PROPERTIES FROM MECHANISMS: AN EQUIVARIANCE PERSPECTIVE ON IDENTIFIABLE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A key goal of unsupervised representation learning is "inverting" a data generating process to recover its latent properties. Existing work that provably achieves this goal relies on strong assumptions on relationships between the latent variables (e.g., independence conditional on auxiliary information). In this paper, we take a very different perspective on the problem and ask, "Can we instead identify latent properties by leveraging knowledge of the mechanisms that govern their evolution?" We provide a complete characterization of the sources of non-identifiability as we vary knowledge about a set of possible mechanisms. In particular, we prove that if we know the exact mechanisms under which the latent properties evolve, then identification can be achieved up to any equivariances that are shared by the underlying mechanisms. We generalize this characterization to settings where we only know some hypothesis class over possible mechanisms, as well as settings where the mechanisms are stochastic. We demonstrate the power of this mechanism-based perspective by showing that we can leverage our results to generalize existing identifiable representation learning results.These results suggest that by exploiting inductive biases on mechanisms, it is possible to design a range of new identifiable representation learning approaches.

1 INTRODUCTION

Modern unsupervised representation learning techniques can generate images of our world with intricate detail (e.g. Karras et al., 2019; Song et al., 2020; Razavi et al., 2019), and yet, the latent representations from which these images are generated remain entangled and challenging to interpret (Schölkopf et al., 2021; Locatello et al., 2019). At the same time, the success of pre-trained of transformers (Devlin et al., 2018; Brown et al., 2020) shows that advances in our ability to disentangle the underlying generative factors can lead to dramatic improvements in the sample complexity of downstream supervised tasks (Bengio & LeCun, 2007; Bengio et al., 2013). In order to consistently replicate this success, we need methods that can reliably invert the data generating process. This is a challenging task because unsupervised representation learning with independent and identically distributed (IID) data is hopelessly *unidentified*: even in the nice case in which observations are generated via some $p(x|z)$, there exists an infinitely large class of possible latent distributions $\tilde{p}(z|x)$ that are consistent with the observed marginal distribution, $p(x)$, and only one of them corresponds to the true latent distribution $p(z|x)$ (Locatello et al., 2019; Khemakhem et al., 2020a).

If we want to build systems which are able to provably *identify*[1] the true latent variables z that generated our observed data $x = g(z)$ for some observation model g, then we need to exploit structural assumptions that constrain this set of possible distributions. Most of the prior work that can provide such guarantees was developed in the independent component analysis (ICA) literature. The key ICA assumption is that the latent factors are (conditionally) independent and non-Gaussian. Then, if the observation model, $g : \mathcal{Z} \to \mathcal{X}$ is linear and invertible, $z = g^{-1}(x)$ is identified (Comon, 1994), and for nonlinear g there are a number of recent approaches that leverage non-stationarity in the distributions over z to identify g^{-1} (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019;

[1] A problem is identified if there exists a unique solution in the infinite data limit and no constraints on model capacity.

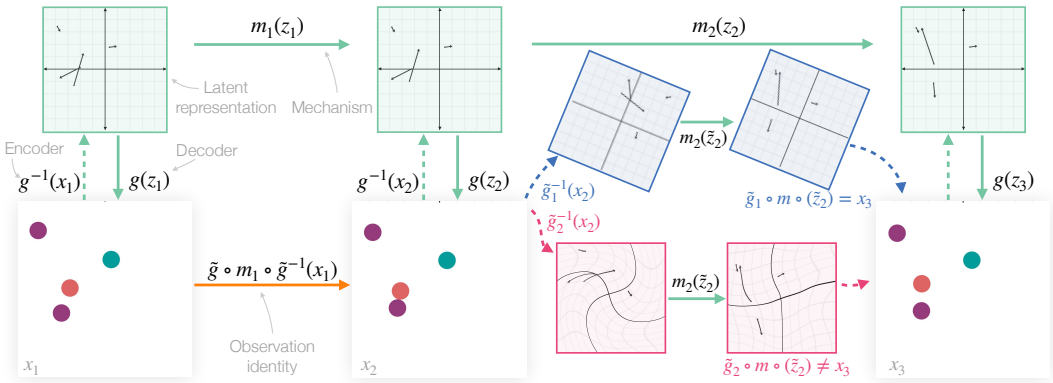

Figure 1: This simple data generating process illustrates that if we know the set of mechanisms that govern the evolution of an environment, this constrains the set of possible representations to any equivariances of these mechanisms. At each time step, we observe an environment of bouncing balls in pixel space as images, x_t. These images are produced by some rendering engine, g, as a function of the true latent representation z_t, which in this case gives positions and velocities of each ball (illustrated by the location and length of the arrows in the latent representation). At each time step, the state evolves according to a mechanism, m_t. Any candidate model that is consistent with the observed data has to satisfy the observation identity. If an encoder produces either the true representation (shown in green), or a representation transformed by some equivariance of the mechanism (e.g. the rotated representation shown in blue) then the observation identity is satisfied. However, models that produce representations that are arbitrary transformations of the true representation (e.g. the warped representation shown in pink) can be discarded because they are not consistent with the observation identity.

Khemakhem et al., 2020a). These results give a tantalizing demonstration that representation learning with identification guarantees is possible, but the requirement that the latent factors are statistically (conditionally) independent is limiting[2] (Higgins et al., 2018; Schölkopf et al., 2021).

In this paper we take a very different approach. We study how the mechanisms that govern an environment's evolution constrain the set of possible latent representations that are consistent with the data. As a simple concrete example, consider the bouncing ball environment shown in Figure 1. The latent state can be completely described by a vector, z, containing the position, velocity and acceleration of the balls[3], and given this latent state, the images, x, shown in Figure 1, are produced via some rendering engine $g : \mathcal{Z} \to \mathcal{X}$. Our task is to leverage sequences of observations $x_1, \ldots, x_T$ and knowledge of m to recover $z_1, \ldots, z_T$. If we can show that this task has a unique solution that is consistent with the observations and mechanisms, then the problem is identified.

Our main result is that when we know the true mechanism, m, the system is identified up to any equivariances of the mechanism; or equivalently, in Section 2.2 we prove that we can identify z up to any invertible transformation $a : \mathcal{Z} \to \mathcal{Z}$ that commutes with m, such that the composition $m \circ a = a \circ m$. For example, in the bouncing balls environment shown in Figure 1, the laws of physics that govern the system are equivariant with respect to your choice of units of measurement—changing from representing z in meters to inches leaves the output of the mechanism unchanged up to a corresponding unit change—and hence we can only hope to identify z up to some scaling function $a(z)$ which corresponds to an arbitrary choice of units of measurement. Interestingly, when the environment evolves according to multiple known mechanisms, then the sources of non-identifiability are even further constrained: such a system is identified up to any equivariances that are shared across all n mechanisms.

[2] As a simple example of dependence between latent variables, assume the bouncing balls shown in Figure 1 have different masses indicated by their colors. If the initial conditions were such that all the balls have the same momentum, then mass and velocity will be inversely correlated.

[3] A complete generative model of these images would also need to track the colors and shapes of the elements; we will return to this issue in the discussion of the results.

Perfectly knowing the true mechanisms and when they are applied is unlikely, but in Section 2.3 we show that we can relax that assumption to a setting where we instead know a hypothesis class of possible mechanisms that could have been applied. This weaker assumption leads to an additional source of non-identifiability: the mechanisms in our hypothesis class can *imitate* each other if there exists an invertible transformation a such that for any two mechanisms m_1 and m_2 in our hypothesis class, $m_2 = a^{-1} \circ m_1 \circ a$. For example, if we are in a setting where our hypothesis class includes both a product mechanism, $m_1(z) = \prod_i z_i$, and a sum mechanism, $m_2(z) = \sum_i z_i$, then m_1 can imitate m_2 if $a(z) = \exp(z)$, since $\sum_i z_i = \log(\prod_i \exp(z_i))$. As before, this result is complete, in the sense that these two sources of non-identifiability—equivariance and imitation—are the only sources of non-identifiability in such systems. This gives us a natural way of thinking about the way knowledge of deterministic mechanisms constrains a representation learning task: with complete knowledge, the only source of non-identifiability is any equivariances inherent in the mechanism, but as we relax our assumptions to allow for a set of possible mechanisms, we also have have to contend with non-identifiability via imitation. That said, the relationship between the size of the hypothesis class and the size of the set of a's that commute via imitation is not necessarily monotonic: for environments governed by multiple mechanisms, a larger hypothesis class can lead to fewer imitations.

Section 3, shows that we can derive analogous results for stochastic mechanisms, $m(z, U)$, where the mechanism defines a conditional distribution $p(z_{t+1}|z_t)$. This generalization gives us a way of comparing our mechanism-based perspective with existing identifiability results. We demonstrate this in Section 4 by showing that it is possible to view the distributional assumptions made in Klindt et al. (2020) as a particular choice of mechanism, and by doing so, we can leverage our theory to give alternative proofs of these results. This strategy required weaker distributional assumptions, thereby generalizing their result. Finally, we give a mechanism-based perspective on the related work in Section 5 and Section 6 concludes with a discussion of the open problems that need to be addressed in order to reliably leverage this approach in practice.

2 Mechanism based identification

2.1 Data generation process

The state of the system at time $t \in \{1, \cdots, T\}$ is given by $z_t \in \mathcal{Z} \subseteq \mathbb{R}^d$. At each time t, we observe $g(z_t) = x_t \in \mathcal{X} \subseteq \mathbb{R}^n$, which is some transformation of the latent z_t. We can think of $g : \mathcal{Z} \to \mathcal{X}$ as a function that transforms the (typically low dimensional) state variables to the (typically high dimensional) observed variables; for example, in the bouncing ball environment described in the introduction, g is the rendering engine that produces the images shown in Figure 1. We assume g is injective with respect to $\mathbb{R}^n$—i.e. $g(z_1) = g(z_2)$ implies $z_1 = z_2$; or equivalently, any change to the underlying state, z, is reflected in some pixel-level change to the observation x—and we make g bijective by restricting its inverse g^{-1} to any x on the data manifold which we denote $\mathcal{X}$ (i.e. $\mathcal{X}$ is the image of g). The state transition from time t to $t+1$ is governed by a mechanism $m_t : \mathcal{Z} \to \mathcal{Z}$. There may be multiple mechanisms in a given environment. For example in Figure 1 the transition from z_1 to z_2 does not involve any collisions so the state evolves according to Newton's first law of motion (Newton, 1687); the transition from z_2 to z_3 involves a collision between two of the balls that is described by Newton's third law. Together m_t and g describe the data-generation from time t to $t+1$ as follows,

$$x_t \leftarrow g(z_t), \qquad z_{t+1} \leftarrow m_t(z_t). \tag{1}$$

If we fix the initial conditions, z_0, this data generation process is deterministic. In Section 2.2, we provide results when the underlying mechanism is known and in Section 2.3, we we extend those results to the case when the mechanisms are not known. In Section 3, we extend our results to stochastic mechanisms, $m(Z_t, U)$ that take samples from some distribution $U \sim \text{Uniform}(0, 1)$ as input. These stochastic mechanisms can represent any conditional distribution $P(Z_{t+1}|Z_t)$ over states.[4]

[4]For example, if $Z_{t+1}|Z_t \sim \mathcal{N}(Z_t, \sigma)$ then $m(z, U) = z + \sigma\Phi^{-1}(U)$ where Φ is the cumulative density of a standard Gaussian. Strictly, representing a joint (Z_{t+1}, Z_t) as $(m(Z_t, U), Z_t)$ requires a restriction to nice spaces. See Austin (2015, Lemma 3.1).

2.2 IDENTIFYING ENCODERS WHEN THE UNDERLYING MECHANISM IS KNOWN

We begin in the simplest version of the system described by equation (1): assume that at each time t the same mechanism $m : \mathcal{Z} \to \mathcal{Z}$ is used, and we know this mechanism. From these assumptions we can derive an identity that describes how the observations x_t and x_{t+1} are related,

$$z_{t+1} = m(z_t), \quad g^{-1}(x_{t+1}) = m \circ g^{-1}(x_t), \quad x_{t+1} = g \circ m \circ g^{-1}(x_t). \tag{2}$$

It may be helpful to think of this identity as describing an autoencoder where we require that the encoder $g^{-1}(x_t)$ inverts the data generating process to produce some latent z_t from x_t, and that the decoder has to reproduce x_{t+1} from $m(z_t)$; i.e. the representation transformed by the mechanism. Importantly, this identity describes the true data generating process, rather than some model of it. Our hypothesis class over the possible encoder / decoder functions is the set, $\mathcal{G}$, of all bijective functions from $\mathcal{Z} \to \mathcal{X}$. By assuming that bijectivity we are essentially assuming that the reconstuction task is solved:[5] $\mathcal{G}$ is the set of all autoencoders that perfectly reproduce the data, such that for any x on the data manifold $\mathcal{X}$, and any an encoder / decoder pair $(\tilde{g}^{-1}, \tilde{g})$ with $\tilde{g} \in \mathcal{G}$, their composition is the identity function, $x = \tilde{g} \circ \tilde{g}^{-1} \circ x$. Because we assumed that the true g is bijective, we know that it is in our search space, $\mathcal{G}$. We can constrain this set using our knowledge of the mechanism by only considering solutions that also satisfy the identity given in equation 2, such that for every pair of observations (x_t, x_{t+1}), an analogous identity holds,

$$x_{t+1} = \tilde{g} \circ m \circ \tilde{g}^{-1}(x_t) \tag{3}$$

Now suppose that we have access to observations x_t from the entire set $\mathcal{X}$, then the above identities have to hold for all $x_t, \in \mathcal{X}$ with corresponding x_{t+1} from equation 2, so we can conclude that the follow functions are equal,

$$g \circ m \circ g^{-1} = \tilde{g} \circ m \circ \tilde{g}^{-1} \tag{4}$$

This relationship will hold for any of the possible decoders $\tilde{g}$ (and corresponding encoders $\tilde{g}^{-1}$) that that are observationally equivalent given our assumptions. We denote the set of all such decoders, $\mathcal{G}_{\mathsf{id}} = \{\tilde{g} \mid \tilde{g} \in \mathcal{G}, g \circ m \circ g^{-1} = \tilde{g} \circ m \circ \tilde{g}^{-1}\}$. If $\mathcal{G}_{\mathsf{id}} = \{g\}$ then we have shown that the problem is exactly identified, which means that if we manage to find a $\tilde{g}$ that satisfies equation 2, then $\tilde{g} = g$; but if $\mathcal{G}_{\mathsf{id}}$, also includes other functions $\tilde{g} \neq g$, then these functions are the sources of non-identifiability.

Equivariant mechanisms To see an example of a setting where the problem is not exactly identified, consider a mechanism which is *equivariant* with respect to some bijective transformation $a : \mathcal{Z} \to \mathcal{Z}$. A mechanism m is said to be *equivariant* w.r.t a if $a \circ m = m \circ a$. For example, a mechanism derived from Newtonian mechanics may be equivariant with respect to your choice of units of measurement, such that $m(cz) = cm(z)$ for some scaling constant c. Similarly, if a mechanism transforms sets of items, any permutation of z would lead to a corresponding permutation of the mechanism's output.

If we have a mechanism that is equivariant with respect to some transformation a (where a is not identity map), that implies that there exists a function $\tilde{g} \neq g$ in $\mathcal{G}_{\mathsf{id}}$, so the problem is not exactly identified. We can see this as follows,

$$g \circ m \circ g^{-1} = g \circ a^{-1} \circ a \circ m \circ g^{-1} = g \circ a^{-1} \circ m \circ a \circ g^{-1} = \tilde{g} \circ m \circ \tilde{g}^{-1}$$

where the first equality uses the fact that $a^{-1} \circ a$ is the identity function, the second applies the definition of equivariance, and the final equality defines $\tilde{g} := g \circ a^{-1}$ and $\tilde{g}^{-1} := a \circ g^{-1}$. This shows that if the mechanism is equivariant, an encoder, $\tilde{g}^{-1}$, can output a transformed $\tilde{z} = a(z)$, and the decoder, $\tilde{g}$, inverts this transformation before producing its output, thereby leaving the observed variables, x, unchanged. If we denote the set of all equivariances of the mechanism $\mathcal{E} = \{a \mid a \text{ is a bijection}, \ a \circ m = m \circ a\}$, then we can define the set of all such sources of non-identification as, $\mathcal{G}_{\mathsf{eq}} = \{\tilde{g} \mid \tilde{g} = g \circ a^{-1}, \ a \in \mathcal{E}\}$. This is a natural source of non-identification: given that we are relying on the mechanism to constrain the encoder $\tilde{g}^{-1}$, it is unsurprising we cannot prevent transformations that are not affected by the mechanism. The more interesting observation, which we will show below, is that this set is the *only* source of non-identification when the mechanism is known, and hence we recover the true z up to equivariances in the mechanism.

To state this theorem, we first define the notion of identifiability up to an equivalence class defined by a family of bijections $\mathcal{A}$, where $a \in \mathcal{A}$ is a map $a : \mathcal{Z} \to \mathcal{Z}$.

[5] This is obviously a strong assumption—learning autoencoders that perfectly reconstructed the data is not at all easy—but it focuses the discussion on the identification issues that remain after reconstruction is solved.

Definition 1. ***Identifiability up to $\mathcal{A}$.*** *If the learned encoder $\tilde{g}^{-1}$ and the true encoder g^{-1} are related by some bijection $a \in \mathcal{A}$, such that $\tilde{g}^{-1} = a \circ g^{-1}$ (or equivalently $\tilde{g} = g \circ a^{-1}$), then $\tilde{g}^{-1}$ is said to learn g^{-1} up to bijections in $\mathcal{A}$. We denote this $\tilde{g}^{-1} \sim_{\mathcal{A}} g^{-1}$.*

Suppose, for example, $\mathcal{A}$ is a family of a permutations. Identifiability up to $\mathcal{A}$ implies that the true latent variables will be recovered, but that they would not necessarily be ordered in the same way as they were in the original data generation process. In this setting where the mechanism, m, is known, the following theorem shows that $\mathcal{A}$ is just $\mathcal{E}$, the set of equivariances of m.

Theorem 1. *If the data generation process follows equation 2, then the set of autoencoders that are consistent with the mechanism's transitions $\mathcal{G}_{\mathsf{id}} = \mathcal{G}_{\mathsf{eq}}$, the set of decoders transformed by some $a \in \mathcal{E}$ (the equivariances of m), and hence, the encoders that solve the identity in equation 4 identify true encoder g^{-1} up to equivariances of m, such that $\tilde{g}^{-1} \sim_{\mathcal{E}} g^{-1}$.*

The proof is given in Section A.2 of the appendix. From Theorem 1, we can derive a number of observations. First, notice that if we have a standard autoencoder[6] then the mechanism is the identity map, $m(z) = z$, and its set of equivariances $\mathcal{E}$ is any invertible function a, and hence Theorem 1 shows that the encoder is essentially unconstrained. However, if the mechanism is any non-trivial function, $m(z') \neq z'$ for some z', then the space of possible encoders is significantly reduced to just those invertible transformations that commute with m. If the system involves multiple known mechanisms, $\{m_1, \ldots, m_T\}$, where at each time $z_{t+1} = m_t(z_t)$, then the encoder is even further constrained. Define the set of all the mechanisms that are used at least once in the evolution of the system as $\mathcal{M}^* = \cup_{t=1}^{T} \{m_t\}$. Suppose $\mathcal{E}^i$ denotes the equivariances of $m_i \in \mathcal{M}^*$. Define the equivariances that are shared across all the mechanisms as $\mathcal{E}^* = \cap_i \mathcal{E}^i$;

Corollary 1. *If the data generation process follows equation 1, then the encoders that satisfy identity in equation 4 for all $m \in \mathcal{M}^*$ identify true encoder g^{-1} up to the equivariances shared across all the mechanisms in $\mathcal{M}^*$, $\mathcal{E}^*$ ($\tilde{g}^{-1} \sim_{\mathcal{E}^*} g^{-1}$).*

The proof of the above claim is in Section A.4. This corollary implies a blessing that comes with more complex environments: if an object is transformed by multiple mechanisms which are diverse (in the sense that they share few equivariances), then it becomes easier to identify. Given access to inputs and outputs of a mechanism, if we cannot tell apart whether some transformation was applied to the input or the output, then the mechanism is equivariant with respect to the transformation. Together Theorem 1 and Corollary 1, state that we can learn to invert the data generation process but we cannot distinguish latents that were transformed by equivariances shared across the mechanisms.

2.2.1 Identifying encoders for affine mechanisms

Theorem 1 shows us that an encoder $\tilde{g}^{-1}$ is identified up to any equivariances of the known mechanism, but given some mechanism, it does not tell us what equivariances may exhibit. This section gives an example of how one might go about finding all sources of equivariance for a given mechanism. We derive the equivariances for affine mechanisms, and in doing so we show conditions under which affine mechanisms lead to identification up to some fixed offset. Affine mechanisms are broadly applicable because with a sufficiently short time interval, they approximate a wide variety of physical systems as the Euler discretization of some linear ordinary differential equation. In such systems, the mechanism is given by $m(z) = Mz_t + b_t$ where $M \in \mathbb{R}^{d \times d}$ is an invertible diagonalizable matrix (with eigendecomposition given as $M = S\Lambda S^{-1}$, where S is the matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues), $b_t \in \mathbb{R}^d$ is the offset parameter at time t, and the analog of equation 2 is,

$$x_{t+1} = g(Mg^{-1}(x_t) + b_t).$$

We search for some encoder $\tilde{g}^{-1}$ such that this relationship holds for all x and t. Define an offset function $o(z) = z + p$, where the offset function shifts the latent by a vector p. Define $\mathcal{O}$ to be the set of all the offset functions. We show that the encoder is identified up to $\mathcal{O}$ when we have at least two distinct offset terms b_t and a regularity condition (Assumption 2).[7]

Assumption 1. *In the data generation process in equation 1, we set $m(z_t) = Mz_t + b_t$, where M is invertible and diagonalizable. We assume that the offset b_t takes at least two distinct values, which we denote by b^1 and b^2.*

[6] With the constraint that the encoder is the inverse of the decoder such that $\tilde{g}^{-1}$ is bijective.

[7] We conjecture that the regularity condition holds for all analytic functions and is thus not needed. Since we do not have a proof of this claim, we include it as an assumption.

Assumption 2. *$a : \mathcal{Z} \to \mathcal{Z}$ is analytic and satisfies the following assumption. For each component $i \in \{1, \cdots, d\}$ of $a_i(z)$ and each $b \in \mathbb{R}^d$, define the set $\mathcal{S}^i = \{\theta \mid \nabla a_i(z+b) = \nabla a_i(z) + \nabla^2 a_i(\theta) b, z \in \mathbb{R}^d\}$. Each set $\mathcal{S}^i$ has a non-zero Lebesgue measure in $\mathbb{R}^d$.*

Theorem 2. *If the data generation process follows equation 1 with affine mechanisms, $m(z) = Mz_t + b_t$, Assumptions 1, 2 hold, the eigenvalues of the mechanism M are all distinct, and each component of $S^{-1}(b^1 - b^2)$ is non-zero, then the encoders that solve the identity in equation 4 identify true encoder g^{-1} up to offsets $\mathcal{O}$ such that $\tilde{g}^{-1} \sim_{\mathcal{O}} g^{-1}$.*

The proof is given in Section A.5 in the Appendix. The above theorem shows the power of using multiple mechanisms. It can be shown that if there is only one mechanism, then we cannot do better than linear identification. However, if we use two mechanisms as is the case in the above theorem, the constraint of shared equivariances (Theorem 1 and Corollary 1) enforces almost exact identifiability (only offset-based errors remain).

2.3 Identifying encoders when the mechanisms are not known

We have seen in the previous section that with complete knowledge of the mechanisms under which a system evolves, we can learn an encoder up to equivariances. In practice, however, we are unlikely to have such complete knowledge. In this section, the system still evolves according to some deterministic mechanism, $z_{t+1} \leftarrow m_t(z_t)$, but we assume that you only know some hypothesis class $\mathcal{M}$ of possible mechanisms which could have been used, without knowing which $m_t \in \mathcal{M}$ is used at every time step. A candidate solution now needs to propose both an encoder $\tilde{g}^{-1}$ and a mechanism $\tilde{m}_t \in \mathcal{M}$ for every (x_t, x_{t+1}) pair such that,

$$x_{t+1} = \tilde{g} \circ \tilde{m}_t \circ \tilde{g}^{-1}(x_t). \tag{5}$$

As before, this relationship holds for all $x_t \in \mathcal{X}$, where x_{t+1} is generated from m_t, so any candidate solution that is consistent with the x's that we observe must satisfy

$$g \circ m_t \circ g^{-1} = \tilde{g} \circ \tilde{m}_t \circ \tilde{g}^{-1}. \tag{6}$$

We define the set of all decoders $\tilde{g}$ (with corresponding encoders $\tilde{g}^{-1}$) that solve equation 6 as $\tilde{\mathcal{G}}_{\mathsf{id}} = \{\tilde{g} | \tilde{g} \text{ is a bijection }, g \circ m_t \circ g^{-1} = \tilde{g} \circ \tilde{m}_t \circ \tilde{g}^{-1}\}$. This set looks very much like the set $\mathcal{G}_{\mathsf{id}}$ that we defined in Section 2.2, but the fact that we have to select $\tilde{m} \in \mathcal{M}$ rather than knowing the true m implies a new source of non-identifiability: imitator mechanisms.

We partition the hypothesis class of mechanisms, $\mathcal{M} = \mathcal{M}^* \cup \mathcal{M}'$, into the mechanisms that are used at least once in the evolution of the system, $\mathcal{M}^*$ ($\mathcal{M}^* = \cup_{t=1}^{T} \{m_t\}$), and mechanisms which are hypothesized but not used, $\mathcal{M}'$.

Definition 2. ***Equivariances and imitators w.r.t $\mathcal{M}$.*** *Define a set of functions $\tilde{\mathcal{E}}$ that satisfy commutativity w.r.t the set of mechanisms $\mathcal{M}$ in the following sense. The set $\tilde{\mathcal{E}}$ comprises of all the bijections, $a(\cdot)$, that satisfy the following condition. If for each $m_1 \in \mathcal{M}^*$, $\exists\ m_2 \in \mathcal{M}$ such that $a \circ m_1 = m_2 \circ a$, then $a \in \tilde{\mathcal{E}}$.*

The set $\tilde{\mathcal{E}}$ consists of two types of elements. We illustrate this through an example of set $\mathcal{M} = \mathcal{M}^* = \{m^1, m^2\}$. If a is a bijection that commutes with both m^1 and m^2, i.e., $a \circ m^1 = m^1 \circ a$ and $a \circ m^2 = m^2 \circ a$, then $a \in \tilde{\mathcal{E}}$. From this we can see that $\tilde{\mathcal{E}}$ consists of elements in the intersection of the equivariances of the respective mechanisms. Alternatively, if a satisfies $a \circ m_1 = m_2 \circ a$ and $a \circ m_2 = m_1 \circ a$, then we say m_2 "imitates" m_1 and vice-versa because you can produce m_1's output from m_2 for any z using the following relationship $m_1 = a^{-1} \circ m_2 \circ a$. Further simplifcation of this yields that $a^2 = a \circ a$ is an equivariance of both m_1 and m_2. This example shows that when we know the list of mechanisms $\mathcal{M} = \mathcal{M}^*$ but do not know which mechanism is used, the set $\tilde{\mathcal{E}}$ can be expressed in terms of the equivariances of the mechanisms. For further details see the Appendix Section A.11. Define the set of maps that are identified up to $\tilde{\mathcal{E}}$ as $\tilde{\mathcal{G}}_{\mathsf{eq}} = \{\tilde{g} \mid \tilde{g} = g \circ a^{-1},\ a \in \tilde{\mathcal{E}}\}$.

Theorem 3. *If the data generation process follows equation 1, then $\tilde{\mathcal{G}}_{\mathsf{id}} = \tilde{\mathcal{G}}_{\mathsf{eq}}$ and hence, the set of all the encoders that satisfy equation 6 identify true encoders g^{-1} up to bijections that commute with the set $\mathcal{M}$, $\tilde{\mathcal{E}}$ ($\tilde{g}^{-1} \sim_{\tilde{\mathcal{E}}} g^{-1}$).*

The proof of the above is in Section A.6 of the Appendix. Equivariances and imitators play similar roles in the way that they relax constraints on the encoder $\tilde{g}^{-1}$—any bijection that commutes with

either is a source of non-identifiability—but they are different from the perspective of how we should think about designing representation learning algorithms. Recall that $\mathcal{M}$ is composed of two sets of mechanisms, $\mathcal{M} = \mathcal{M}^* \cup \mathcal{M}'$, mechanisms in $\mathcal{M}^*$ that are used at least once in the evolution of the environment and those mechanisms in $\mathcal{M}'$ which are hypothesized but never used. Equivariances are dictated only by $\mathcal{M}^*$, which characterizes the evolution of the environment. Any increases to the number of distinct mechanisms in $\mathcal{M}^*$ will potentially decrease the number of equivariances shared by all mechanisms. This can only be achieved by modifying the environment in some way, either through an explicit intervention that modifies it's mechanisms or by collecting data from multiple environments with diverse of mechanisms. For example, in bouncing balls example given in Figure 1, one could change the environment by varying the mass of the balls or observing it under different gravity conditions; or one could intervene by, say, changing the shape of balls in the system such that you get a different bouncing mechanism.

Imitators, by contrast, are a function of both the mechanisms $\mathcal{M}^*$ that were used and those that were hypothesized, $\mathcal{M}'$. An imitator is just some mechanism that can imitate another via some bijection, so one would expect that as we grow the number of mechanisms in $\mathcal{M}$, the size of the set of imitators can only grow; but interestingly, this is not always the case for mechanisms from $\mathcal{M}^*$. Recall that any $a \in \tilde{\mathcal{E}}$ produces an encoder of the form $\tilde{g}^{-1} = a \circ g^{-1}$, so the same transformation has to be used for all imitations and equivariances among the mechanisms in $\mathcal{M}^*$. Because of this, it is possible that increasing the size of $\mathcal{M}^*$ reduces the number of imitators. For example, if there is some mechanism, m_i, that does not commute with any non-trivial a in $\tilde{\mathcal{E}}$ (either by imitation or equivariance), then adding m_i to $\mathcal{M}^*$ will make the problem exactly identified. On the other hand, growing the size of the set of unused mechanisms, $\mathcal{M}'$, can result in significant non-identifiability. For example, if $\mathcal{M}$ consisted of a flexible class of functions (e.g. a multi-layer perceptron) then it would be easy to construct imitators of the form $m_1 = a^{-1} \circ m_2 \circ a$.

Illustrating Theorem 3. We consider the same setting as in Theorem 2. For each $t \in \{1, \cdots, d+1\}$, $m(z_t) = Mz_t + b_t$ and $x_t = g(z_t)$. We only know that the mechanism is affine and only the offset b_t is changing, but parameters M and b_t are not known. Let us construct the set $\tilde{\mathcal{E}}$ corresponding to the above setting. We can show that the set $\tilde{\mathcal{E}}$ consists of affine functions (See the Appendix A.7 for details). From Theorem 3, we can thus conclude that for this data generation process even with *very little knowledge of the mechanism, we get linear identifiability*. This is weaker than the offset based identifiability in Theorem 2, but there we were required to know the entire affine mechanism.

3 Stochastic mechanisms

The results thus far relied on the assumption that the evolution of a system could be described in terms of deterministic mechanisms. This deterministic approach models settings where the full latent state is observable (via some unknown encoder g^{-1}) at a short enough time interval that there is no uncertainty about the system's evolution. To generalize to cases where there is some uncertainty about the latent state's evolution, we now develop analogous identification results for stochastic mechanisms that induce conditional distributions over latent states.The systems evolves as

$$X_t \leftarrow g(Z_t), \quad Z_{t+1} \leftarrow m_t(Z_t, U_t), \tag{7}$$

where each U_t is noise with each component sampled independently from standard uniform distribution Uniform$[0,1]$, $Z_1 \sim \mathbb{P}_Z$, $m_t : \mathcal{Z} \times [0,1]^d \to \mathcal{Z}$, and the decoder $g : \mathcal{Z} \to \mathcal{X}$ is a diffeomorphism (i.e. a smooth bijection with an invertible Jacobian, see definition A.1 (Kass & Vos, 2011)).

In this section, we will focus on the case where the true mechanism is unknown; the case where mechanism is known is a special case with no imitator, i.e., $\tilde{m}_t = m_t$ for all t. The goal is to search for an encoder $\tilde{g}^{-1}$, which is a diffeomorphism, that generates $\hat{X}_{t+1} = \tilde{g} \circ \tilde{m}_t(\tilde{g}^{-1}(x_t), \hat{U}_t)$, where $\hat{U}_t$ is a random vector with each component from Uniform$[0,1]$. In the deterministic case, we had required that any candidate encoder, $\tilde{g}^{-1}$, was point-wise consistent with the pairs of observations (x_{t+1}, x_t). Here, encoders are only required to match the observed conditional distributions. An encoder that is consistent with the observed data can be used to generate $\hat{X}_{t+1}$ such that the distribution of $\hat{X}_{t+1}|X_t = x_t$ matches the distribution of $X_{t+1}|X_t = x_t$ for all $x_t \in \mathcal{X}$,

$$\begin{aligned} X_{t+1}|X_t = x_t &\overset{d}{=} \hat{X}_{t+1}|X_t = x_t \\ g \circ m_t\big(g^{-1}(x_t), U_t\big) &\overset{d}{=} \tilde{g} \circ \tilde{m}_t\big(\tilde{g}^{-1}(x_t), \hat{U}_t\big) \end{aligned} \tag{8}$$

Now, define the set of all candidate decoders $\tilde{g}$ (with corresponding encoders $\tilde{g}^{-1}$) that solve the above equation 7 as $\mathcal{G}^{\mathsf{s}}_{\mathsf{id}}$. We can extend the notion of equivariance and imitation to the stochastic case by replacing equality in value by equality in distribution, and show that, as before, these are the only sources of non-identifiability. In the special case where m_t is known, $\tilde{\mathcal{E}}^{\mathsf{s}}$, defined below, only contains maps that result from equivariance. We continue to use the $\mathcal{M}^*$ – set of mechanisms that are used at least once in the evolution of the system and $\mathcal{M}$ – hypothesis class of all the mechanisms.

Definition 3. ***Equivariance and imitators in distribution w.r.t*** $\mathcal{M}$. *Define a set of functions $\mathcal{E}^{\mathsf{s}}$ that satisfy commutativity w.r.t the set of distributions $\mathcal{M}$ in the following sense. The set $\mathcal{E}^{\mathsf{s}}$ comprises all the diffeomorphisms a that satisfy the following condition: $a \in \mathcal{E}^{\mathsf{s}}$ if and only if for each $m \in \mathcal{M}^*$, $\exists\ m^{'} \in \mathcal{M}$ such that for all $z \in \mathcal{Z}$, $a \circ m(z, U) \overset{d}{=} m^{'}(a(z), U)$, where each component of U is sampled independently from Uniform$[0, 1]$.*

Define the set of maps that identify true g up to $\mathcal{E}^{\mathsf{s}}$ as $\mathcal{G}^{\mathsf{s}}_{\mathsf{eq}} = \{\tilde{g} | \tilde{g} = g \circ a^{-1}, a \in \mathcal{E}^{\mathsf{s}}\}$

Theorem 4. *If the data generation process follows equation 7, then $\mathcal{G}^{\mathsf{s}}_{\mathsf{id}} = \mathcal{G}^{\mathsf{s}}_{\mathsf{eq}}$ and hence, the set of encoders that solve equation 8 identify the true g^{-1} up to the equivariances and imitators in distribution w.r.t $\mathcal{M}$, $\mathcal{E}^{\mathsf{s}}$($\tilde{g}^{-1} \sim_{\mathcal{E}^{\mathsf{s}}} g^{-1}$).*

The proof of the above is provided in Section A.8 in the Appendix. Theorem 4 is consistent the results that we developed for the deterministic case (Theorem 3), but because the mechanism is allowed to be stochastic, it allows us to generalize known results based on distributional assumptions; we give an example of this in the next section.

4 A MECHANISM-BASED PERSPECTIVE ON EXISTING RESULTS

Our primary motivation for understanding how mechanistic knowledge can aid identification, is to develop methods that do not require independence assumptions over the latent variables. However, independence assumptions are not incompatible with the mechanism-based perspective: they simply define a particular kind of mechanism, which then implies identification up to the mechanism's associated equivariances and imitators. We demonstrate this by re-deriving recent identification results from Klindt et al. (2020) using the mechanisms implied by their respective distributional assumptions. We begin by describing the data generation process used by Klindt et al. (2020) as a stochastic mechanism of the form of equation 7. For each $t \in \{1, \cdots, T\}$

$$Z_{t+1} = Z_t + V_t, \quad V_t = f(U_t), \quad U_t \sim \text{Uniform}[0, 1]^d \tag{9}$$

where f is an inverse CDF such that each component of $V_t \in \mathbb{R}^d$ is sampled from the generalized Laplace distribution centred at zero with norm parameter $\alpha \neq 2$, $Z_1 \sim \mathbb{P}_Z$. Next, we want to use Theorem 4 to derive all the solutions to the identity in equation 8.

Theorem 5. *If the data generation process follows equation 9, and $\mathbb{P}_Z$ and the parameters defining f are known (same assumption as in Klindt et al. (2020)), then the solution to equation 8 leads to identifying the true representations up to permutation, sign-flip and offset.*

The proof is given in Section A.9 in the Appendix. The above theorem generalizes Theorem 1 from Klindt et al. (2020) as we do not require $\alpha < 2$ and rather we work with $\alpha \neq 2$. Alternatively, analogous results to those in Klindt et al. (2020) can be derived in settings where we do not know the distribution of V_t but instead assume that we observe X_t often enough that the difference between Z_t and Z_{t+1} is small. In particular, suppose that the data generation process follows equation 9 except each component of V_t is an i.i.d. draw from a non-Gaussian with zero mean and $|V_t| < \delta$. Then as $\delta \to 0$ the true latent is identified up to permutation, sign-flip and offset. See Section A.10 for details.

5 RELATED WORKS

Non-linear independent component analysis (ICA) is a highly unidentified problem; several works (Hyvärinen & Pajunen, 1999; Locatello et al., 2019) have shown that it is impossible to invert the data generation process without placing restrictions on the data and models. In recent years, a lot of progress has been made on the problem of non-linear identification. Hyvarinen & Morioka (2016; 2017) provide the first proofs for identification in non-linear ICA. Hyvarinen & Morioka (2016)

showed that if the latent variables are mutually independent, with each component evolving in time following a non-stationary time series without temporal dependence, then non-linear identification is possible. Hyvarinen & Morioka (2017) showed that non-linear identification is also possible if the latent variables are mutually independent, with each component evolving in time following a stationary time series with temporal dependence. In Hälvä & Hyvarinen (2020), the authors combine non-stationarity (Hyvarinen & Morioka, 2016) and temporal dependency (Hyvarinen & Morioka, 2017) and extend identifiability guarantees in somewhat more general settings. Khemakhem et al. (2020a); Hyvarinen et al. (2019); Khemakhem et al. (2020b) further generalized the previous results; in these works instead of using time the authors require observation of auxiliary information. Klindt et al. (2020) departs from other non-linear ICA works as it explicitly exploits the sparsity in the transitions of the latent variables (further details on Klindt et al. (2020) can be found in the previous section.). In Zimmermann et al. (2021), the authors show that minimizing contrastive losses commonly used in self-supervised learning can also guarantee identification provided the data (contrastive pairs) follow a specific choice of data generation process (e.g., contrastive pair is generated from a von Mises-Fisher distribution). In another line of work Locatello et al. (2020); Shu et al. (2019), the authors study the role of weak supervision in assisting disetanglement. In a recent work, Gresele et al. (2021), propose to add new form of constraints to non-linear ICA. The constraint is based on the observation that the decoder g that gives rise to the image x is composed of simpler functions that are mutually algorithmically independent; the authors exploit this inductive bias on the structure of g to invert the data generation process. In Section 4 we argued that many of the existing distributional assumptions could be interpreted as particular choices of stochastic mechanisms; for more details, see Table 1 in the appendix, where we describe the form of the respective mechanisms. In short, prior work has focused on identification guarantees under assumptions on the dependence between the different random variables, which is in sharp contrast to our approach, which focuses on identification under varying degrees of the knowledge of mechanisms that govern latent dynamics.

Equivariance There is significant recent interest in leveraging equivariance assumptions to design more efficient deep network architectures; for a recent survey, see Bronstein et al. (2021). The general recipe of this line of work is to design functions (deep network architectures) that enforce equivariances. Our setting inverts this recipe, in that we have some known function, $m(z)$, and we are interested in finding all of its equivariances. The relationship between distributions and their equivariances has a long history in the statistics literature (see e.g. Eaton, 1989, and the references therein). Our characterization of stochastic equivariances was inspired by the functional representations given in Bloem-Reddy & Teh (2020). Finally, the importance of group symmetries in representation learning was discussed in Higgins et al. (2018). Higgins et al. focus on the relationship between the symmetries of the environment and a model's representations, whereas we focus on how symmetries in the environment's transition function constrain the representation. The two perspectives are complementary, and in future work we hope to unify them.

6 Discussion and limitations

This paper has presented the first systematic study of how mechanisms governing the dynamics of high-level variables can be used to identify these variables from low-level observations, and up to what equivariances, which depend on the mechanisms. We show that this perspective is both powerful—yielding significant constraints in the space of possible representation—and that it generalizes many known approaches. There are, however, a number of open problems that need to be addressed to make this approach practical. First, we have not specified a particular loss function to select $\tilde{g}$; the observation identity immediately implies a reconstruction-based loss, but other losses such as contrastive losses may be more beneficial. Second, there is not yet a general approach to enforcing invertibility while mapping between different sized spaces, so maintaining bijectivity is a challenge. And finally, while we provide a characterization of identifiability when mechanisms are not known, a natural follow-up question is how to use those insights to build methods that allow one to learn both the mechanism and the representation.

REFERENCES

Tim Austin. Exchangeable random measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 51(3), Aug 2015. ISSN 0246-0203. doi: 10.1214/13-aihp584. URL http://dx.doi.org/10.1214/13-AIHP584.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *J. Mach. Learn. Res.*, 21:90–1, 2020.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL https://arxiv.org/abs/2104.13478.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Morris H DeGroot. *Probability and statistics*. Pearson, 2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Morris L Eaton. Group invariance applications in statistics. In *Regional conference series in Probability and Statistics*, pp. i–133. JSTOR, 1989.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *arXiv preprint arXiv:2106.05200*, 2021.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, 2020.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Robert E Kass and Paul W Vos. *Geometrical foundations of asymptotic inference*, volume 908. John Wiley & Sons, 2011.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020a.

Ilyes Khemakhem, Ricardo Pio Monti, Diederik P Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *arXiv preprint arXiv:2002.11537*, 2020b.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Boris Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Isaac Newton. *Newton's Principia : the Mathematical Principles of Natural Philosophy.* 1687.

Bogdan Nica. The mazur-ulam theorem. *arXiv preprint arXiv:1306.2380*, 2013.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pp. 14866–14876, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. *arXiv preprint arXiv:2102.08850*, 2021.

Approach	Assumptions
Time contrastive (Hyvarinen & Morioka, 2016)	$Z_t \leftarrow U_t$, each U_t^j is independent and non-stationary
Permutation contrastive (Hyvarinen & Morioka, 2017)	Each $Z_{t+1}^j \leftarrow m(Z_t^j, U_t^j)$, stationary and not quasi-Gaussian
Slow VAE (Klindt et al., 2020)	$Z_{t+1} \leftarrow Z_t + f(U_t)$, each $f(U_t^i)$ is independent and generalized Laplace
Conditional VAE (Khemakhem et al., 2020a; Hyvarinen et al., 2019)	$Z \leftarrow m(O, U)$, all components of Z are independent conditional on O
Independently modulated component analysis (Khemakhem et al., 2020b)	$Z \leftarrow m(O, U)$, m has a special structures allowing to relax conditional independence
Contrastive learning (Zimmermann et al., 2021)	$\tilde{Z} \leftarrow m(Z, U)$, m is such that $\tilde{Z} \sim$ conditional von Mises-Fisher

Table 1: Table comparing different works and the assumptions made for identifiability.

A Appendix

A.1 Comparison to existing approaches

A.2 Proof of Theorem 1

Proof. First we show that $\mathcal{G}_{\mathsf{id}} \subseteq \mathcal{G}_{\mathsf{eq}}$.

Consider a $\tilde{g} \in \mathcal{G}_{\mathsf{id}}$. For each $x \in \mathcal{X}$

$$\begin{aligned} g \circ m \circ g^{-1}(x) &= \tilde{g} \circ m \circ \tilde{g}^{-1}(x) \\ \tilde{g}^{-1} \circ \Big(g \circ m \circ g^{-1}(x)\Big) &= \tilde{g}^{-1} \circ \Big(\tilde{g} \circ m \circ \tilde{g}^{-1}(x)\Big) \\ \tilde{g}^{-1} \circ \Big(g \circ m \circ g^{-1}(x)\Big) &= m \circ \tilde{g}^{-1}(x) \end{aligned} \tag{10}$$

Since g is invertible we can substitute x in the above equation with $x = g(z)$ and obtain for each $z \in \mathcal{Z}$

$$\begin{aligned} \Big(\tilde{g}^{-1} \circ g\Big) \circ m \circ \Big(g^{-1} \circ g(z)\Big) &= m \circ \Big(\tilde{g}^{-1} \circ g(z)\Big) \\ \Big(\tilde{g}^{-1} \circ g\Big) \circ m(z) &= m \circ \Big(\tilde{g}^{-1} \circ g\Big)(z) \end{aligned} \tag{11}$$

Define $\tilde{g}^{-1} \circ g = a$. Observe that a is invertible and from equation 11 we gather that $a \in \mathcal{E}$. Also, since $\tilde{g} = g \circ a^{-1}$, we can conclude that $\tilde{g} \in \mathcal{G}_{\mathsf{eq}}$, which proves the first part of the claim. For the second part, we show that $\mathcal{G}_{\mathsf{eq}} \subseteq \mathcal{G}_{\mathsf{id}}$.

Consider a $\tilde{g} \in \mathcal{G}_{\mathsf{eq}} = \{\tilde{g} \mid \tilde{g} = g \circ a^{-1},\ a \in \mathcal{E}\}$. By definition, can express $\tilde{g} = g \circ a^{-1}$. For each $x \in \mathcal{X}$ we write

$$\begin{aligned} \tilde{g} \circ m \circ \tilde{g}^{-1}(x) &= \big(g \circ a^{-1}\big) \circ m \circ \big(g \circ a^{-1}\big)^{-1}(x) \\ \big(g \circ a^{-1}\big) \circ m \circ \big(a \circ g^{-1}\big)(x) &= \big(g \circ a^{-1}\big) \circ m \circ a\big(g^{-1}(x)\big) \\ \big(g \circ a^{-1}\big) \circ a \circ (m(g^{-1}(x)\big) &= g \circ m \circ g^{-1}(x) \end{aligned} \tag{12}$$

Observe that $\tilde{g}$ is both a bijection and satisfies the identity in equation 2. Therefore, $\tilde{g} \in \mathcal{G}_{\mathsf{id}}$. This proves the second part of the claim. Therefore, $\mathcal{G}_{\mathsf{id}} = \mathcal{G}_{\mathsf{eq}}$.

□

A.3 Linear mechanism M and linear G.

In this section, we follow up the analysis that we did more abstractly for general functions g and m in the case when g and m are linear maps. We write the data generation process as follows.

$$\begin{aligned} z_{t+1} &= Mz_t + b, \\ x_t &= Gz_t, \end{aligned} \tag{13}$$

where $M \in \mathbb{R}^{d\times d}$ is a matrix describing the mechanism, $b \in \mathbb{R}^d$ is the offset parameter, $G \in \mathbb{R}^{d\times d}$ is an invertible matrix determining how the data transforms from latent space to the observable space. We write the eigendecomposition of M as follows $M = S\Lambda S^{-1}$, where S is the matrix of eigenvectors and Λ is a diagonal matrix of the eigenvalues. On the same lines as the equation 2, we can obtain an identity between x_t and x_{t+1} as follows.

$$\begin{aligned} z_{t+1} &= Mz_t + b \\ G^{-1}x_{t+1} &= MG^{-1}x_t + b \\ x_{t+1} &= GMG^{-1}x_t + Gb \end{aligned} \tag{14}$$

If the learner knows M and b, it tries to solve for an invertible $\tilde{G}$ that satisfies

$$x_{t+1} = \tilde{G}M\tilde{G}^{-1}x_t + \tilde{G}b \tag{15}$$

Theorem 6. *If the data generation process follows equation 13 and the eigenvalues of the mechanism M are all distinct and each component of the vector $S^{-1}b$ is non-zero, then the only solution to the identity in equation 15 is the true mechanism G.*

Proof. We take the difference of the equations equation 14 and equation 15 to get the following condition. For each $x_t \in \mathbb{R}^d$

$$(GMG^{-1} - \tilde{G}M\tilde{G}^{-1})x_t + (G - \tilde{G})b = 0 \tag{16}$$

Substitute $x_t = 0$ in the above to get

$$(G - \tilde{G})b = 0 \tag{17}$$

We plug the above condition in equation 17 back into equation 16 to get the following condition. For each $x_t \in \mathbb{R}^d$

$$(GMG^{-1} - \tilde{G}M\tilde{G}^{-1})x_t = 0 \tag{18}$$

If we select d linearly independent vectors $x_t \in \mathbb{R}^d$ and substitute in the above equation 18, we get

$$\begin{aligned} &GMG^{-1} - \tilde{G}M\tilde{G}^{-1} = 0 \\ &G^{-1}\Big(GMG^{-1} - \tilde{G}M\tilde{G}^{-1}\Big) = 0 \\ &MG^{-1} - G^{-1}\tilde{G}M\tilde{G}^{-1} = 0 \\ &\Big(MG^{-1} - G^{-1}\tilde{G}M\tilde{G}^{-1}\Big)\tilde{G} = 0 \\ &MG^{-1}\tilde{G} = G^{-1}\tilde{G}M \end{aligned} \tag{19}$$

Let $A = G^{-1}\tilde{G}$. We substitue A and the eigendecomposition of M ($M = S\Lambda S^{-1}$) in equation 19 to get

$$\begin{aligned} &M = AMA^{-1}, \\ &S\Lambda S^{-1} = AS\Lambda S^{-1}A^{-1}, \\ &\Lambda = \Big(S^{-1}AS\Big)\Lambda\Big(S^{-1}A^{-1}S\Big), \\ &\Lambda = C\Lambda C^{-1}, \\ &\Lambda C = C\Lambda, \end{aligned} \tag{20}$$

where $C = S^{-1}AS$, $\Lambda = \mathsf{diag}\Big(\lambda_1, \cdots, \lambda_d\Big)$. We further simplify equation 20 and compare each element of the matrix ΛC and $C\Lambda$ to get the following condition. For all $i, j \in \{1, \cdots, d\}$

$$\begin{aligned} &[\Lambda C]_{ij} = C_{ij}\lambda_i, \quad [C\Lambda]_{ij} = C_{ij}\lambda_j \\ &C_{ij}(\lambda_i - \lambda_j) = 0 \end{aligned} \tag{21}$$

Consider the above equation 21 for $i \neq j$ to get

$$C_{ij}(\lambda_i - \lambda_j) = 0 \implies C_{ij} = 0 \text{ (we use the assumption that } \lambda_i \neq \lambda_j) \tag{22}$$

From the above, it follows that C is a diagonal matrix. We obtain an expression for A in terms of C matrix below.

$$\begin{aligned} &S^{-1}AS = C \\ &A = SCS^{-1} \end{aligned} \tag{23}$$

From equation 17 we get that

$$\begin{aligned} (G - \tilde{G})b = 0 &\overset{\tilde{G}=GA}{\implies} (I - A)b = 0 \implies (I - SCS^{-1})b = 0 \\ \underbrace{\implies}_{\text{Left multiply } S^{-1}} \quad & S^{-1}b - CS^{-1}b \implies (C - I)S^{-1}b = 0 \end{aligned} \tag{24}$$

Since C is a diagonal matrix, we can simplify the above condition further to get $(c_{ii} - 1)(S^{-1}b)_i = 0$. Since $(S^{-1}b)_i \neq 0 \implies c_{ii} = 1$, we obtain $C = I \implies G^{-1}\tilde{G} = I \implies G = \tilde{G}$.

□

A.4 Extending Theorem 1 to multiple known mechanisms

Together m_t and g describe the data-generation from time t to $t+1$ as follows,

$$x_t \leftarrow g(z_t) \qquad z_{t+1} \leftarrow m_t(z_t). \tag{25}$$

In this section, we assume that the learner knows the mechanisms that are used at different times. The learner's goal is to solve the following identity. For each $m \in \mathcal{M}^*$

$$g \circ m \circ g^{-1} = \tilde{g} \circ m \circ \tilde{g}^{-1} \tag{26}$$

Define the set of bijections $\tilde{g}$ that solve the above identity (equation 26) as $\mathcal{G}^*_{\mathsf{id}}$. Define the set of bijections that commute with each $m \in \mathcal{M}^*$ as $\mathcal{E}^* = \{a \mid a \text{ is a bijection, and for each } m \in \mathcal{M}^*, a \circ m = m \circ a\}$. Define the set of bijections that identify true g up to the set $\mathcal{E}^*$ as $\mathcal{G}^*_{\mathsf{eq}} = \{\tilde{g} \mid \tilde{g} = g \circ a^{-1},\ a \in \mathcal{E}^*\}$.

Theorem 7. *If the data generation process follows equation 25, then $\mathcal{G}^*_{\mathsf{id}}$ the set of all the solutions to the identity in equation 26 is equal to $\mathcal{G}^*_{\mathsf{eq}}$. In other words, the set of all the maps that identify true g up to the equivariances shared across mechanisms in $\mathcal{M}$, i.e., $\tilde{g}^{-1} \sim_{\mathcal{E}^*} g^{-1}$.*

Proof. First we show that $\mathcal{G}^*_{\mathsf{id}} \subseteq \mathcal{G}^*_{\mathsf{eq}}$.

Consider a $\tilde{g} \in \mathcal{G}^*_{\mathsf{id}}$. For each $x \in \mathcal{X}$ and for each $m \in \mathcal{M}^*$

$$\begin{aligned} &g \circ m \circ g^{-1}(x) = \tilde{g} \circ m \circ \tilde{g}^{-1}(x) \\ &\tilde{g}^{-1} \circ \Big(g \circ m \circ g^{-1}(x)\Big) = \tilde{g}^{-1} \circ \Big(\tilde{g} \circ m \circ \tilde{g}^{-1}(x)\Big) \\ &\tilde{g}^{-1} \circ \Big(g \circ m \circ g^{-1}(x)\Big) = m \circ \tilde{g}^{-1}(x) \end{aligned} \tag{27}$$

Since g is invertible we can substitute x in the above equation with $x = g(z)$ and obtain for each $z \in \mathcal{Z}$

$$\begin{aligned}\left(\tilde{g}^{-1} \circ g\right) \circ m \circ \left(g^{-1} \circ g(z)\right) &= m \circ \left(\tilde{g}^{-1} \circ g(z)\right) \\ \left(\tilde{g}^{-1} \circ g\right) \circ m(z) &= m \circ \left(\tilde{g}^{-1} \circ g\right)(z)\end{aligned} \tag{28}$$

Define $\tilde{g}^{-1} \circ g = a$. Observe that a is invertible and from equation 28 we gather that $a \in \mathcal{E}^*$. Also, since $\tilde{g} = g \circ a^{-1}$, we can conclude that $\tilde{g} \in \mathcal{G}^*_{\mathsf{eq}}$, which proves the first part of the claim.

In the second part, we need to show that $\mathcal{G}^*_{\mathsf{eq}} \subseteq \mathcal{G}^*_{\mathsf{id}}$.

Consider a $\tilde{g} \in \mathcal{G}^*_{\mathsf{eq}} = \{\tilde{g} \mid \tilde{g} = g \circ a^{-1},\ a \in \mathcal{E}^*\}$. By definition, can express $\tilde{g} = g \circ a^{-1}$. For each $x \in \mathcal{X}$ and for each $m \in \mathcal{M}^*$ we write

$$\begin{aligned}&\tilde{g} \circ m \circ \tilde{g}^{-1}(x) = \left(g \circ a^{-1}\right) \circ m \circ \left(g \circ a^{-1}\right)^{-1}(x) \\ &\left(g \circ a^{-1}\right) \circ m \circ \left(a \circ g^{-1}\right)(x) = \left(g \circ a^{-1}\right) \circ m \circ a\left(g^{-1}(x)\right) \text{ (since } a \text{ commutes with each } m \in \mathcal{M}) \\ &\implies \left(g \circ a^{-1}\right) \circ a \circ \left(m(g^{-1}(x)\right) = g \circ m \circ g^{-1}(x)\end{aligned} \tag{29}$$

Observe that $\tilde{g}$ is both a bijection and satisfies the identity in equation 26. Therefore, $\tilde{g} \in \mathcal{G}^*_{\mathsf{id}}$. This proves the second part of the claim. Therefore, $\mathcal{G}^*_{\mathsf{id}} = \mathcal{G}^*_{\mathsf{eq}}$.

□

A.5 Proof of Theorem 2

Proof. We try to find all the bijections a that are in the set $\mathcal{E}$. If $a \in \mathcal{E}$, then it satisfies $a \circ m = m \circ a$. We replace m with the affine mechanism to obtain the following condition. For each $z \in \mathbb{R}^d$

$$a(Mz + b) = Ma(z) + b \tag{30}$$

Next, we take gradient of the function in the LHS and RHS of the above equation 30 separately w.r.t z. Consider the jth component of $a(Mz + b)$ denoted as $a_j(Mz + b)$. We first take the gradient of $a_j(Mz + b)$ w.r.t z

$$\nabla_z a_j(Mz + b) = \left(\frac{dy}{dz}\right)^{\mathsf{T}} \nabla_y a_j(y), \tag{31}$$

where $y = Mz + b$, $\nabla_y a_j(y)$ is the gradient of a_j w.r.t y and $\frac{dy}{dz}$ denotes the Jacobian of y w.r.t x. We simplify the above further to get

$$\nabla_z a_j(Mz + b) = M^{\mathsf{T}} \nabla_y a_j(y) = M^{\mathsf{T}} \nabla_y a_j(Mz + b) \tag{32}$$

We can write the above for each component of a as follows.

$$\begin{aligned}&\left[\nabla_z a_1(Mz + b), \cdots, \nabla_z a_d(Mz + b)\right] = \left[M^{\mathsf{T}} \nabla_y a_1(Mz + b), \cdots, M^{\mathsf{T}} \nabla_y a_d(Mz + b)\right] \\ &= M^{\mathsf{T}}[\nabla_y a_1(Mz + b), \cdots, \nabla_y a_d(Mz + b)] = M^{\mathsf{T}} J^{\mathsf{T}}(Mz + b),\end{aligned} \tag{33}$$

where $J(Mz + b)$ is the Jacobian of a computed at $Mz + b$. Next, we take the gradient of the jth component of the RHS in equation 30

$$\nabla_z \left[m_j^{\mathsf{T}} a(z) + b_j\right] = \sum_i m_{ji} \nabla_z a_i(z) = \left[\nabla a_1(z), \cdots, \nabla a_d(z)\right] \begin{bmatrix} m_{j1} \\ m_{j2} \\ \vdots \\ m_{jd} \end{bmatrix} \tag{34}$$

We can write the above for each component in the RHS of equation 30 as follows

$$\Big[\nabla_z\big[m_1^\mathsf{T} a(z)+b_1\big],\cdots,\nabla_z\big[m_d^\mathsf{T} a(z)+b_d\big]\Big] = \big[\nabla a_1(z),\cdots,\nabla a_d(z)\big]\begin{bmatrix} m_{11},\cdots,m_{d1} \\ m_{12},\cdots,m_{d2} \\ \vdots \quad\quad \vdots \\ m_{1d},\cdots,m_{dd}\end{bmatrix} = J^\mathsf{T}(z)M^\mathsf{T} \tag{35}$$

We equate the gradient of LHS and RHS in equation 30 using the expressions derived in equation 33 and equation 35 to obtain

$$a(Mz+b) = Ma(z)+b \implies M^\mathsf{T}J^\mathsf{T}(Mz+b) - J^\mathsf{T}(z)M^\mathsf{T} = 0 \tag{36}$$

We write the same expression at another offset $b^{'} \neq b$ below

$$a(Mz+b^{'}) = Ma(z)+b \implies M^\mathsf{T}J^\mathsf{T}(Mz+b^{'}) - J^\mathsf{T}(z)M^\mathsf{T} = 0 \tag{37}$$

Taking the difference of equation 36 and equation 37 we get $M^\mathsf{T}J^\mathsf{T}(Mz+b) = M^\mathsf{T}J^\mathsf{T}(Mz+b^{'})$. Since M is invertible, we get $J(Mz+b) = J(Mz+b^{'})$. Consider row j of this identity. For each $z \in \mathbb{R}^d$, $\exists\ \theta \in \mathbb{R}^d$

$$\nabla a_j(Mz+b) - \nabla a_j(Mz+b^{'}) = 0 \implies \nabla^2 a_j(\theta) = 0 \tag{38}$$

From Assumption 2, it follows that $\nabla^2 a_j(\theta) = 0$ over a measurable set. Since a_j is analytic, we can conclude that $\nabla^2 a_j(\theta) = 0$ over the entire space (Mityagin, 2015). This implies $a(z) = Az+p$. Plug $a(z) = Az+p$ into $a(Mz+b) = Ma(z)+b$ to get

$$A(Mz+b)+p = MA(z+p)+b \implies (AM-MA)z + (A-I)b + (I-M)p = 0 \tag{39}$$

We write the same expression for offset $b^{'}$

$$A(Mz+b^{'})+p = MA(z+p)+b^{'} \implies (AM-MA)z + (A-I)b^{'} + (I-M)p = 0 \tag{40}$$

We take the difference of equation 39 and equation 40 to get

$$(A-I)(b-b^{'}) = 0 \tag{41}$$

Substitue $z=0$ in equation 39 to get

$$(A-I)b + (I-M)p = 0 \tag{42}$$

Substitute the above condition in equation 42 into equation 39 to get the following. For each z

$$(AM-MA)z = 0 \tag{43}$$

The above equation 43 is the same as equation 19 and the equation 41 is the same as equation 24 with b replaced by $b-b^{'}$. Following the same analysis as in the proof of linear case from here on we get that latent variables are exactly identified. We provide the same analysis below for completeness. By choosing d linearly independent z and substituing in equation 43 we get the following

$$\begin{aligned}
&MA = AM \\
&M = AMA^{-1}, \\
&S\Lambda S^{-1} = AS\Lambda S^{-1}A^{-1}, \\
&\Lambda = \Big(S^{-1}AS\Big)\Lambda\Big(S^{-1}A^{-1}S\Big), \\
&\Lambda = C\Lambda C^{-1}, \\
&\Lambda C = C\Lambda,
\end{aligned} \tag{44}$$

where $C = S^{-1}AS$, $M = S\Lambda S^{-1}$, $\Lambda = \mathsf{diag}\Big(\lambda_1, \cdots, \lambda_d\Big)$. We further simplify equation 20 and compare each element of the matrix ΛC and $C\Lambda$ to get the following condition. For all $i, j \in \{1, \cdots, d\}$

$$\begin{aligned} &[\Lambda C]_{ij} = C_{ij}\lambda_i, \;\; [C\Lambda]_{ij} = C_{ij}\lambda_j \\ &C_{ij}(\lambda_i - \lambda_j) = 0 \end{aligned} \tag{45}$$

Consider the above equation 45 for $i \neq j$ to get

$$C_{ij}(\lambda_i - \lambda_j) = 0 \implies C_{ij} = 0 \text{ (we use the assumption that } \lambda_i \neq \lambda_j) \tag{46}$$

From the above, it follows that C is a diagonal matrix. We obtain an expression for A in terms of C matrix below.

$$\begin{aligned} &S^{-1}AS = C \\ &A = SCS^{-1} \end{aligned} \tag{47}$$

From equation 41 we get that

$$\begin{aligned} &(G - \tilde{G})(b - b^{'}) = 0 \overset{\tilde{G}=GA}{\implies} (I - A)(b - b^{'}) = 0 \implies (I - SCS^{-1})b = 0 \\ &\underbrace{\implies}_{\text{Left multiply } S^{-1}} \quad S^{-1}(b - b^{'}) - CS^{-1}(b - b^{'}) \implies (C - I)S^{-1}(b - b^{'}) = 0 \end{aligned} \tag{48}$$

Since C is a diagonal matrix, we can simplify the above condition further to get $(c_{ii} - 1)(S^{-1}(b - b^{'}))_i = 0$. Since $(S^{-1}(b - b^{'}))_i \neq 0 \implies c_{ii} = 1$, we obtain $C = I \implies A = I \implies a(z) = z + p$. This proves that the latents are identified up to an offset. □

A.6 Proof of Theorem 3

Proof. We first show that $\tilde{\mathcal{G}}_{\mathsf{id}} \subseteq \tilde{\mathcal{G}}_{\mathsf{eq}}$. Consider a $\tilde{g} \in \tilde{\mathcal{G}}_{\mathsf{id}}$. We rewrite equation 6 below. For all $x \in \mathcal{X}$

$$\begin{aligned} &g \circ m_t \circ g^{-1}(x) = \tilde{g} \circ \tilde{m}_t \circ \tilde{g}^{-1}(x) \\ &(\tilde{g}^{-1} \circ g) \circ (m_t \circ g^{-1}(x)) = \tilde{m}_t \circ \tilde{g}^{-1}(x) \end{aligned} \tag{49}$$

Since g is bijective, we can write $x = g(z)$ to get

$$(\tilde{g}^{-1} \circ g) \circ m_t(z) = \tilde{m}_t \circ \tilde{g}^{-1} \circ g(z) \tag{50}$$

Since the above equality holds for all $z \in \mathcal{Z}$ we can conclude that

$$\begin{aligned} &(\tilde{g}^{-1} \circ g) \circ m_t = \tilde{m}_t \circ (\tilde{g}^{-1} \circ g), \\ &a \circ m_t = \tilde{m}_t \circ a, \end{aligned} \tag{51}$$

The above conclusion in equation 51 holds for all $m_t \in \mathcal{M}^*$. Therefore, a in equation 51 and $\{\tilde{m}_t\}_{t=1}^{T}$ (where $\tilde{m}_t \in \mathcal{M}$) together satisfy the condition that for all $m \in \mathcal{M}^*$ $a \circ m = \tilde{m} \circ a$, where $\tilde{m} \in \mathcal{M}$. We can rewrite $\tilde{g}^{-1} \circ g = a$ as $\tilde{g} = g \circ a^{-1}$. From this it follows that $\tilde{g} \in \tilde{\mathcal{G}}_{\mathsf{eq}}$. This proves the first part of the theorem.

Now let us consider the second part of the theorem. Consider a $\tilde{g} \in \tilde{\mathcal{G}}_{\mathsf{eq}}$. We can write $\tilde{g} = g \circ a^{-1}$, where $a \in \tilde{\mathcal{E}}$. At time t, some mechanism $m_t \in \mathcal{M}^*$ is used to transform the latents. Since $a \in \tilde{\mathcal{E}}$, select the mechanism $\tilde{m}_t \in \mathcal{M}$ for which $a \circ m_t = \tilde{m}_t \circ a$ and as a consequence

$$g \circ m_t \circ g^{-1} = g \circ a^{-1} \circ \tilde{m}_t \circ a \circ g^{-1} = \tilde{g} \circ \tilde{m}_t \circ \tilde{g}^{-1} \tag{52}$$

In the first equality above, we use $a \circ m_t = \tilde{m}_t \circ a$ and in the second equality we use the definition of $\tilde{g}$. We can repeat the above exercise for all t and corresponding m_t using the same a. Therefore, $\tilde{g}$ is in $\tilde{\mathcal{G}}_{\mathsf{id}}$. This shows the second part of the theorem, i.e., $\tilde{\mathcal{G}}_{\mathsf{eq}} \subseteq \tilde{\mathcal{G}}_{\mathsf{id}}$.

□

A.7 LEVERAGING THEOREM 3 WHEN MECHANISM IS LINEAR AND g IS NON-LINEAR

We write the data generation process as follows. For each $t \in \{1, \cdots, d+1\}$

$$\begin{aligned} z_{t+1} &= M_t z_t + b_t, \\ x_t &= g(z_t), \end{aligned} \tag{53}$$

Let us construct the set $\tilde{\mathcal{E}}$ corresponding to the above setting. For each $z \in \mathbb{R}^d$,

$$\begin{aligned} a(Mz+b) = M^{'}a(z) + b^{'} \implies M^{\mathsf{T}}J^{\mathsf{T}}(Mz+b) - J^{\mathsf{T}}(z)M^{',\mathsf{T}} = 0, \\ a(Mz+\tilde{b}) = M^{'}a(z) + \tilde{b}^{'} \implies M^{\mathsf{T}}J^{\mathsf{T}}(Mz+b^{'}) - J^{\mathsf{T}}(z)M^{',\mathsf{T}} = 0, \end{aligned} \tag{54}$$

where (M, b) and $(M, \tilde{b})$ are the true mechanisms and $(M^{'}, b^{'})$ and $(M', \tilde{b}^{'})$ are the imtitating mechanisms chosen by the learner, J is the Jacobian of a. Note here the learner only exploits the knowledge that b changes to $\tilde{b}$, which is why it keeps $M^{'}$ fixed and only changes the offset. We take the difference of the RHS in the above two equations to get

$$M^{\mathsf{T}}J^{\mathsf{T}}(Mz+b) = M^{\mathsf{T}}J^{\mathsf{T}}(Mz+b^{'}) \tag{55}$$

Since M is invertible we get $J(Mz+b) - J(Mz+b^{'}) = 0$ for all z. We can follow the same justification as was used in equation 37 to conclude that $J(z)$ is constant and a is thus an affine map. We substitue the affine map $Az + p$ back into equation 54 to get the following. For all z

$$AMz + Ab + p = M^{'}Az + b^{'} + M^{'}p \tag{56}$$

Substitute $z = 0$ to get $b^{'} = Ab + p - M^{'}p$. Substitute this condition back into the above equation, we get $AM = M^{'}A \implies M^{'} = AMA^{-1}$. We can carry out the same excerise for the second equation in equation 54 and get $M^{'} = AMA^{-1}$ and $\tilde{b}^{'} = A\tilde{b} + p - M^{'}p$.

A.8 PROOF OF THEOREM 4

Before stating the proof of Theorem 4, we state two existing results that we use.

Result 1. (Change of variables formula (DeGroot, 2012)) Given a continuous random variable $X \in \mathbb{R}^d$ with pdf p_X and its transformation $Y = f(X)$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism,[8] then $p_Y(f(x))|\mathsf{det}(J_f(x))| = p_X(x)$, where J_f is the Jacobian of f computed at x.

Lemma 1. *If X and Y are two continuous random variables that take values in $\mathbb{R}^d$ that are equal in distribution, i.e., $X \overset{d}{=} Y$. If $f : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism, then $f(X) \overset{d}{=} f(Y)$.*

Proof. Since X and Y are equal in distribution, they have the same pdfs, i.e. $p_X(x) = p_Y(x)$ for all $x \in \mathbb{R}^d$. We can use the change of variables formula in Resut 1 above to get the following. Let $W = f(X)$, $p_X(f^{-1}(w))|\mathsf{det}(J_{f^{-1}}(w))| = p_W(w)$ and let $V = f(Y)$, $p_Y(f^{-1}(v))|\mathsf{det}(J_{f^{-1}}(v))| = p_V(v)$. Comparing the two expressions when $w = v$ we get $p_W(w) = p_V(w)$. This proves the result. □

We stated Result 1 and Lemma 1 for continuous random variables. When the random variables are discrete, Lemma 1 holds for any function f.

Lemma 2. *(Kass & Vos, 2011) If $f : \mathcal{Z} \to \mathcal{Z}$ and $g : \mathcal{Z} \to \mathcal{Z}$ are diffeomorphisms, then $f \circ g$ is a diffeomorphism.*

[8] http://math.mit.edu/~larsh/teaching/F2007/handouts/changeofvariables.pdf

Proof. We first show that $\mathcal{G}^{\mathsf{s}}_{\mathsf{id}} \subseteq \mathcal{G}^{\mathsf{s}}_{\mathsf{eq}}$.

From the identity in equation 8 we get that $\tilde{g}, \{\tilde{m}_t\}_{t=1}^T$ satisfy the following for all $t \in \{1, \cdots, T\}$

$$\begin{aligned} g \circ m_t\big(g^{-1}(x_t), U_t\big) &\stackrel{d}{=} \tilde{g} \circ \tilde{m}_t\big(\tilde{g}^{-1}(x_t), \hat{U}_t\big) \\ \Big(\tilde{g}^{-1} \circ g\Big) \circ m_t\big(g^{-1}(x_t), U_t\big) &\stackrel{d}{=} \tilde{m}_t\big(\tilde{g}^{-1}(x_t), \hat{U}_t\big) \end{aligned} \tag{57}$$

In the second step in the above equation, we transformed the random variables in the first step using the same transform $\tilde{g}^{-1}$. $\tilde{g}^{-1}$ is a diffeomorphism; we compose both sides of the first step LHS and RHS with $\tilde{g}^{-1}$. We use Lemma 1 to get from the first step to the second step in the above equation equation 57. In the above equation x_t is a fixed value and the only source of randomness is from U_t in LHS and $\hat{U}_t$ in the RHS. We substitute $x_t = g(z_t)$ to further simplify the above expression in equation 57

$$\begin{aligned} \Big(\tilde{g}^{-1} \circ g\Big) \circ m_t\big(g^{-1} \circ g(z_t), U_t\big) &\stackrel{d}{=} \tilde{m}_t\big(\tilde{g}^{-1} \circ g(z_t), \hat{U}_t\big) \\ \Big(\tilde{g}^{-1} \circ g\Big) \circ m_t\big(z_t, u_t\big) &\stackrel{d}{=} \tilde{m}_t\big(\tilde{g}^{-1} \circ g(z_t), \hat{U}_t\big) \end{aligned} \tag{58}$$

Substitute $a = \tilde{g}^{-1} \circ g$ in the above to get the following

$$\begin{aligned} a \circ m_t\big(z_t, U_t\big) &\stackrel{d}{=} \tilde{m}_t\big(a(z_t), \hat{U}_t\big) \\ a \circ m_t\big(z_t, U_t\big) &\stackrel{d}{=} \tilde{m}_t\big(a(z_t), U_t\big) \end{aligned} \tag{59}$$

From Lemma 2 it follows that a in the above is a diffeomorphism. Since we assume that $\cup_{t=1}^T \{m_t\} = \mathcal{M}^*$ it follows that a in equation 51 and $\{\tilde{m}_t\}_{t=1}^T$ (where $\tilde{m}_t \in \mathcal{M}$) together satisfy the condition for membership in $\mathcal{E}^{\mathsf{s}}$. Since $\tilde{g} = g \circ a^{-1}$ we obtain that $\tilde{g} \in \mathcal{G}^{\mathsf{s}}_{\mathsf{eq}}$.

We now show that $\mathcal{G}^{\mathsf{s}}_{\mathsf{eq}} \subseteq \mathcal{G}^{\mathsf{s}}_{\mathsf{id}}$. Consider a $\tilde{g} \in \mathcal{G}^{\mathsf{s}}_{\mathsf{eq}}$. We use $\tilde{g} = g \circ a^{-1}$, where $a \in \mathcal{E}^{\mathsf{s}}$ to simplify the following random variable

$$g \circ m_t(g^{-1}(x_t), U_t) = \Big(g \circ a^{-1} \circ a\Big) \circ m_t(g^{-1}(x_t), U_t) = \tilde{g} \circ a \circ m_t(g^{-1}(x_t), U_t) \tag{60}$$

Since $a \in \mathcal{E}^{\mathsf{s}}$ we have

$$a \circ m_t(g^{-1}(_t), U_t) \stackrel{d}{=} \tilde{m}_t\big(a \circ g^{-1}(x_t), U_t\big)$$

From Lemma 2 it follows that $\tilde{g}$ is a diffeomorphism. From Lemma 1 it follows that

$$\tilde{g} \circ a \circ m_t(g^{-1}(x_t), U_t) \stackrel{d}{=} \tilde{g} \circ \tilde{m}_t(a \circ g^{-1}(x_t), U_t) = \tilde{g} \circ \tilde{m}_t(\tilde{g}^{-1}(x_t), U_t) \tag{61}$$

Combining equation 60 and equation 61 and using the fact that $\hat{U}_t \stackrel{d}{=} U_t$ we get

$$g \circ m_t(g^{-1}(x_t), U_t) \stackrel{d}{=} \tilde{g} \circ \tilde{m}_t(\tilde{g}^{-1}(x_t), U_t) \stackrel{d}{=} \tilde{g} \circ \tilde{m}_t(\tilde{g}^{-1}(x_t), \hat{U}_t) \tag{62}$$

From the definition of $\mathcal{E}^{\mathsf{s}}$ it follows with the same choice of a the condition continues to hold for all $m_t \in \mathcal{M}^*$. Therefore, $\tilde{g} \in \mathcal{G}^{\mathsf{s}}_{\mathsf{id}}$. This proves the second part of the theorem.

□

A.9 PROOF OF THEOREM 5

Proof. In Theorem 4, we showed that all the solutions to the identity in equation 8 can be characterized in terms of the equivariances in distribution defined by the set $\mathcal{E}^{\mathsf{s}}$. Let us analyze the set $\mathcal{E}^{\mathsf{s}}$ for the class of mechanisms considered in Klindt et al. (2020). Consider a $a \in \mathcal{E}^{\mathsf{s}}$. For each $z \in \mathbb{R}^d$

$$a(z + V) \stackrel{d}{=} a(z) + \hat{V}, \tag{63}$$

Define $\hat{Y} = a(z) + \hat{V}$. Since $\hat{V} \stackrel{d}{=} V$ we write the probability density function (pdf) of $\hat{Y}$ as

$$f_{\hat{Y}}(y) = f_V(y - a(z)) \tag{64}$$

Define $Y = a(z + V)$. $a : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism. We use the change of variables result (Result 1) to write the pdf Y as follows. For each $y \in \mathbb{R}^d$

$$f_Y(y) = \frac{1}{\left|\mathsf{det}\big(J\big(a^{-1}(y)\big)\big)\right|} f_V\big(a^{-1}(y) - z\big), \tag{65}$$

where $J(a^{-1}(y))$ is the Jacobian of a computed at $a^{-1}(y)$, and det is the determinant.

We substitute equation 64 and equation 65 in the equivariance condition in equation 63 to obtain the following. For each $y \in \mathbb{R}^d$

$$\begin{aligned} Y &\stackrel{d}{=} \hat{Y} \\ f_V(y - a(z)) &= \frac{f_V(a^{-1}(y) - z)}{|\mathsf{det}(J(a^{-1}(y)))|} \\ f_V(a(w) - a(z)) &= \frac{f_V(w - z)}{|\mathsf{det}(J(w))|}, \end{aligned} \tag{66}$$

where $w = a^{-1}(y)$. In the above we equated the conditionals for each z, we now equate the marginals.

$$\begin{aligned} g \circ (Z + V) &\stackrel{d}{=} \tilde{g} \circ (\hat{Z} + \hat{V}) \\ a \circ (Z + V) &\stackrel{d}{=} \hat{Z} + \hat{V} \end{aligned} \tag{67}$$

We follow Klindt et al. (2020) and assume $\hat{Z} \stackrel{d}{=} Z$ and $\hat{V} \stackrel{d}{=} V$. Therefore, $Z + V \stackrel{d}{=} \hat{Z} + \hat{V}$. We use this condition to restate equation 67 as

$$a \circ (Z + V) \stackrel{d}{=} Z + V \implies a(W) \stackrel{d}{=} W, \tag{68}$$

where $W = Z + V$. We translate equation 68 into the condition on the pdfs as follows. For each $w \in \mathbb{R}^d$

$$f_W(w)|\mathsf{det}(J(w))| = f_W(w) \implies |\mathsf{det}(J(w))| = 1 \tag{69}$$

Substituting the above equation equation 69 into equation 66 we get

$$\begin{aligned} &f_V(a(w) - a(z)) = \frac{f_V(w - z)}{|\mathsf{det}(J(w))|} \implies f_V(a(w) - a(z)) = f_V(w - z) \\ &\|a(w) - a(z)\|_\alpha = \|w - z\|_\alpha, \end{aligned} \tag{70}$$

where in the last condition in the above expression we exploit the fact that f_V is a generalized Laplacian distribution. From Mazur-Ulam theorem Nica (2013) it follows that a is affine. We now write a as a matrix A with offset vector q and simplify the condition in equation 63.

For each $z \in \mathbb{R}^d$ we have

$$\begin{aligned} A(z+V)+q &\stackrel{d}{=} Az+\hat{V}+q \\ &\implies AV \stackrel{d}{=} \hat{V} \\ \implies \mathbb{E}[AVV^{\mathsf{T}}A^{\mathsf{T}}] = \mathbb{E}[\hat{V}\hat{V}^{\mathsf{T}}] &\implies AA^{\mathsf{T}} = \mathsf{I} \end{aligned} \tag{71}$$

Since A is a square matrix and AA^{T} it follows that $A^{\mathsf{T}}A = \mathsf{I}$. Therefore, A is an orthonormal matrix. Observe that all the elements of $\hat{V}$ are independent. Since $AV \stackrel{d}{=} \hat{V}$ it follows that all the elements of AV are also independent. Define $AV = Q$. Observe that A is an orthonormal matrix that is multiplied with a vector V with all independent elements (each of which is non-Gaussian as $\alpha \neq 2$) and outputs a vector that has all independent components. From Theorem 11 in Comon (1994) we get that A is a composition of permutation and scaling. Since A is also orthonormal, each term in the diagonal scaling matrix can only be 1 or -1. Therefore, $A = \Pi\Lambda$, where Π is a permutation matrix and Λ is a diagonal matrix with $+1, -1$ elements. Finally, $a(z) = \Pi\Lambda z + q$.

□

A.10 Alternative identification result for small transitions

Let us analyze the set $\mathcal{E}^{\mathsf{s}}$ for this class of mechanisms. We assume that the learner knows that the mechanism is additive, and that the noise components are all independent. In the analysis below we consider bijections that are analytic (each component of the bijection is an analytic function). Consider an $a \in \mathcal{E}^{\mathsf{s}}$

$$a(z+V) \stackrel{d}{=} a(z) + \hat{V}. \tag{72}$$

We write the first-order approximation of the above identity below

$$\begin{aligned} &a(z) + \mathsf{J}(z)V \stackrel{d}{=} a(z) + \hat{V} \\ &\mathsf{J}(z)V \stackrel{d}{=} \hat{V} \end{aligned} \tag{73}$$

where $V \stackrel{d}{\neq} \hat{V}$. Note that the set of solutions a to equation 72 and equation 73 become equal in the limit of $\delta \to 0$, where δ is the bound on each component of $|V|$. We analyze the solution to equation 73 below.

$$\begin{aligned} &\mathbb{E}[(\mathsf{J}(z)V)(\mathsf{J}(z)V)^{\mathsf{T}}] = \mathsf{J}(z)\mathbb{E}[VV^{\mathsf{T}}]\mathsf{J}(z)^{\mathsf{T}} = \sigma^2\mathsf{J}(z)\mathsf{J}(z)^{\mathsf{T}} \\ &\sigma^2\mathbb{E}[\hat{V}\hat{V}^{\mathsf{T}}] = \sigma^2\mathsf{I} \\ &\sigma^2\mathsf{J}(z)\mathsf{J}(z)^{\mathsf{T}} = \sigma^2\mathsf{I} \implies \mathsf{J}(z)\mathsf{J}(z)^{\mathsf{T}} = \mathsf{I} \implies \mathsf{J}(z)^{\mathsf{T}}\mathsf{J}(z) = \mathsf{I} \end{aligned} \tag{74}$$

Since $\mathsf{J}(z)V \stackrel{d}{=} \hat{V}$ and each component of $\hat{V}$ is independent, we can deduce that all the components of $\mathsf{J}(z)V$ are independent as well. From Theorem 11 in Comon (1994), we can deduce that $\mathsf{J}(z)$ is composed of permutation times a diagonal matrix. Since the matrix is orthonormal, each scaling component can only be ± 1. We can apply this same analysis at another point $\tilde{z}$ in the neighborhood of z and continue to find that the jacobian matrix is a permutation times diagonal matrix (that describes sign flips). Note that the permutation matrix times scaling used to express the Jacobian cannot change between the points $\tilde{z}$ and z (if it does change then that violates the Jacobian's continuity). Since the Jacobian is equal to a fixed permutation times a fixed scaling matrix over a neighborhood, we can extend this to the entire space (here we use the fact that the a is analytic and Mityagin (2015)).

A.11 Analyzing $\tilde{\mathcal{E}}$ when $\mathcal{M} = \mathcal{M}^*$

In this section, we analyze imitators when we know the set of mechanisms that are deployed, we do not know which mechanism is used when. If $a \in \tilde{\mathcal{E}}$ and $\mathcal{M} = \mathcal{M}^*$. From the definition of a, it follows that for each $m \in \mathcal{M}^*$, $\exists\, m^{'} \in \mathcal{M}^*$ such that $a \circ m = m^{'} \circ a$. We claim that two distinct $m \in \mathcal{M}^*$ cannot share the same $m^{'}$ (imitator). Suppose there was a common $m^{'}$ imitating m and $\tilde{m}$.

$$\begin{aligned} a \circ m &= m^{'} \circ a \\ a \circ \tilde{m} &= m^{'} \circ a \end{aligned} \tag{75}$$

We take the difference of the above two equations to get

$$a \circ m = a \circ \tilde{m} \tag{76}$$

Since a is a bijection, we can conclude that $m = \tilde{m}$, which is a contradiction of the fact that m and $\tilde{m}$ are distinct.

This claim implies that for a given a there is an injective map from $\mathcal{M}^*$ to $\mathcal{M}^*$. If the set $\mathcal{M}^*$ is finite, then from Pigeonhole principle it follows that this injective map is a bijection.

Let us index the mechanism $\mathcal{M}^* = \{m^1, \cdots, m^n\}$. We call the bijection map $\pi : \mathcal{M}^* \to \mathcal{M}^*$

Consider the element i. We claim $\exists\ l \in \{1, \cdots, n\}\ \pi^l(i) = i$. We write the chain starting from i as $i \to \pi(i) \to \pi^2(i), \cdots, \pi^k(i)$. Since the chain $(\pi(i) \to \pi^2(i), \cdots, \pi^k(i))$ has n steps there have to be at least two elements that are equal. Suppose $p > q$ and $\pi^p(i) = \pi^q(i)$

$$\pi^p(i) = \pi^q(i) \implies \pi^{p-1}(i) = \pi^{q-1}(i) \implies \cdots ..\pi^{p-q}(i) = i \tag{77}$$

In the above at each step we use the fact that π is a bijection and that shows the claim that $\pi^l(i) = i$. We now use this observation to carry out the following simplification

$$\begin{aligned} m^i &= a^{-1} \circ m^{\pi(i)} \circ a \\ m^{\pi(i)} &= a^{-1} \circ m^{\pi^2(i)} \circ a \\ &\vdots \\ m^{\pi^k(i)} &= a^{-1} \circ m^i \circ a \end{aligned} \tag{78}$$

Substituting the second equation $m^{\pi(i)}$ into first, and then the third $m^{\pi^2(i)}$ and so on we get

$$m^i = a^{-k} \circ m^i \circ a^k \tag{79}$$

Therefore, for each m^i, $\exists\ k$ such that a^k is its equivariance.

To summarize, if $a \in \tilde{\mathcal{E}}$ and $\mathcal{M} = \mathcal{M}^*$, where $\mathcal{M}$ is a finite set, then for each mechanism $m \in \mathcal{M}$, $\exists k \in \{1, \cdots, |\mathcal{M}|\}$ such that a^k is its equivariance.