has a solution $\mathbf{c}^*$ in $\mathcal{C}_i$. Then,

$$
c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) = 0 , \quad \forall \varphi \in \mathbb{R}^{|\text{pa}(i)|} .
\tag{283}
$$

Multiplying the left side by $r_p(n_i)$ and the right side by $r_q(n_i + \delta(\varphi))$, we obtain that for all $n_i \in \mathbb{R}$ and $\varphi \in \mathbb{R}^{|\text{pa}(i)|}$

$$
r_p(n_i) \cdot \left[c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) \right] = r_q(n_i + \delta(\varphi)) \cdot \left[c_i^* - (\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) \right] = 0 ,
\tag{284}
$$

which implies that for all $\mathbf{c} \in \mathcal{C}_i$, $n_i \in \mathbb{R}$ and $\varphi \in \mathbb{R}^{|\text{pa}(i)|}$

$$
r_p(n_i) \cdot \left[c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) \right] = r_q(n_i + \delta(\varphi)) \cdot \left[c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) \right] .
\tag{285}
$$

This contradicts (275), and equivalently Assumption 2. Hence, the proof is complete.

D.2.3   PROOF OF LEMMA 20

Approach. We will use the same argument as at the beginning of the proof of Theorem 34. Specifically, we will show that for any node i and two-layer NNs f_i and $\bar{f}_i$ with weight matrices $\mathbf{W}^i$ and $\bar{\mathbf{W}}^i$, the following continuum of equations admit their solutions $\mathbf{c}$ in $\mathbb{R}^n \setminus \mathcal{C}_i$.

$$
\begin{cases}
c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} \bar{f}_i(\varphi) = 0 \\
\\
c_i - \mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = 0
\end{cases}
, \qquad \forall \varphi \in \mathbb{R}^{|\text{pa}(i)|} ,
\tag{286}
$$

in which shorthand φ is used for $\mathbf{z}_{\text{pa}(i)}$. Hence, by invoking Theorem 34, Assumption 2 holds.

Definitions. Define $d_i \triangleq |\text{pa}(i)|$. Since $\max\{\text{rank}(\mathbf{W}^i), \text{rank}(\bar{\mathbf{W}}^i)\} = d_i$, without loss of generality, suppose that $\mathbf{W}^i \in \mathbb{R}^{w_i \times d_i}$ has rank d_i. The rest of the proof follows similarly for the case of $\text{rank}(\bar{\mathbf{W}}^i) = d_i$. We use shorthand $\{\mathbf{W}, w\}$ for $\{\mathbf{W}^i, w_i\}$ when it is obvious from context.

Parameterization. Note that f_i can be represented by different parameterizations, some containing more hidden nodes than others. Without loss of generality, let w be the fewest number of nodes that can represent f_i. This implies that the entries of $\boldsymbol{\nu}$ are nonzero. Otherwise, if $\nu_i = 0$, we can remove i-th hidden node and still have the same f_i. Similarly, the rows of $\mathbf{W}$ are distinct. Otherwise, if there exist rows $\mathbf{W}_i = \mathbf{W}_j$ for distinct $i, j \in [w]$,

removing j-th hidden node and using $(\nu_i + \nu_j)$ in place of ν_i results the same function as f_i with $(w-1)$ hidden nodes. Similarly, we have $\mathbf{W}_i \neq \mathbf{0}$ for all $i \in [w]$. Otherwise, we have

$$\nu_i \cdot \sigma(\mathbf{W}_i \cdot \varphi) + \nu_0 = (\nu_0 + \frac{\nu_i}{2}) \ , \tag{287}$$

and by removing i-th hidden node and using $(\nu_0 + \frac{\nu_i}{2})$ instead of ν_0, we reach the same function as f_i with $(w-1)$ hidden nodes. Finally, we have $\mathbf{W}_i + \mathbf{W}_j \neq \mathbf{0}$. Otherwise, we have

$$\nu_i \cdot \sigma(\mathbf{W}_i \varphi) + \nu_j \sigma(\mathbf{W}_j \cdot \varphi) + \nu_0 = (\nu_i - \nu_j) \cdot \sigma(\mathbf{W}_i \cdot \varphi) + (\nu_0 + \nu_j) \ , \tag{288}$$

and by removing j-th hidden node and using $(\nu_i - \nu_j)$ instead of ν_i and $(\nu_0 + \nu_j)$ instead of ν_0, we reach the same function as f_i with $(w-1)$ hidden nodes. In summary, we have

$$\mathbf{W}_i \neq \mathbf{0} \ , \quad \nu_i \neq 0 \ , \quad \forall i \in [w] \ , \quad \text{and} \quad \mathbf{W}_i \pm \mathbf{W}_j \neq \mathbf{0} \ , \quad \forall i, j \in [w] : i \neq j \ . \tag{289}$$

We will show that there does not exist $\mathbf{c} \in \mathcal{C}_i$ such that $\mathbf{c}^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = c_i$ for all $\varphi \in \mathbb{R}^{d_i}$. Assume the contrary, and assume there exist $\mathbf{c}^* \in \mathcal{C}_i$ such that

$$(\mathbf{c}^*)^\top \cdot \nabla_{\mathbf{z}} f_i(\varphi) = c_i^* \ , \quad \forall \varphi \in \mathbb{R}^{d_i} \ . \tag{290}$$

This is equivalent to showing that there exists nonzero $\mathbf{b}^* \in \mathbb{R}^{d_i}$ such that

$$(\mathbf{b}^*)^\top \cdot \nabla_\varphi f_i(\phi) = c_i^* \ , \quad \forall \varphi \in \mathbb{R}^{d_i} \ . \tag{291}$$

Based on (261), the gradient of $f_i(\varphi)$ is

$$\nabla_\varphi f_i(\varphi) = \mathbf{W}^\top \cdot \mathrm{diag}(\boldsymbol{\nu}) \cdot \dot{\sigma}(W \cdot \varphi) \ , \tag{292}$$

where $\mathrm{diag}(\boldsymbol{\nu})$ is the diagonal matrix with $\boldsymbol{\nu}$ as its diagonal elements, and $\dot{\sigma}$ is the derivative of the sigmoid function, applied element-wise to its argument. Hence, based on (290), the contradiction premise is equivalent to having a nonzero $\mathbf{b}^* \in \mathbb{R}^{d_i}$ such that

$$\left[\dot{\sigma}(\mathbf{W} \cdot \varphi)\right]^\top \cdot \mathrm{diag}(\boldsymbol{\nu}) \cdot \mathbf{W} \cdot \mathbf{b}^* = c_i \ , \quad \forall \varphi \in \mathbb{R}^{d_i} \ . \tag{293}$$

We note that since $\mathbf{W}$ is full-rank and ν_i is nonzero for all $i \in [w]$, the matrix $\mathrm{diag}(\boldsymbol{\nu}) \cdot \mathbf{W}$ is full-rank as well and it has a trivial null space. Subsequently, $\mathbf{b}^* \in \mathbb{R}^{d_i}$ is nonzero if and only if $(\mathrm{diag}(\boldsymbol{\nu}) \cdot \mathbf{W} \cdot \mathbf{b}^*) \in \mathbb{R}^w$ is nonzero. We will use the following lemma to show that (293) cannot be true. This establishes that the contradiction premise is not correct, and completes the proof.

Lemma 35 *Let $\mathbf{u} \in \mathbb{R}^p$ have nonzero entries with distinct absolute values, i.e., $u_i \neq 0$ and $|u_i| \neq |u_j|$ for all $i \neq j$, and $\alpha \in \mathbb{R}$ be a constant. Then, for every nonzero vector $\mathbf{c} \in \mathbb{R}^p$, there exists $\alpha \in \mathbb{R}$ such that $[\dot{\sigma}(\alpha \mathbf{u})]^\top \cdot \mathbf{c} \neq a$.*

Proof: See Appendix D.2.4.

Let us define $\xi \triangleq \mathbf{W} \cdot \varphi$. We will show that there exists $\varphi^* \in \mathbb{R}^{d_i}$ such that $\xi = \mathbf{W} \cdot \varphi^*$ satisfies the conditions in Theorem 35. Then, using Theorem 35 with the choice of $\mathbf{u} = \mathbf{W} \cdot \varphi^*$, $d = \mathrm{diag}(\boldsymbol{\nu}) \cdot \mathbf{W} \cdot \mathbf{b}^*$, and $a = c_i$, we find that there exists $\alpha \in \mathbb{R}$ such that

$$\left[\dot{\sigma}(\alpha \cdot \mathbf{W} \cdot \varphi^*)\right]^\top \cdot \mathrm{diag}(\boldsymbol{\nu}) \cdot \mathbf{W} \cdot \mathbf{b}^* \neq c_i \ . \tag{294}$$

Hence, (293) is false since it is violated for $\varphi = \alpha\varphi^*$ and the proof is completed. We show the existence of such φ^* as follows. We first construct the set of φ values for which conditions of Theorem 35 on w are *not* satisfied. The set in question is the union of the following cases: (i) $\xi_i = \mathbf{W}_i \cdot \varphi = 0$ for some $i \in [w]$, (ii) $|\xi_i| = |\xi_j|$ for some distinct $i, j \in [w]$, or equivalently, $(\mathbf{W}_i \pm \mathbf{W}_j)\varphi = 0$. Note that $\mathbf{W}_i \neq \mathbf{0}$ and $\mathbf{W}_i \pm \mathbf{W}_j \neq \mathbf{0}$ by (289). For a nonzero vector $\mathbf{y} \in \mathbb{R}^{d_i}$, the set $\{\varphi \in \mathbb{R}^{d_i} : \mathbf{y}^\top\varphi = 0\}$ is a $(d_i - 1)$-dimensional subspace of $\mathbb{R}^{d_i}$. Then, there is w number of $(d_i - 1)$-dimensional subspaces that fall under case (i), and $w(w - 1)$ number of $(d_i - 1)$-dimensional subspaces that fall under case (ii). Therefore, there are w^2 lower-dimensional subspaces for which the conditions of Theorem 35 do not hold. However, $\mathbb{R}^{d_i}$ cannot be covered by a finite number of lower-dimensional subspaces of itself. Therefore, there exists $\varphi^* \in \mathbb{R}^{d_i}$ such that $\xi = \mathbf{W} \cdot \varphi^*$ satisfies the conditions of Theorem 35, and the proof is completed.

D.2.4 PROOF OF LEMMA 35

Assume the contrary and suppose that there exists a nonzero $\mathbf{c}$ and α for a given $\mathbf{u}$. Define the function $g_u(\alpha) \triangleq \mathbf{c}^\top \cdot \dot{\sigma}(\alpha\mathbf{u})$. Note that

$$\dot{\sigma}(x) = \frac{1}{1 + e^{-x} + e^x} \tag{295}$$

is an even analytic function, and its Taylor series expansion at 0 has the domain of convergence $\{x \in \mathbb{R} : |x| < \frac{\pi}{2}\}$. Thus, for all $\alpha \in (-\frac{\pi}{2\max_i |u_i|}, \frac{\pi}{2\max_i |u_i|})$,

$$a = g_u(\alpha) = \sum_{j=1}^{p} c_j \cdot \dot{\sigma}(\alpha u_j) = \sum_{i=0}^{\infty} \gamma_i \sum_{j=1}^{p} c_j(\alpha u_j)^{2i} = \sum_{i=0}^{\infty}(\gamma_i \sum_{j=1}^{p} c_j u_j^{2i})\alpha^{2i} . \tag{296}$$

Note that $g_u(\alpha)$ is a constant function of $\alpha \in (-\frac{\pi}{2\max_i |u_i|}, \frac{\pi}{2\max_i |u_i|})$. Thus, its Taylor coefficients, i.e., $(\gamma_i \sum_{j=1}^{p} c_j u_j^{2i})$, are zero for all $i \in \mathbb{N}^+$. However, the coefficients of even powers in Taylor expansion of $\dot{\sigma}$, i.e., γ_i's, are nonzero (Weisstein, 2002). Therefore, we have $\sum_{i=1}^{p}(c_j u_j^{2i}) = 0$ for all $i \in \mathbb{N}^+$. Next, construct the following system of linear equations,

$$\begin{bmatrix} u_1^2 & u_1^4 & \dots & u_1^{2w} \\ \vdots & \vdots & \dots & \vdots \\ u_p^2 & u_p^4 & \dots & u_p^{2w} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} = \mathbf{0} . \tag{297}$$

This is equivalent to

$$\text{diag}([u_1^2, \dots, u_p^2]^\top) \cdot \underbrace{\begin{bmatrix} 1 & u_1^2 & \dots & u_1^{2(p-1)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & u_p^2 & \dots & u_p^{2(p-1)} \end{bmatrix}}_{\text{Vandermonde}} \cdot \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} = \mathbf{0} . \tag{298}$$

Note that the Vandermonde matrix in (298) has determinant $\prod_{1 \leq i \leq j < p}(u_i^2 - u_j^2)$, which is nonzero since $|u_i| \neq |u_j|$ for $i \neq j$. Multiplying an invertible matrix with a diagonal invertible matrix generates another invertible matrix, and $\mathbf{c}$ must be a zero vector which is a contradiction. Hence, there does not exist such $\mathbf{c}$ for which $\mathbf{c}^\top \cdot \dot{\sigma}(\alpha\mathbf{u})$ is constant for every $\alpha \in \mathbb{R}$.

D.3 Assumptions of (Zhang et al., 2023)

We clarify that Assumption 2 of Zhang et al. (2023), which is referred to as "linear interventional faithfulness", *requires* nonlinear SCMs. Subsequently, both of their main results (Theorem 1 and Theorem 2) require this assumption. First, note that Zhang et al. (2023, p.6) implicitly explain that their results are for nonlinear SCMs: *"In general, we show in Appendix B that a large class of nonlinear SCMs and soft interventions satisfy this assumption."* Next, we show that this assumption is violated for linear additive noise models as follows. First, we quote their assumption:

Linear interventional faithfulness (Zhang et al., 2023, Assumption 2): Intervention I on node i satisfies linear interventional faithfulness if for every $j \in \text{ch}(i) \cup \{i\}$ such that $\text{pa}(j) \cap \text{de}(i) = \emptyset$, it holds that $\mathbb{P}(Z_j + \mathbf{c}^\top \mathbf{Z}_S) \neq \mathbb{P}^I(Z_j + \mathbf{c}^\top \mathbf{Z}_S)$ for all constant vectors $\mathbf{c} \in \mathbb{R}^{|S|}$ where $S = [n] \setminus \text{de}(i)$.

Linear SCMs violate linear interventional faithfulness. Next, we show that linear SCMs violate the above linear interventional faithfulness assumption. Consider an intervention on node i, and let $j \in \text{ch}(i)$ such that $\text{pa}(i) \cap \text{de}(i) = \emptyset$. Let

$$Z_j = \mathbf{w}^\top \mathbf{Z}_{\text{pa}(j)} + N_j \ , \tag{299}$$

where $\mathbf{w} \in \mathbb{R}^{|\text{pa}(j)|}$ denotes the weight vector. Note that we have $\text{pa}(j) \subseteq S = [n] \setminus \text{de}(i)$ since $\text{pa}(i) \cap \text{de}(i) = \emptyset$. Consider vector $\mathbf{c} \in \mathbb{R}^{|S|}$ such that $\mathbf{c}_{\text{pa}(j)} = -\mathbf{W}$ and $\mathbf{c}_{S \setminus \text{pa}(j)} = \mathbf{0}$ where $\mathbf{c}_{\text{pa}(j)}$ denotes the entries corresponding to nodes in $\text{pa}(j)$. Then, we have

$$Z_j + \mathbf{c}^\top \mathbf{Z}_S = N_j \ , \tag{300}$$

which remains invariant after an intervention on node i and yields $\mathbb{P}(Z_j + \mathbf{c}^\top \mathbf{Z}_S) = \mathbb{P}^I(Z_j + \mathbf{c}^\top \mathbf{Z}_S)$. Therefore, linear SCMs violate linear interventional faithfulness.

Appendix E. Empirical Evaluations: Details and Additional Results

E.1 Implementation Details of LSCALE-I Algorithm

Preprocessing: Dimensionality reduction. Since $\mathbf{X} = \mathbf{G} \cdot \mathbf{Z}$, we can compute $\text{im}(\mathbf{G})$ using n random samples of $\mathbf{X}$ almost surely. Specifically, $\text{im}(\mathbf{G})$ equals the column space of the sample covariance matrix of n samples of $\mathbf{X}$. In LSCALE-I, as a preprocessing step, we compute this subspace and express samples of $\mathbf{X}$ and $\mathbf{s_X}$ in this basis. This procedure effectively reduces the dimension of $\mathbf{X}$ and $\mathbf{s_X}$ from d to n. Then, we perform steps of LSCALE-I using this n dimensional observed data.

Implementation details of Algorithm 1. In Stage L1 of Algorithm 1, choosing any $\mathbf{h} \in \text{col}(\mathbf{R_X^m})$ suffices. To make our algorithm deterministic, we choose the top eigenvector of $\text{col}(\mathbf{R_X^m})$ that corresponds to the largest eigenvalue. In Stage L2, we use a nonzero threshold $\lambda_{\mathcal{G}}$, i.e.,

$$\hat{\text{pa}}(m) \triangleq \left\{ i \neq m : \mathbb{E}\left[\left| d_{\hat{\mathbf{Z}}}^m(\hat{\mathbf{Z}}; \hat{\mathbf{H}}) \right|_i \right] \geq \lambda_{\mathcal{G}} \right\} \ . \tag{301}$$

In Table 16, we list all $\lambda_{\mathcal{G}}$ values used in the experiments.

Table 16: Choice of thresholding parameter $\lambda_{\mathcal{G}}$ for latent graph estimation.

Experiment	perfect scores	noisy scores
Table 3	0.001	0.1
Table 6	0.001	0.05
Table 7	0.0001	0.001
Table 8	0.001	0.1
Table 9	0.01	0.5

Implementation details of Algorithm 2. The latent graph estimation part of Algorithm 2 involves rank tests, i.e., determining the dimension of the intersection of column spaces. Therefore, we also need a threshold, denoted by λ_{eigv} to determine the numerical ranks in practice. In experiments reported in Table 8, we set $\lambda_{\mathrm{eigv}} = 0.01$.

Using the algorithms of related work. For the comparisons in Section 7.3, we used the shared sources codes of Squires et al. (2023)[8] and Zhang et al. (2023)[9].

E.2 Implementation Details of GSCALE-I Algorithm

We perform experiments for the coupled interventions setting, and solve the following optimization problem in Step G2 of GSCALE-I,

$$\min_{h} \left\| \mathbf{D}_{\mathrm{t}}(h) - \mathbf{I}_{n \times n} \right\|_{1,1} + \lambda \mathbb{E}\left[\left\| h^{-1}(h(\mathbf{X})) - \mathbf{X} \right\|_2^2 \right] . \tag{302}$$

where $\lambda > 0$ is a regularization parameter to ensure injectivity. In the following, first, we show why this problem is equivalent to solving (83). Then, we describe the computation of the ground truth score differences for $\mathbf{X}$ and discuss other implementation details.

Implementation steps. We use n_{s} samples from the observational environment to compute empirical expectations. Since encoder h is parameterized by $\mathbf{H}$, we use gradient descent to learn this matrix. To do so, we minimize the loss described above in (302). We denote the final parameter estimate by $\mathbf{H}^*$ and the encoder parameterized by $\mathbf{H}^*$ by h^*.

In the simulation results reported in Section 7.2, we set $\lambda = 1$ and solve (302) using RMSprop optimizer with learning rate 10^{-3} for 3×10^4 steps for $n = 5$ and 4×10^4 steps for $n = 8$. We also use early stopping when the training converges before the maximum number of steps. Recall that the latent graph estimate $\hat{\mathcal{G}}$ is constructed using $\mathbb{1}\{\mathbf{D}(h^*)\}$. Similarly to (301), We use a threshold $\lambda_{\mathcal{G}}$ to obtain the graph from the upper triangular part of $\mathbf{D}(h^*)$.

Architecture details and hyperparameter selection for image experiments. In Table 17 and Table 18, we list the architecture details and training hyperparameters used in the image experiments in Section 7.5. Specifically, Table 17 specifies the autoencoder architecture we used to estimate the latent variables, and Table 18 details the classification-based score difference network we trained. Finally, Table 19 lists the first level CNN-based autoencoder we used for image experiments with 2 balls.

---

8. https://github.com/csquires/linear-causal-disentanglement-via-interventions
9. https://github.com/uhlerlab/discrepancy_vae

Table 17: 2-step autoencoder architecture for learning latent representations of images.

Step	Details
Step 1 Encoder	Input: Image $\mathbf{X} \in \mathbb{R}^{64 \times 64 \times 3}$ Flatten FC(256), ReLU, LayerNorm FC(64) $\triangleright$ Intermediate representation $\mathbf{Y}$
Step 1 Decoder	FC(256), ReLU, LayerNorm FC($64 \times 64 \times 3$), Sigmoid $\triangleright$ Reconstructed image $\hat{\mathbf{X}}_{\mathrm{c}}$
Autoencoder-1	Training: 100 epochs, Adam optimizer, batch size:16 Learning rate: 10^{-3}, weight decay: 0.01 Minimize $\mathbb{E}\big[\|\mathbf{X} - \hat{\mathbf{X}}_{\mathrm{c}}\|^2\big]$ $\triangleright$ recons. loss
Step 2 Encoder	Input: $\mathbf{Y} \in \mathbb{R}^{64}$ FC(256), ReLU, LayerNorm FC(6) $\triangleright$ Latent causal variables $\hat{\mathbf{Z}}$
Step 2 Decoder	FC(256), ReLU, LayerNorm FC(64) $\triangleright$ Reconstructed $\hat{\mathbf{Y}}$
Autoencoder-2	Training: 100 epochs, Adam optimizer, batch size: 16, weight decay: 0.01 Learning rate: 0.95 decay/epoch for 75 epochs, reset to 10^{-3} for 25 epochs Minimize $\|\mathbf{D}_{\mathrm{t}}(h) - \mathbf{I}_{n \times n}\|_{1,1} + \lambda \mathbb{E}\big[\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2\big]$ $\triangleright$ score loss + recons. loss

Table 18: Architecture for LDR (log-density-ratio) model for score-difference estimation.

Layers:	Input: Image $\mathbf{X} \in \mathbb{R}^{64 \times 64 \times 3}$, class (intervention) label $y \in \{0, 1\}$ Conv(3, 32, kernel size=3, stride=1, padding=1), ReLU, BatchNorm, Dropout(0.1) MaxPool(kernel size=2, stride=2) Conv(32, 64, kernel size=3, stride=1, padding=1), ReLU, BatchNorm, Dropout(0.1) MaxPool(kernel size=2, stride=2) Flatten, FC(128), ReLU FC(1)
Training:	10 epochs, Adam optimizer, weight decay: 0.01, batch size: 16 Learning rate: 10^{-5} Minimize binary cross entropy between sigmoid of negation of output and labels

E.3   Experimental Methodology Comparison with DiscrepancyVAE

One major difference between the LSCALE-I algorithm and the publicly available DiscrepancyVAE implementation is that DiscrepancyVAE tries to learn the intervention effects as an encoding of the intervention labels. Specifically, for each intervention target, the VAE model learns a vector of latent intervention targets (a neural network with softmax activation) to denote which latent node is intervened on. In practice, learning an intervention encoding runs the risk of learning encodings where different interventions affect the same latent node, consequently, having nodes that are not affected by any intervention encoding. We demonstrate this phenomenon for DiscrepancyVAE in Figure 8, plotting the learned intervention embeddings for each single-node intervention. We observe that despite hav-

Table 19: Autoencoder architecture for learning latent representations of images - Step 1, CNN variant.

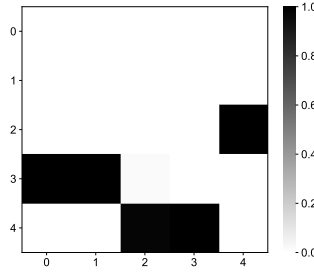| Step	Details
Encoder:	Input: Image $\mathbf{X} \in \mathbb{R}^{64 \times 64 \times 3}$
	Conv(3, 64, kernel size=3, stride=1, padding=1), BatchNorm, ReLU, MaxPool(kernel size=2, stride=2)
	Conv(64, 128, kernel size=3, stride=1, padding=1), BatchNorm, ReLU, MaxPool(kernel size=2, stride=2)
	Conv(128, 256, kernel size=3, stride=1, padding=1), BatchNorm, ReLU, MaxPool(kernel size=2, stride=2)
	Conv(256, 512, kernel size=3, stride=1, padding=1), BatchNorm, ReLU, MaxPool(kernel size=2, stride=2)
	Flatten, FC(n)
Decoder:	FC($n, 512 \cdot 4 \cdot 4$), Unflatten to (512, 4, 4)
	Upsample(scale=2), ConvTranspose(512, 256, kernel size=3, stride=1, padding=1), BatchNorm, ReLU
	Upsample(scale=2), ConvTranspose(256, 128, kernel size=3, stride=1, padding=1), BatchNorm, ReLU
	Upsample(scale=2), ConvTranspose(128, 64, kernel size=3, stride=1, padding=1), BatchNorm, ReLU
	Upsample(scale=2), ConvTranspose(64, 3, kernel size=3, stride=1, padding=1), BatchNorm, ReLU
	Sigmoid
Training:	100 epochs, Adam optimizer, Cosine annealing with warm restart at 1/5th of epochs, batch size: 256
	Learning rate: 10^{-3}, weight decay: 0.01
	Minimize $\mathbb{E}\big[\|\mathbf{X} - \hat{\mathbf{X}}_c\|^2\big]$ $\triangleright$ recons. loss

Figure 8: Intervention encoding learned in the sample run in Figure 3.

ing exhaustive single-node interventions, the intervention embedding network fails to learn exhaustive intervention targets/effects. Therefore, the learned graph is not immediately interpretable as a DAG on n causal latent variables since the learned intervention targets (i.e., node labels) are not surjective.

In contrast, in LSCALE-I, the effect of an intervention is estimated using score function differences. While the distinction may seem small, this approach ensures that the estimated graph's nodes are always interpretable since score differences between observational and interventional environments directly correspond to the intervention targets.

E.4 Visualizations for Intervention Extrapolation on Biological Data

In Figure 9, we plot certain double-node intervention samples, both actual and sampled using our score-difference Langevin method in a UMAP (McInnes et al., 2018) plot. We plot the double-node interventions that Zhang et al. (2023) plots in their double-node intervention extrapolation results that have more than 1000 samples for both single and double-node environments.sec:discrvae-invert-enc-problem We see across intervention pairs, score-based intervention extrapolation method leads to a sampling pattern that is relatively compliant
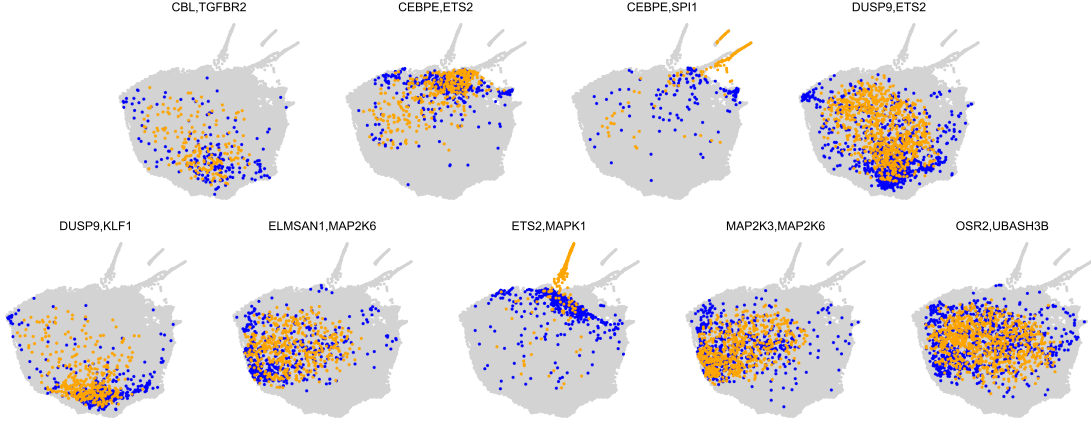
83

Figure 9: UMAP visualizations for double-node intervention extrapolation. Actual samples are in blue, generated samples are in orange.

with the actual double node interventional data. However, we observe that in some pairs, intervention effects are overestimated (particularly for CEBPE+SPI and ETS2+MAPK1).

E.5   Details of Score Functions Used in the Experiments

Score function of the linear Gaussian model.   For a Gaussian random vector $\mathbf{Y} \sim \mathcal{N}(\mu, \Sigma)$, the score function of $\mathbf{Y}$ is given by

$$\mathbf{s_Y}(\mathbf{y}) = -\Sigma^{-1} \cdot (\mathbf{y} - \mu) \ . \tag{303}$$

Using the linear model specified in (102), in environment $\mathcal{E}^0$,

$$\mathbf{Z} = (\mathbf{I}_{n \times n} - \mathbf{A})^{-1} \cdot N_i \ . \tag{304}$$

Since N_i is a zero-mean Gaussian random vector and $(\mathbf{I}_{n \times n} - \mathbf{A})^{-1}$ is a full rank matrix, $\mathbf{Z}$ is a zero-mean Gaussian random vector. Hence,

$$\mathbf{s}(\mathbf{z}) = -\big[\mathrm{Cov}(\mathbf{Z})\big]^{-1} \cdot \mathbf{z} \ . \tag{305}$$

Note that the covariance of $\mathbf{Z}$ has the following form

$$\mathrm{Cov}(\mathbf{Z}) = (\mathbf{I}_{n \times n} - \mathbf{A})^{-1} \cdot \mathrm{diag}\big([\sigma_1^2, \ldots, \sigma_n^2]^\top\big) \cdot \big((\mathbf{I}_{n \times n} - \mathbf{A})^{-1}\big)^\top \ . \tag{306}$$

Similarly, the score function of $\mathbf{Z}$ in $\mathcal{E}^m$ is given by

$$\mathbf{s}^m(z) = -\big[\mathrm{Cov}(\mathbf{Z}^m)\big]^{-1} \cdot \mathbf{z} \ , \qquad \forall m \in [n] \ , \tag{307}$$

in which $\mathbf{Z}^m$ denotes the latent variables $\mathbf{Z}$ in environment $\mathcal{E}^m$. Finally, by setting $f = \mathbf{G}$, Theorem 9 specifies the score differences of $\mathbf{X}$ in terms of $\mathbf{Z}$ under different environment pairs as

$$\mathbf{s_X}(\mathbf{x}) - \mathbf{s}_\mathbf{X}^m(\mathbf{x}) = \big(\mathbf{G}^\dagger\big)^\top \cdot \big[\mathbf{s}(\mathbf{z}) - \mathbf{s}^m(\mathbf{z})\big] \ . \tag{308}$$

84

Score estimation for a linear Gaussian model. For all environments $\mathcal{E}^m \in \mathcal{E}$, given n_{s} i.i.d. samples of $\mathbf{X}$, we first compute the sample covariance matrix denoted by $\hat{\Sigma}^m$. Then, we compute the sample precision matrix as

$$\hat{\Theta}^m \triangleq (\hat{\Sigma}^m)^\dagger \, , \tag{309}$$

which leads to the score function estimate given by

$$\hat{\mathbf{s}}_{\mathbf{X}}^m(\mathbf{x}) \triangleq -\hat{\Theta}^m \cdot \mathbf{x} \, . \tag{310}$$

Score function of the quadratic model. Following (16), score functions s^m and $\tilde{s}^m$ are decomposed as follows.

$$\mathbf{s}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log q_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \, , \tag{311}$$

$$\text{and} \quad \tilde{\mathbf{s}}^m(\mathbf{z}) = \nabla_{\mathbf{z}} \log \tilde{q}_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_{\mathbf{z}} \log p_i(z_i \mid \mathbf{z}_{\mathrm{pa}(i)}) \, . \tag{312}$$

For additive noise models, the terms in (311) and (312) have closed-form expressions. Specifically, using (131) and denoting the score functions of the noise terms $\{N_i : i \in [n]\}$ by $\{r_i : i \in [n]\}$, we have

$$[\mathbf{s}(\mathbf{z})]_i = r_i(n_i) - \sum_{j \in \mathrm{ch}(i)} \frac{\partial f_i(\mathbf{z}_{\mathrm{pa}(j)})}{\partial z_i} \cdot r_j(n_j) \, . \tag{313}$$

Recall that we consider a quadratic latent model with

$$f_i(\mathbf{z}_{\mathrm{pa}(i)}) = \sqrt{\mathbf{z}_{\mathrm{pa}(i)}^\top \cdot \mathbf{Q}_i \cdot \mathbf{z}_{\mathrm{pa}(i)}} \, , \quad \text{and} \quad N_i \sim \mathcal{N}(0, \sigma_i^2) \, , \tag{314}$$

which implies

$$\frac{\partial f_j(\mathbf{z}_{\mathrm{pa}(j)})}{\partial z_i} = \frac{[\mathbf{Q}_j]_i \cdot \mathbf{z}_{\mathrm{pa}(j)}}{\sqrt{\mathbf{z}_{\mathrm{pa}(j)}^\top \cdot \mathbf{Q}_j \cdot \mathbf{z}_{\mathrm{pa}(j)}}} \, , \quad \text{and} \quad r_i(n_i) = -\frac{n_i}{\sigma_i^2} \, , \quad \forall i \in [n] \, . \tag{315}$$

Components of the score functions $\mathbf{s}^m$ and $\tilde{\mathbf{s}}^m$ can be computed similarly. Subsequently, using Theorem 9 of Theorem 8, we can compute the score differences of observed variables as follows.

$$\mathbf{s}_{\mathbf{X}}(\mathbf{x}) - \mathbf{s}_{\mathbf{X}}^m(\mathbf{x}) = \left[[J_g(\mathbf{z})]^\dagger \right]^\top \cdot [\mathbf{s}(\mathbf{z}) - \mathbf{s}^m(\mathbf{z})] \, , \tag{316}$$

$$\mathbf{s}_{\mathbf{X}}(\mathbf{x}) - \tilde{\mathbf{s}}_{\mathbf{X}}^m(\mathbf{x}) = \left[[J_g(\mathbf{z})]^\dagger \right]^\top \cdot [\mathbf{s}(\mathbf{z}) - \tilde{\mathbf{s}}^m(\mathbf{z})] \, , \tag{317}$$

$$\mathbf{s}_{\mathbf{X}}^m(\mathbf{x}) - \tilde{\mathbf{s}}_{\mathbf{X}}^m(\mathbf{x}) = \left[[J_g(\mathbf{z})]^\dagger \right]^\top \cdot [\mathbf{s}^m(\mathbf{z}) - \tilde{\mathbf{s}}^m(\mathbf{z})] \, . \tag{318}$$

References

E. Acartürk, B. Varıcı, K. Shanmugam, and A. Tajer. Sample complexity of interventional causal representation learning. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

K. Ahuja, J. S. Hartford, and Y. Bengio. Properties from mechanisms: An equivariance perspective on identifiable representation learning. In *Proc. International Conference on Learning Representations*, virtual, April 2022.

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional causal representation learning. In *Proc. International Conference on Machine Learning*, Honolulu, Hawaii, July 2023.

K. Ahuja, A. Mansouri, and Y. Wang. Multi-domain causal representation learning via weak distributional invariances. In *Proc. International Conference on Artificial Intelligence and Statistics*, Valencia, Spain, May 2024.

S. Bing, U. Ninad, J. Wahl, and J. Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Proc. Causal Learning and Reasoning*, Los Angeles, CA, April 2024.

W. M. Boothby. *An introduction to differentiable manifolds and Riemannian geometry, Revised*, volume 120. Gulf Professional Publishing, 2003.

J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen. Weakly supervised causal representation learning. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

L. A. Gilbert, M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C. Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, 159(3):647–661, 2014.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(2), 2012.

H. Hälvä and A. Hyvärinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Proc. Conference on Uncertainty in Artificial Intelligence*, virtual, August 2020.

A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proc. International Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, April 2017.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999.

A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proc. International Conference on Artificial Intelligence and Statistics*, Naha, Japan, April 2019.

M. Jain, A. Denton, S. Whitfield, A. Didolkar, B. Earnshaw, J. Hartford, et al. Automated discovery of pairwise interactions from unstructured data. *arXiv:2409.07594*, 2024.

Y. Jiang and B. Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

J. Jin and V. Syrgkanis. Learning linear causal representations from general environments: Identifiability and intrinsic ambiguity. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proc. International Conference on Artificial Intelligence and Statistics*, virtual, August 2020a.

I. Khemakhem, R. Monti, D. Kingma, and A. Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2020b.

B. Kivva, G. Rajendran, P. Ravikumar, and B. Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

S. Lachapelle, P. R. López, Y. Sharma, K. Everett, R. L. Priol, A. Lacoste, and S. Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv:2401.04890*, 2024.

T. E. Lee, J. A. Zhao, A. S. Sawhney, S. Girdhar, and O. Kroemer. Causal reasoning in simulation for structure and transfer learning of robot manipulation policies. In *Proc. IEEE International Conference on Robotics and Automation*, Xi'an, China, May 2021.

A. Li, Y. Pan, and E. Bareinboim. Disentangled representation learning in non-markovian causal systems. In *Proc. Advances on Neural Information Processing Systems*, Vancouver, Canada, December 2024a.

Z. Li, Y. Shen, K. Zheng, R. Cai, X. Song, M. Gong, Z. Hao, Z. Zhu, G. Chen, and K. Zhang. On the identification of temporally causal representation with instantaneous dependence. *arXiv:2405.15325*, 2024b.

W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. Causal component analysis. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Intervention design for causal representation learning. In *UAI 2022 Workshop on Causal Representation Learning*, Eidhoven, Netherlands, 2022.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *Proc. International Conference on Learning Representations*, Kigali, Rwanda, May 2023a.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Biscuit: Causal representation learning from binary interactions. In *Proc. Uncertainty in Artificial Intelligence*, Pittsburgh, PA, August 2023b.

Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. van den Hengel, K. Zhang, and J. Q. Shi. Identifiable latent polynomial causal models through the lens of change. In *Proc. International Conference on Learning Representations*, Vienna, Austria, May 2024.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. International Conference on Machine Learning*, Long Beach, CA, June 2019.

F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *Proc. International Conference on Machine Learning*, virtual, April 2020.

L. McInnes, J. Healy, N. Saul, and L. Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

F. Montagna, A. A. Mastakouri, E. Eulig, N. Noceti, L. Rosasco, D. Janzing, B. Aragam, and F. Locatello. Assumption violations in causal discovery and the robustness of score matching. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023a.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In *Proc. Conference on Causal Learning and Reasoning*, Tübingen, Germany, April 2023b.

H. Morioka and A. Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *Proc. International Conference on Machine Learning*, Vienna, Austria, July 2024.

T. M. Norman, M. A. Horlbeck, J. M. Replogle, A. Y. Ge, A. Xu, M. Jost, L. A. Gilbert, and J. S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2009.

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 *Proc. International Conference on Machine Learning*, Baltimore, MD, July 2022.

S. Saengkyongam, E. Rosenfeld, P. Ravikumar, N. Pfister, and J. Peters. Identifying representations for intervention extrapolation. In *Proc. International Conference on Learning Representations*, Vienna, Austria, May 2024.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021.

X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(1): 10994–11048, 2022.

R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole. Weakly supervised disentanglement with guarantees. In *Proc. International Conference on Learning Representations*, virtual, May 2020.

L. Simon. Introduction to geometric measure theory. *Tsinghua Lectures*, 2014.

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

C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. Linear causal disentanglement via interventions. In *Proc. International Conference on Machine Learning*, Honolulu, Hawaii, July 2023.

N. Sturma, C. Squires, M. Drton, and C. Uhler. Unpaired multi-domain causal representation learning. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

A. Tejada-Lapuerta, P. Bertin, S. Bauer, H. Aliee, Y. Bengio, and F. J. Theis. Causal machine learning for single-cell genomics. *Nature Genetics*, pages 1–12, 2025.

B. Varıcı, E. Acartürk, K. Shanmugam, and A. Tajer. Linear causal representation learning from unknown multi-node interventions. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning with interventions. *arXiv:2301.08230*, 2023.

J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021.

J. von Kügelgen, M. Besserve, W. Liang, L. Gresele, A. Kekić, E. Bareinboim, D. M. Blei, and B. Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

E. W. Weisstein. Sigmoid function. *https://mathworld. wolfram. com/*, 2002.

R. Welch, J. Zhang, and C. Uhler. Identifiability guarantees for causal disentanglement from purely observational data. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2024.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. International Conference on Machine Learning*, Bellevue, Washington, July 2011.

D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. Multi-view causal representation learning with partial observability. In *Proc. International Conference on Learning Representations*, Vienna, Austria, May 2024.

D. Yao, D. Rancati, R. Cadei, M. Fumero, and F. Locatello. Unifying causal representation learning with the invariance principle. In *Proc. International Conference on Learning Representations*, Singapore, April 2025.

W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *Proc. International Conference on Learning Representations*, virtual, April 2022.

J. Zhang, C. Squires, K. Greenewald, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

K. Zhang, S. Xie, I. Ng, and Y. Zheng. Causal representation learning from multiple distributions: A general setting. In *Proc. International Conference on Machine Learning*, Vienna, Austria, July 2024.

Z. Zhu, F. Locatello, and V. Cevher. Sample complexity bounds for score-matching: Causal discovery and generative modeling. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.