A Unifying Framework for Unsupervised Concept Extraction

Chandler Squires
Machine Learning Department, CMU

Pradeep Ravikumar
Machine Learning Department, CMU

Abstract

Techniques for *concept extraction*, such as sparse autoencoders and transcoders, aim to extract high-level symbolic concepts from low-level nonsymbolic representations. When these extracted concepts are used for downstream tasks such as model steering and unlearning, it is essential to understand their guarantees, or lack thereof. In this work, we present a unified theoretical framework for unsupervised concept extraction, in which we frame the task of concept extraction as identifying a generative model. We present a general meta-theorem for identifiability, which reduces the problem of establishing identifiability guarantees to the problem of characterizing the intersection of two sets. As we demonstrate on a range of widely-used approaches, this meta-theorem substantially simplifies the task of proving such guarantees, thus paving the way for the development of new, principled approaches for concept extraction.

1 INTRODUCTION

As large neural networks have become the de facto standard for many machine learning tasks, there has been growing interest in understanding the internal representations that they learn. Such an understanding promises to enable several downstream capabilities, e.g. intervening on these representations can be an effective approach for post-training tasks such as model steering, unlearning, and representation alignment (Turner et al., 2023; Ashuach et al., 2025; Sucholutsky et al., 2023).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

A key challenge to gaining such an understanding is that neural representations are often *entangled* and *distributed*, rather than *disentangled* and *local*. More formally, consider a pre-trained encoder $\Phi\colon \mathcal{X} \to \mathcal{Z}$ mapping from a high-dimensional input space $\mathcal{X}$ to a neural *feature space* $\mathcal{Z}$. For example, Φ may be the first 10 layers of an image classification model, which maps an image $x \in \mathcal{X}$ to a neural activation vector $z = \Phi(x) \in \mathcal{Z}$. Typically, the individual coordinates of z are *polysemantic*, i.e., each coordinate z_i represents a combination of many human-interpretable features rather than a single one (Elhage et al., 2022).

To address this challenge, a common approach is to extract high-level *concepts* from these neural representations, e.g. by training a sparse autoencoder to reconstruct the neural representations (Cunningham et al., 2023). Such approaches specify a *concept space* $\mathcal{C}$ and find some *concept extractor* $h\colon \mathcal{Z} \to \mathcal{C}$. The concept extractor is typically taken to be injective, so that an intervention on the concepts (i.e., a function $f_c\colon \mathcal{C} \to \mathcal{C}$) can be translated into a corresponding intervention $f_z = h^{-1} \circ f_c \circ h$ on the neural features.

Unfortunately, many existing approaches for concept extraction do not reliably recover a "canonical" set of features, e.g. for different random seeds, the same sparse autoencoder can recover very different concept extractors h and h' from the same data (Leask et al., 2025; Paulo and Belrose, 2025). This ambiguity both hinders interpretability and casts doubt on the prospect of using the extracted concepts for downstream tasks, since the same concept-level intervention will result in different feature-level interventions (Méloux et al., 2025; Song et al., 2025). To avoid these issues, it is essential to treat the task of concept extraction more rigorously, and develop new methods with appropriate uniqueness guarantees.

In this work, we consider the task of concept extraction from a rigorous statistical viewpoint, using the tools of *identifiability theory*. We treat concepts as latent random variables $\mathbf{C}$ sampled from some unknown distribution Q belonging to some known class

$\mathcal{Q}$, and neural features as observed random variables $\mathbf{Z}$ which depend on $\mathbf{C}$ in an unknown fashion.

In the deterministic setting (i.e., $\mathbf{Z} = g(\mathbf{C})$ for some unknown, injective g), the task of concept extraction is thus equivalent to identifying $h = g^{-1}$ up to simple symmetries, e.g. relabelings of the concepts. In this setting, the choice of class $\mathcal{Q}$ can be seen as an *explicit* description of the *implicit* assumptions imposed by different sparse autoencoders (Hindupur et al., 2025). More generally, our framework permits *stochastic* relationships between concepts and features, opening the door to a variety of more realistic cases.

Contributions Our contributions are four-fold:

1. We introduce a new theoretical framework for unsupervised concept extraction based on *latent concept generative models (LC-GMs)*, which are models that can be decomposed into a *concept generator* Q and a *mixing kernel* K (Sec. 3).
2. We present general meta-theorems for the identifiability of a class $\mathcal{M}$ of LC-GMs (Sec. 4). Most importantly, our *intersection theorem* (Thm. 4.2) characterizes identifiability via the intersection of two *valid transition sets*, extending prior work by Xi and Bloem-Reddy (2023) to model classes which allow for stochastic mixing.
3. To highlight the meta-theoretical utility of our framework, we show how our theorems can recover diverse existing identifiability results, e.g. in dictionary learning, independent component analysis, and mixture modeling (Sec. 5, App. F).
4. To highlight the potential practical utility of our framework, we discuss its implications for method development, including connections between concept extraction, generative modeling, and amortized posterior inference (Sec. 6).

2 NOTATION

As is common in identifiability theory, our results are given in the *population setting*, i.e., we work directly with probability distributions rather than random variables and samples. As such, the mathematical context for this work is measure-theoretic probability, see App. A for key definitions and references.

In this paper, we assume all spaces are ***Polish Borel spaces***. These spaces enjoy several useful properties, which we introduce as needed. Two basic examples include the sets $\mathbb{R}^d$ and $\{1, 2, \dots, d\}$ with their standard topologies and σ-algebras. In this section, we let $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ be Polish Borel spaces.

Distributions For any measure μ on $\mathcal{V}$ (including probability distributions), we define the ***support*** of μ, denoted $\operatorname{supp}(\mu)$, as the smallest closed set $F \subseteq \mathcal{V}$ such that $\mu(\mathcal{V} \setminus F) = 0$. We let $\mathcal{P}(\mathcal{V})$ denote the set of probability distributions on $\mathcal{V}$. We say that Q is ***absolutely continuous*** with respect to a measure μ if $\mu(A) = 0$ implies $Q(A) = 0$ for any measurable set A, written $Q \ll \mu$. If $Q \ll \mu$ and $\mu \ll Q$, we write $Q \sim \mu$. Finally, given a point $v \in \mathcal{V}$, we let $\delta_v \in \mathcal{P}(\mathcal{V})$ denote the delta distribution at v.

Markov Kernels We let $\mathcal{K}(\mathcal{U} \to \mathcal{V})$ denote the set of Markov kernels from $\mathcal{U}$ to $\mathcal{V}$. For a Markov kernel K from $\mathcal{U}$ to $\mathcal{V}$, we let $K(\cdot \mid u)$ denote the associated probability distribution under $u \in \mathcal{U}$. Given two Markov kernels $K, K' \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$ and a measure μ on $\mathcal{U}$, we write $K \sim_\mu K'$ to denote that K and K' are equal μ-almost everywhere (abbreviated μ-a.e.).

For $K \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$, we define the ***pushforward operator*** $K_\sharp \colon \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{V})$ as follows:

$$K_\sharp \colon Q \mapsto \textstyle\int_{\mathcal{U}} K(\cdot \mid u) \cdot Q(du).$$

A ***posterior kernel*** of K with respect to Q is a Markov kernel $L \in \mathcal{K}(\mathcal{V} \to \mathcal{U})$ which satisfies the condition $K(dv \mid du) \cdot Q(du) = L(du \mid v) \cdot P(dv)$, where $P = K_\sharp(Q)$. In Polish Borel spaces, a posterior kernel always exists and is unique P-a.e.; we denote any such kernel as $K_Q^\dagger$.

We say that K is ***measure-separating*** if its associated pushforward operator $K_\sharp$ is injective. We let $\mathcal{K}_{\text{sep}}(\mathcal{U} \to \mathcal{V})$ denote the set of measure-separating Markov kernels from $\mathcal{U}$ to $\mathcal{V}$.[1] As a special case, we let $\mathcal{F}_{\text{inj}}(\mathcal{U} \to \mathcal{V})$ denote the set of injective measurable functions from $\mathcal{U}$ to $\mathcal{V}$, and we let $\mathcal{F}_{\text{c-inj}}(\mathcal{U} \to \mathcal{V})$ denote the set of continuous injective measurable functions from $\mathcal{U}$ to $\mathcal{V}$. For a measurable function $g \colon \mathcal{U} \to \mathcal{V}$, we let $E_g \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$ denote the ***Dirac kernel*** associated with g, i.e., $E_g \colon u \mapsto \delta_{g(u)}$.

Given two Markov kernels $K \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$ and $L \in \mathcal{K}(\mathcal{V} \to \mathcal{W})$, we let $LK \in \mathcal{K}(\mathcal{U} \to \mathcal{W})$ denote their composition, defined as

$$(LK) \colon u \mapsto \textstyle\int_{\mathcal{V}} L(\cdot \mid v) \cdot K(dv \mid u).$$

For Polish Borel spaces, pushforward and composition can be interchanged, i.e., $(LK)_\sharp = L_\sharp \circ K_\sharp$. We refer to this property as ***functoriality***. Since a probability distribution $Q \in \mathcal{P}(\mathcal{U})$ can be treated as Markov kernel from the empty set to $\mathcal{U}$, we also use the notation $KQ = K_\sharp(Q)$. In particular, by functoriality, $(LK)(Q) = L(KQ)$.

[1] More generally, we say K is ***measure-separating on*** $\mathcal{P}' \subseteq \mathcal{P}$ if the restriction $K_\sharp|_{\mathcal{P}'}$ is injective, and we denote the set of such kernels as $\mathcal{K}_{\text{sep}}(\mathcal{U} \to \mathcal{V}; \mathcal{P}')$.

Composition Notation For two sets of Markov kernels $\mathcal{K} \subseteq \mathcal{K}(\mathcal{U} \to \mathcal{V})$ and $\mathcal{L} \subseteq \mathcal{K}(\mathcal{V} \to \mathcal{W})$, we let

$$\mathcal{L} \circ \mathcal{K} := \{LK : L \in \mathcal{L}, K \in \mathcal{K}\}.$$

In App. A, we give an extension of this notation when $\mathcal{L}$ or $\mathcal{K}$ are replaced by sets of functions, or when $\mathcal{K}$ is replaced by a set of probability distributions.

Matrix-Vector Notation For several examples, we use discrete spaces $\mathcal{U}$ and $\mathcal{V}$. In such settings, distributions and Markov kernels can be conveniently represented by (labelled) vectors and matrices, respectively. In this paper, we use column-stochastic convention, i.e., distributions are column vectors, and Markov kernels are column-stochastic matrices. For example, consider $\mathcal{U} = \{a, b\}$ and $\mathcal{V} = \{u, v\}$. Then

$$Q = \begin{matrix} a \\ b \end{matrix} \begin{pmatrix} \beta_a \\ \beta_b \end{pmatrix} \quad \text{and} \quad K = \begin{matrix} u \\ v \end{matrix} \overset{\begin{matrix} a & b \end{matrix}}{\begin{pmatrix} \beta_{au} & \beta_{bu} \\ \beta_{av} & \beta_{bv} \end{pmatrix}}$$

are the distribution $Q = \beta_a \cdot \delta_a + \beta_b \cdot \delta_b$ and the Markov kernel K with $K(\cdot \mid a) = \beta_{au} \cdot \delta_u + \beta_{av} \cdot \delta_v$ and $K(\cdot \mid b) = \beta_{bu} \cdot \delta_u + \beta_{bv} \cdot \delta_v$, respectively.

3 LATENT CONCEPT GENERATIVE MODELS

Let $\mathcal{C}$ and $\mathcal{Z}$ be Polish Borel spaces, which we call a ***concept space*** and ***feature space***, respectively. Though we typically consider $\mathcal{C} \subseteq \mathbb{R}^d$ or $\mathcal{C} = \{1, 2, \dots, d\}$, we note that the Polish Borel condition is quite weak. Thus, our results can be applied to many complex spaces, e.g. some function spaces or spaces of structured objects such as graphs. A ***concept distribution*** is a distribution over $\mathcal{C}$, and a ***mixing kernel*** is a Markov kernel from $\mathcal{C}$ to $\mathcal{Z}$.

3.1 Basic Definitions

As described in Sec. 1, we consider the task of concept extraction from a statistical viewpoint, in which the concepts are latent random variables which generate the observed neural features. We formalize this generation process as a ***latent concept generative model (LC-GM)*** M from $\mathcal{C}$ to $\mathcal{Z}$, which is a tuple $\mathsf{M} = (Q, K)$ with $Q \in \mathcal{P}(\mathcal{C})$ and $K \in \mathcal{K}(\mathcal{C} \to \mathcal{Z})$.[2]

Given a model $\mathsf{M} = (Q, K)$, the ***induced feature distribution*** of M, denoted $P^{\mathsf{M}} \in \mathcal{P}(\mathcal{Z})$, is

$$P^{\mathsf{M}} := KQ$$

i.e., P^{M} represents the marginal distribution over features $\mathbf{Z}$ when first sampling concepts $\mathbf{C}$ from Q, then sampling $\mathbf{Z}$ from the distribution $K(\cdot \mid \mathbf{C})$.

Similarly, the ***induced concept extractor***, denoted $H^{\mathsf{M}} \in \mathcal{K}(\mathcal{Z} \to \mathcal{C})$, is defined as $H^{\mathsf{M}} = K_Q^\dagger$, where $K_Q^\dagger$ is the posterior kernel of K with respect to Q. We illustrate these two concepts with an example:

Example 3.1 (Induced objects)**.** *Consider the concept space $\mathcal{C} = \mathbb{R}^d$ and the feature space $\mathcal{Z} = \mathbb{R}^p$.*

For a left-invertible matrix $\mathsf{G} \in \mathbb{R}^{p \times d}$, let $\mathsf{M} = (Q, K)$, where $Q = \mathcal{N}(\mathbf{0}, \mathsf{I}_d)$ and $K \colon \mathbf{c} \mapsto \delta_{\mathsf{G} \cdot \mathbf{c}}$. The induced feature distribution of M is $P^{\mathsf{M}} = \mathcal{N}(\mathbf{0}, \mathsf{G} \cdot \mathsf{G}^\top)$, and the induced concept extractor is $H^{\mathsf{M}} \colon \mathbf{z} \mapsto \delta_{\mathsf{H} \cdot \mathbf{z}}$, where $\mathsf{H} \in \mathbb{R}^{d \times p}$ is the left inverse of G.

We let $\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$ denote the set of latent concept generative models from $\mathcal{C}$ to $\mathcal{Z}$, i.e.,

$$\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z}) := \mathcal{P}(\mathcal{C}) \times \mathcal{K}(\mathcal{C} \to \mathcal{Z}).$$

Throughout this section, let $\mathcal{C}$ and $\mathcal{C}'$ be concept spaces, let $\mathsf{M} = (Q, K)$ be a LC-GM from $\mathcal{C}$ to $\mathcal{Z}$, and let $\mathsf{M}' = (Q', K')$ be a LC-GM from $\mathcal{C}'$ to $\mathcal{Z}$.

3.2 Feature Equivalence

When performing concept extraction in the population setting, we observe a distribution $P^* \in \mathcal{P}(\mathcal{Z})$ over features, and aim to find some LC-GM M such that $P^{\mathsf{M}} = P^*$. Intuitively, such a model M represents a possible explanation for the observed distribution P, and a possible concept extractor H^{M}.

However, multiple LC-GMs can induce the same feature distribution, but induce different concept extractors. In particular, we say M and M' are ***feature equivalent*** if $P^{\mathsf{M}} = P^{\mathsf{M}'}$. As the following simple example shows, feature equivalence does not imply $H^{\mathsf{M}} = H^{\mathsf{M}'}$, even up to relabeling.

Example 3.2 (Feature equivalence)**.** *Consider the spaces $\mathcal{C} = \{a, b\}$, $\mathcal{C}' = \{c, d\}$, and $\mathcal{Z} = \{u, v\}$. Let $\mathsf{M} = (Q, K)$, and let $\mathsf{M}' = (Q', K')$, for*

$$K = \begin{matrix} u \\ v \end{matrix} \overset{\begin{matrix} a & b \end{matrix}}{\begin{pmatrix} 2/3 & 1/3 \\ 0 & 1 \end{pmatrix}}, \qquad Q = \begin{matrix} a \\ b \end{matrix} \begin{pmatrix} 3/4 \\ 1/4 \end{pmatrix},$$

$$K' = \begin{matrix} u \\ v \end{matrix} \overset{\begin{matrix} c & d \end{matrix}}{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}, \qquad Q' = \begin{matrix} c \\ d \end{matrix} \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

Then $P^{\mathsf{M}} = P^{\mathsf{M}'} = \frac{1}{2}(\delta_u + \delta_v)$. However, $H^{\mathsf{M}} \neq H^{\mathsf{M}'}$:

$$H^{\mathsf{M}} = \begin{matrix} a \\ b \end{matrix} \overset{\begin{matrix} u & v \end{matrix}}{\begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}} \quad H^{\mathsf{M}'} = \begin{matrix} c \\ d \end{matrix} \overset{\begin{matrix} u & v \end{matrix}}{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}.$$

[2] We note that LC-GMs are deep latent variable models (*c.f.* Diederik and Max (2019, Section 1.7)) where *both* the "prior" Q and the "stochastic decoder" K are unknown.

3.3 Blackwell Coarsening and Equivalence

In Sec. 4, we will consider restricted model classes $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$, and our main interest will be developing *necessary* conditions for feature equivalence. To motivate these restrictions, we first examine an important *sufficient* condition for feature equivalence.

Inspired by connections to Blackwell's theory of experiments (Blackwell, 1953), we introduce the ***Blackwell coarsening relation*** on LC-GMs. Returning to the general setting with potentially distinct concept spaces $\mathcal{C}$ and $\mathcal{C}'$, consider a ***transition kernel*** $T \in \mathcal{K}(\mathcal{C} \to \mathcal{C}')$. We say M is ***Blackwell coarser than*** M' ***under*** T if the following conditions hold:

1. **Measure refinement**: $Q' = TQ$.
2. **Kernel coarsening**: $K \sim_Q K'T$.

For convenience, we will often simply state that M is Blackwell coarser than M', denoted $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$, if there exists some $T \in \mathcal{K}(\mathcal{C} \to \mathcal{C}')$ such that these relations hold.[3] It is straightforward to show that Blackwell coarsening is sufficient for feature equivalence:

Proposition 3.1. *If $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$, then M and M' are feature equivalent. In particular, if $Q' = TQ$ and $K \sim_Q K'T$, then $KQ = K'Q'$.*

The proof follows directly from substitution and functoriality; see App. C.1. Intuitively, the Blackwell coarsening relation captures a tradeoff between two explanations for randomness in a feature distribution $P^* = P^{\mathsf{M}} = P^{\mathsf{M}'}$. When $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$, K is "more stochastic" than K' (a *garbling* of K in Blackwell's terms), whereas Q' has higher entropy than Q. Put differently, M emphasizes a kind of measurement uncertainty, while M' emphasizes natural variability. This tradeoff is sharply illustrated by our example:

Example 3.3 (Blackwell coarsening)**.** *Let M and M' be the models from Ex. 3.2. It is straightforward to check that M is Blackwell coarser than M' under*

$$T = \begin{matrix} \\ c \\ d \end{matrix} \begin{matrix} \begin{matrix} a & b \end{matrix} \\ \begin{pmatrix} 2/3 & 0 \\ 1/3 & 1 \end{pmatrix} \end{matrix}.$$

As this discussion and the example should make clear, Blackwell coarsening is not a symmetric relation, e.g., there is no Markov kernel $T' \in \mathcal{K}(\mathcal{C}' \to \mathcal{C})$ such that $K' = KT'$ for Ex. 3.2. In the special case where the relation does hold in both directions (i.e., $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$, and $\mathsf{M}' \preceq_{\text{Bl}} \mathsf{M}$), we say that M and M' are ***Blackwell equivalent***, which we denote by $\mathsf{M} =_{\text{Bl}} \mathsf{M}'$.

[3] Using $\preceq_{\text{Bl}}$ is justified since Blackwell coarsening is transitive: if $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$ under T and $\mathsf{M}' \preceq_{\text{Bl}} \mathsf{M}''$ under T', then $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}''$ under $T'T$, see App. C.1.

Converse At first pass, it may seem that Blackwell coarsening could serve as a necessary condition for feature equivalence, at least under conditions such as injectivity of $K_\sharp$ and $K'_\sharp$. However, a simple modification of Ex. 3.2, given in App. B, shows that this is not the case: two LC-GMs M and M' may be feature equivalent with neither $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$ nor $\mathsf{M}' \preceq_{\text{Bl}} \mathsf{M}$.

Despite this null result, we can already obtain a partial converse to Prop. 3.1, which will be a key ingredient for our main results. In particular, if the kernel coarsening condition already holds for M and M' which are feature equivalent, then the measure refinement condition automatically holds:

Lemma 3.1. *Assume that K' is measure-separating, and that $K \sim_Q K'T$ for some $T \in \mathcal{K}(\mathcal{C} \to \mathcal{C}')$.*

If $P^{\mathsf{M}} = P^{\mathsf{M}'}$, then $Q' = TQ$, i.e., $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$.

The proof follows directly from injectivity of K' and functoriality; see App. C.1.

4 MAIN RESULTS

We are finally ready to move on to our main results, which give *necessary* conditions for the feature equivalence of two models $\mathsf{M} = (Q, K)$ and $\mathsf{M}' = (Q', K')$ in the same model class $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$. As is typical in most settings, we consider model class of the form $\mathcal{M} = \mathcal{Q} \times \mathcal{K}$, where $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{C})$ and $\mathcal{K} \subseteq \mathcal{K}(\mathcal{C} \to \mathcal{Z})$, so that any constraints on the concept distributions and the mixing kernels are logically independent.

In Sec. 4.1, we define a natural condition on the kernel class $\mathcal{K}$ under which Blackwell equivalence is indeed a necessary condition for feature equivalence. We then use the result in Sec. 4.2 to establish an easy-to-use *intersection theorem* for identifiability.

4.1 Blackwell Reducibility

As seen in Sec. 3.3, a main obstacle to guaranteeing Blackwell equivalence of M and M' is the potential incomparability of the mixing kernels K and K'. To avoid this issue, we define a novel condition on $\mathcal{K}$.

Definition 4.1. *Let $\mathcal{I}$ be a Polish Borel space and $\mathcal{R} := \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{I}) \circ \mathcal{P}(\mathcal{C})$. Consider a* **base kernel** *$B \in \mathcal{K}_{\text{sep}}(\mathcal{I} \to \mathcal{Z})$. We say that $\mathcal{K} \subseteq \mathcal{K}(\mathcal{C} \to \mathcal{Z}; \mathcal{R})$ is* **Blackwell reducible (BR) through** *B if*

$$\mathcal{K} \subseteq \{B\} \circ \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{I}).$$

We say $\mathcal{K}$ is **continuous Blackwell reducible (CBR) through** *B if*

$$\mathcal{K} \subseteq \{B\} \circ \mathcal{F}_{\text{c-inj}}(\mathcal{C} \to \mathcal{I}),$$

Put differently, Blackwell reducibility of $\mathcal{K}$ means that, for all $K \in \mathcal{K}$, we have $K = BE_\phi$ for some injective function $\phi\colon \mathcal{C} \to \mathcal{I}$. By this definition, if $\mathcal{K}$ is Blackwell reducible, then any subset $\mathcal{K}' \subseteq \mathcal{K}$ is also BR (and similarly for CBR). As we show in App. B, the BR condition rules out previous problematic cases: $\mathcal{K} = \{K, K'\}$ is not BR for the mixing kernels K, K' in Ex. 3.2. Assuming that $\mathcal{K}$ is BR, we obtain a (specialized) converse of Prop. 3.1, where Blackwell equivalence and feature equivalence coincide.

Theorem 4.1. *Let $\mathcal{M} = \mathcal{P}(\mathcal{C}) \times \mathcal{K}$ for some BR class of kernels $\mathcal{K}$, and let $\mathsf{M}, \mathsf{M}' \in \mathcal{M}$ with $\mathsf{M} = (Q, K)$ and $\mathsf{M}' = (Q', K')$. If $P^{\mathsf{M}} = P^{\mathsf{M}'}$, then $\mathsf{M} =_{\text{Bl}} \mathsf{M}'$.*

If we also assume $\mathcal{K}$ is a CBR class of kernels and $\operatorname{supp}(Q) = \operatorname{supp}(Q') = \mathcal{C}$, then M is a Blackwell coarsening of M' under E_τ for some $\tau \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{C})$.

See App. C.2 for the proof, which leverages the shared base kernel B to reduce the problem closer to the case of deterministic mixing. Indeed, Thm. 4.1 extends Lemma 2.1 of Xi and Bloem-Reddy (2023), which essentially shows that feature equivalence implies Blackwell equivalence for $\mathcal{K} = \{E_g : g \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{Z})\}$, a prototypical example of a BR class:

Example 4.1 (BR for injective measurable functions)**.** *The set $\mathcal{K} = \{E_g : g \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{Z})\}$ is BR. Specifically, letting $\mathcal{I} = \mathcal{Z}$ and $B\colon \mathbf{z} \mapsto \delta_{\mathbf{z}}$, any $E_g \in \mathcal{K}$ can be written as $E_g = BE_g$. Similarly, the set $\mathcal{K} = \{E_g\colon g \in \mathcal{F}_{\text{c-inj}}(\mathcal{C} \to \mathcal{Z}\}$ is CBR.*

However, Blackwell reducibility is a much broader notion. For example, in latent factor models, a common assumption is that of *additive noise* after mixing, i.e., $\mathbf{Z} = g(\mathbf{C}) + \boldsymbol{\varepsilon}$ for some independent noise $\boldsymbol{\varepsilon} \sim N_{\boldsymbol{\varepsilon}}$. In measure-theoretic terms, this assumption imposes $K(\cdot \mid \mathbf{c}) = N_{\boldsymbol{\varepsilon}} * \delta_{g(\mathbf{c})}$, where $*$ denotes convolution. When $N_{\boldsymbol{\varepsilon}}$ has a well-behaved form, e.g. multivariate Gaussian, this class of mixing kernels is BR.

Example 4.2 (BR for additive Gaussian noise)**.** *Take $\mathcal{Z} = \mathbb{R}^p$ and $\Sigma \in \mathcal{S}^p_{\succeq 0}$, where $\mathcal{S}^p_{\succeq 0}$ denotes the positive semidefinite cone. Let $\mathcal{K}$ denote the set of mixing kernels K such that $K(\cdot \mid \mathbf{c}) = \mathcal{N}(\mathbf{0}, \Sigma) * \delta_{g(\mathbf{c})}$ for some $g \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{Z})$. Then $\mathcal{K}$ is BR.*

Specifically, let $\mathcal{I} = \mathcal{Z}$ and $B\colon \mathbf{z} \mapsto \mathcal{N}(0, \Sigma) * \delta_{\mathbf{z}}$. Then, any $K \in \mathcal{K}$ can be written as $K = BE_g$.

In App. C.2.1, we show that B is indeed measure-separating. While both of these examples were closely based on injective functions from $\mathcal{C}$ to $\mathcal{Z}$, our final example demonstrates that this need not be the case. In particular, Blackwell reducibility can also capture parametric mixture models, as long as each mixture component is associated with a distinct distribution.

Example 4.3 (BR of Gaussian mixtures)**.** *For $\mathcal{C} = [d]$ and $\mathcal{Z} = \mathbb{R}^p$, let $\mathcal{K}$ denote the set of mixing kernels K such that $K(\cdot \mid c) = \mathcal{N}(\mu_c, \Sigma_c)$, where $\mu_c \neq \mu_{c'}$ for $c \neq c'$. Then $\mathcal{K}$ is BR.*

In particular, letting $\mathcal{I} = \mathbb{R}^p \times \mathcal{S}^p_{\succeq 0}$ and $B\colon (\mu, \Sigma) \mapsto \mathcal{N}(\mu, \Sigma)$, any mixing kernel $K \in \mathcal{K}$ can be written as $K = BE_\phi$, where $\phi\colon c \mapsto (\mu_c, \Sigma_c)$.

As seen in App. C.2.2, the main result of Yakowitz and Spragins (1968) can be viewed as a proof that B is measure-separating on the appropriate domain.

4.2 Intersection Theorem

Thm. 4.1 is a powerful tool for characterizing feature equivalence, but is not directly stated in the typical language of identifiability theory. We now bridge this gap by giving a general definition of identifiability and giving a more direct, easy-to-use version of Thm. 4.1 based on the intersection of two *valid transition sets*, extending Theorem 2.2 of Xi and Bloem-Reddy, 2023.

Identifiability Let $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$ be a model class, and let $\sim$ be an equivalence relation on $\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$. We say that $\mathsf{M} \in \mathcal{M}$ ***is identifiable up to $\sim$ in $\mathcal{M}$***, denoted $\mathsf{ID}(\mathcal{M}, \sim, \mathsf{M})$, if $P^{\mathsf{M}} = P^{\mathsf{M}'}$ implies $\mathsf{M} \sim \mathsf{M}'$. As this definition makes clear, identifiability is in general a *ternary* relation, depending crucially on the choice model class $\mathcal{M}$, the choice of equivalence relation $\sim$, and the specific model $\mathsf{M} \in \mathcal{M}$. However, in many practical scenarios, the same identifiability result will hold for all models $\mathsf{M} \in \mathcal{M}$, i.e., we will have $\mathsf{ID}(\mathcal{M}, \sim, \mathsf{M})$ for all $\mathsf{M} \in \mathcal{M}$. In this case, we simply say that $\mathcal{M}$ ***is identifiable up to*** $\sim$, denoted $\mathsf{ID}(\mathcal{M}, \sim)$.

Notably, Blackwell equivalence is precisely an equivalence relation on $\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$, i.e., under this terminology, Thm. 4.1 shows that M is identifiable up to Blackwell equivalence in $\mathcal{M}$ whenever $\mathcal{M} = \mathcal{P}(\mathcal{C}) \times \mathcal{K}$ for some Blackwell reducible class $\mathcal{K}$. Our next goal is to characterize Blackwell equivalence more explicitly.

Valid transition sets The definition of Blackwell coarsening nearly suggests that Blackwell equivalence can be split into two separate relations: one over the concept distributions, and one over the mixing kernels. However, the kernel coarsening condition depends on Q, which is somewhat unsatisfactory for the practical development of identifiability results.

To avoid this issue, we introduce a final restriction on our model classes. Again considering a model class of the form $\mathcal{M} = \mathcal{Q} \times \mathcal{K}$, this new restriction now applies to the set $\mathcal{Q}$ of concept distributions:

Definition 4.2. $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{C})$ *is a* **full support measure class** *if there is a* **base measure** μ *on* $\mathcal{C}$ *with* $\text{supp}(\mu) = \mathcal{C}$ *such that for all* $Q \in \mathcal{Q}$, $Q \sim \mu$.

Under this condition, $K \sim_Q K'$ if and only if $K \sim_\mu K'$; thus, we can discuss kernel coarsening without reference to a specific $Q \in \mathcal{Q}$. To capture the restrictions implied by the measure refinement condition, we define the set of ***valid measure transitions*** of $\mathcal{Q}$ at Q, denoted $\mathcal{T}_Q(\mathcal{Q})$, as

$$\mathcal{T}_Q(\mathcal{Q}) := \{\tau \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{C}) : E_\tau Q \in \mathcal{Q}\}.$$

Hence, $\mathcal{T}_Q(\mathcal{Q})$ captures the set of transition functions τ which stay inside the set $\mathcal{Q}$ when applied to Q, i.e., for which $Q' = E_\tau Q$ is still a valid concept distribution according to the model class. Similarly, to capture the restrictions implied by the kernel coarsening condition, we define the set of ***valid kernel transitions*** of $\mathcal{K}$ at K, denoted $\mathcal{T}_K(\mathcal{K})$, as

$$\mathcal{T}_K(\mathcal{K}) := \{\tau \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{C}) : \exists\, K' \in \mathcal{K},\ K \sim_\mu K' E_\tau\}.$$

Here, $\mathcal{T}_K(\mathcal{K})$ denotes the set of transition functions which produce K upon coarsening some allowable mixing kernel K'. For ease of notation, the dependence of $\mathcal{T}_K(\mathcal{K})$ on the base measure μ is left implicit.

Identifiability of Models Now that we have defined valid transition sets and introduced the measure class assumption on $\mathcal{Q}$, we are ready to adapt Thm. 4.1 into a more usable form. To relate the sets $\mathcal{T}_Q(\mathcal{Q})$ and $\mathcal{T}_K(\mathcal{K})$ to equivalence relations, we consider one of the most common forms of equivalence relations: those induced by a group of functions. In particular, a nonempty set G of bijective measurable functions from $\mathcal{C}$ to $\mathcal{C}$ is called a ***transition group*** if, for all $\tau, \tau' \in G$, we have $\tau^{-1} \in G$ and $\tau \circ \tau' \in G$.

A group of functions G defines a natural equivalence relation $\sim_G$ on $\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$, where $\mathsf{M} \sim_G \mathsf{M}'$ if and only if M is Blackwell coarser than M' under E_τ for some $\tau \in G$. As we show in App. C.3, the group structure of G ensures that $\sim_G$ is indeed an equivalence relation. We can directly relate valid transition sets to identifiability for such relations:

Theorem 4.2. *Let* $\mathcal{M} = \mathcal{Q} \times \mathcal{K}$, *where* $\mathcal{Q}$ *is a full support measure class and* $\mathcal{K}$ *is CBR. Let* $\mathsf{M} = (Q, K) \in \mathcal{M}$, *and let* G *be a transition group.*

If $\mathcal{T}_Q(\mathcal{Q}) \cap \mathcal{T}_K(\mathcal{K}) \subseteq G$, *then* M *is identifiable up to* $\sim_G$ *in* $\mathcal{M}$, *i.e.,* $\mathsf{ID}(\mathcal{M}, \sim_G, \mathsf{M})$.

Proof. Let $\mathsf{M} = (Q, K)$ and $\mathsf{M}' = (Q', K')$ be feature equivalent models, with $\mathsf{M}, \mathsf{M}' \in \mathcal{M}$. By the assumption that $\mathcal{Q}$ is a full support measure class, $\text{supp}(Q) = \text{supp}(Q') = \mathcal{C}$. Thus, by the premise that $\mathcal{K}$ is CBR, Thm. 4.1 gives $Q' = E_\tau Q$ and $K \sim_\mu K' E_\tau$ for some $\tau \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{C})$. Finally, $Q' \in \mathcal{Q}$ implies $\tau \in \mathcal{T}_Q(\mathcal{Q})$, and $K' \in \mathcal{K}$ implies $\tau \in \mathcal{T}_K(\mathcal{K})$. Thus, by the premise of the theorem, we have $\tau \in G$, so M is identifiable up to $\sim_G$ in $\mathcal{M}$, as desired. □

Identifiability of Model Classes Alternatively, when we are interested in identifiability of an entire model class $\mathcal{M}$, rather than identifiability of a specific model $\mathsf{M} = (Q, K)$, we can define transition sets which do not depend on Q and K. Specifically, we let $\mathcal{T}(\mathcal{Q}) = \bigcup_{Q \in \mathcal{Q}} \mathcal{T}_Q(\mathcal{Q})$ and $\mathcal{T}(\mathcal{K}) = \bigcup_{K \in \mathcal{K}} \mathcal{T}_K(\mathcal{K})$, i.e., $\tau \in \mathcal{T}(\mathcal{Q})$ if it is a valid measure transition for some distribution $Q \in \mathcal{Q}$, and similarly for $\tau \in \mathcal{T}(\mathcal{K})$.

Corollary 4.1. *Let* $\mathcal{M} = \mathcal{Q} \times \mathcal{K}$, *where* $\mathcal{Q}$ *is a full support measure class and* $\mathcal{K}$ *is CBR, and let* G *be a transition group. If* $\mathcal{T}(\mathcal{Q}) \cap \mathcal{T}(\mathcal{K}) \subseteq G$, *then* M *is identifiable up to* $\sim_G$, *i.e.,* $\mathsf{ID}(\mathcal{M}, \sim_G)$.

As an important case, $\mathcal{T}(\mathcal{K})$ can be described quite simply for deterministic mixing functions. In particular, let $\mathcal{K} = \{E_g : g \in \mathcal{G}\}$ for some $\mathcal{G} \subseteq \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{Z})$. Then, $\mathcal{T}(\mathcal{K})$ is precisely the *indeterminacy set* defined in Xi and Bloem-Reddy (2023) (see App. C.3):

$$\mathcal{T}(\mathcal{K}) = \{\tau : \tau \sim_\mu g^{-1} \circ g' \text{ for some } g, g' \in \mathcal{G}\}.$$

5 APPLICATIONS

In this section, we highlight the utility of Thm. 4.2 by recovering classic identifiability results for two model classes relevant to unsupervised concept extraction and representation learning. To describe these model classes, we first introduce a clean new notation that makes all modeling assumptions highly transparent.

Predicate notation To compactly express model classes of the form $\mathcal{M} = \mathcal{Q} \times \mathcal{K}$, we introduce ***predicate notation***. We encode constraints on the concept distributions $Q \in \mathcal{Q}$ via ***concept predicates***, which are Boolean-valued functions $\mathsf{P} \colon \mathcal{P}(\mathcal{C}) \to \{\text{True}, \text{False}\}$. For a set $\{\mathsf{P}_j\}_{j=1}^J$ of concept predicates, we let $\mathcal{P}(\mathcal{C}; \mathsf{P}_1, \ldots, \mathsf{P}_J)$ denote the set of concept distributions for which all J predicates are true:

$$\mathcal{P}(\mathcal{C}; \mathsf{P}_1, \ldots, \mathsf{P}_J) := \left\{ Q \in \mathcal{P}(\mathcal{C}) : \bigwedge_{j=1}^J \mathsf{P}_j(Q) \right\},$$

Similarly, a ***mixing predicate*** is a Boolean-valued function $\mathsf{P} \colon \mathcal{K}(\mathcal{C} \to \mathcal{Z}) \to \{\text{True}, \text{False}\}$, and for a set of mixing predicates $\{\mathsf{P}_j\}_{j=1}^J$, we let $\mathcal{K}(\mathcal{C} \to \mathcal{Z}; \mathsf{P}_1, \ldots, \mathsf{P}_J)$ denote the set of mixing kernels K for which all J predicates are true.

This notation plays very nicely with the notion of valid transition sets. Specifically, let $\mathcal{Q}_1 = \mathcal{P}(\mathcal{C}; \mathsf{P}_1)$, $\mathcal{Q}_2 = \mathcal{P}(\mathcal{C}; \mathsf{P}_2)$, and $\mathcal{Q} = \mathcal{P}(\mathcal{C}; \mathsf{P}_1, \mathsf{P}_2)$ for some concept predicates P_1 and P_2. Then a simple application of the definitions shows that $\mathcal{T}(\mathcal{Q}) = \mathcal{T}(\mathcal{Q}_1) \cap \mathcal{T}(\mathcal{Q}_2)$; an analogous result holds for mixing kernels.

Equivalence Relations We let $\text{Relabel}(d)$ denote the set of ***relabeling functions*** on $\mathbb{R}^d$, i.e., $\tau \in \text{Relabel}(d)$ if and only if $\tau \colon c \mapsto \mathsf{S} \cdot c$ for some permutation matrix S. Similarly, we let $\text{Scale}(d)$ denote the set of ***scaling functions*** on $\mathbb{R}^d$, i.e., $\tau \in \text{Scale}(d)$ if and only if $\tau \colon c \mapsto \mathsf{D} \cdot c$ for some diagonal matrix with nonzero diagonal entries.

For $\mathcal{C} = \mathbb{R}^d$, $\sim_{\text{lbl}}$ denotes the equivalence relation $\sim_G$ for $G = \text{Relabel}(d)$, i.e., $\mathsf{M} \sim_{\text{lbl}} \mathsf{M}'$ if the two models are equivalent up to relabeling the concepts. Similarly, $\sim_{\text{scale}}$ denotes the equivalence relation $\sim_G$ for $G = \{g \circ g', g \in \text{Relabel}(d), g' \in \text{Scale}(d)\}$.

5.1 SAEs and Dictionary Learning

Sparse autoencoders (SAEs) and their variants are some of the most widely-used methods for concept extraction from neural features. Typically, sparse autoencoders assume continuous concept and feature spaces $\mathcal{C}$ and $\mathcal{Z}$ and high-dimensional concept spaces, i.e., $\mathcal{C} = \mathbb{R}^d$ and $\mathcal{Z} = \mathbb{R}^p$ for $d > p$. As suggested by the name, SAEs aim for sparsity over the concept space, often by a soft constraint such as ℓ_1 regularization. Although these approaches are not typically framed from the perspective taken here, they are closely related to the classical field of *dictionary learning*, also known as *sparse coding* (Olshausen and Field, 1997; Aharon et al., 2006a; Arora et al., 2014). To illustrate our results, we use Thm. 4.2 to derive a classic result in dictionary learning.

To formalize a strict (rather than soft) sparsity constraint, we let Σ_s^d denote the set of s-sparse vectors in $\mathbb{R}^d$, i.e., $\Sigma_s^d := \{\mathbf{c} \in \mathbb{R}^d : \|\mathbf{c}\|_0 \le s\}$. We endow $\mathcal{C} = \Sigma_s^d$ with the ***canonical stratified measure*** λ_s^d defined in App. D.1, noting $\text{supp}(\lambda_s^d) = \Sigma_s^d$. To employ Thm. 4.2, we use the concept predicate $\mathsf{MeasCls}_\mu$, where $\mathsf{MeasCls}_\mu(Q)$ holds if and only if Q is in the measure class of μ, i.e., $Q \ll \mu$ and $\mu \ll Q$.

In most mainstream SAEs and in the traditional dictionary learning literature, one assumes a *deterministic* and *linear* relationship between concepts and features, a form of the well-known *linear representation hypothesis* (Park et al., 2024; Jiang et al., 2024). This assumption can be encoded by the mixing predicate Linear, where $\mathsf{Linear}(K)$ holds if and only if $K \colon \mathbf{c} \mapsto \delta_{\mathsf{G} \cdot \mathbf{c}}$ for some matrix $\mathsf{G} \in \mathbb{R}^{p \times d}$.

Since $d > p$, the matrix G is not necessarily injective on Σ_s^d. To ensure injectivity, it is essential to further restrict the matrices G. The classical approach to ensure injectivity is to assume a lower bound on the ***spark*** of G. In particular, we let $\text{spk}(\mathsf{G})$ denote the smallest $k \in \mathbb{N}$ such that some k columns of G are linearly dependent. Then, when used with the Linear, the mixing predicate $\mathsf{Spark}_\sigma(K)$ holds if and only if $K \colon \mathbf{c} \mapsto \delta_{\mathsf{G} \cdot \mathbf{c}}$ for some $\mathsf{G} \in \mathbb{R}^{p \times d}$ with $\text{spk}(\mathsf{G}) \ge \sigma$.

Taking these assumptions together, we can define the model class for *spark-constrained dictionary learning*:

$$\begin{aligned} \mathcal{M}_{\text{spk}}(d, p; s, \sigma) &:= \mathcal{Q}_{\text{spk}} \times \mathcal{K}_{\text{spk}}, \text{ for} \\ \mathcal{Q}_{\text{spk}} &:= \mathcal{P}(\Sigma_s^d; \mathsf{MeasCls}_{\lambda_s^d}), \text{ and} \\ \mathcal{K}_{\text{spk}} &:= \mathcal{K}(\Sigma_s^d \to \mathbb{R}^p; \mathsf{Linear}, \mathsf{Spark}_\sigma). \end{aligned} \tag{1}$$

To apply Cor. 4.1, we first note that $\mathcal{Q}_{\text{spk}}$ is a full support measure class. We also need to verify that $\mathcal{K}_{\text{spk}}$ is CBR. As noted in Sec. 4.1, the set of continuous injective measurable functions is CBR, and any subset of a CBR class is also CBR. Thus, to ensure the second condition, it suffices to ensure that $\mathcal{K}_{\text{spk}}$ contains only injective functions.

For general σ and s, this condition need not be true. However, it is well-known that taking $\sigma > 2s$ ensures injectivity.[4] The following result further characterizes $\mathcal{T}(\mathcal{K}_{\text{spk}})$ when $\sigma > 2s$.

Claim 5.1. *Let $K, K' \in \mathcal{K}_{\text{spk}}$ defined in Eq. (1), and assume $\sigma > 2s$. If $K \sim_{\lambda_s^d} K' E_\tau$ for some $\tau \in \mathcal{F}_{\text{inj}}(\Sigma_s^d \to \Sigma_s^d)$, then $\tau \in \text{Relabel}(d) \circ \text{Scale}(d)$.*

More concisely, $\mathcal{T}(\mathcal{K}_{\text{spk}}) = \text{Relabel}(d) \circ \text{Scale}(d)$.

See App. D.2 for the proof of this claim: intuitively, linearity of $\tau \in \mathcal{T}(\mathcal{K}_{\text{spk}})$ follows from linearity of K and K', while the further constraints are induced by a combinatorial argument over s-dimensional subspaces. Thus, by applying Claim 5.1 and Cor. 4.1, we obtain an adaptation of Theorem 3 in Aharon et al. (2006b) to the population setting:

Claim 5.2. *Let $\mathcal{M} = \mathcal{M}_{\text{spk}}(d, p; s, \sigma)$ for $\sigma > 2s$. Then $\mathcal{M}$ is identifiable up to $\sim_{\text{scale}}$.*

5.2 Independent Component Analysis

Although not commonly used for concept extraction, several unsupervised representation learning approaches fit into our latent concept generative modeling framework. These approaches are also designed with the goal of *disentangling* their inputs into high-level concepts (Higgins et al., 2017; Bengio et al.,

[4] For completeness, we give a proof in App. D.2.

2013; R. Chen et al., 2018; Alemi et al., 2017). Again, these approaches are not always framed from a statistical angle, and often lack identifiability guarantees (Locatello et al., 2019). However, many approaches are closely related to the classical statistics problem of *independent component analysis (ICA)*, which fits directly into our framework. As suggested by the name, the core assumption in ICA is that the concept variables are statistically independent. We encode this with the concept predicate Ind, where $\mathsf{Ind}(Q)$ holds if and only if Q is a product distribution.

As in Sec. 5.1, we focus on the common setting of continuous concepts, continuous features, and deterministic mixing. Hence, we let $\mathcal{C} = \mathbb{R}^d$ and $\mathcal{Z} = \mathbb{R}^p$, where $\mathcal{C}$ is endowed with the Lebesgue measure λ. In contrast to Sec. 5.1, ICA usually considers $d \leq p$, and does not assume sparsity on $\mathcal{C}$. For convenience, we also assume a unit variance constraint. When used with Ind, the concept predicate $\mathsf{UnitVar}(Q)$ holds if and only if $\mathsf{Cov}(Q) = \mathsf{I}_d$; this constraint removes scaling ambiguity and yields a simpler proof.

It is well-known that the independence constraint and deterministic mixing are not sufficient for any meaningful identifiability guarantee (Hyvärinen and Pajunen, 1999). In our parlance, with $\mathcal{Q}^d_{\text{ind}} := \mathcal{P}(\mathbb{R}^d; \mathsf{Ind}, \mathsf{UnitVar})$ the set $\mathcal{T}(\mathcal{Q}^d_{\text{ind}})$ is far too large, since several non-trivial functions preserve independence. Hence, identifiability results in ICA must rely on additional assumptions.

Here, we focus on the classical setting of *linear* ICA, which assumes a linear and injective mixing function. We encode this assumption by the mixing predicate $\mathsf{LinearInj}$, and let $\mathcal{K}_{\text{lin}} = \mathcal{K}(\mathbb{R}^d \to \mathbb{R}^p; \mathsf{LinearInj})$. This restriction reduces the space of possible transition kernels: since the composition of two linear maps is linear, we have $\mathcal{T}(\mathcal{K}_{\text{lin}}) = \text{GL}(d)$, where $\text{GL}(d)$ denotes the *general linear group* on $\mathbb{R}^d$.

However, linear mixing still does not provide enough constraint. In App. D.3, we recount the fact that $\mathcal{T}(\mathcal{K}_{\text{lin}}) \cap \mathcal{T}(\mathcal{Q}^d_{\text{ind}}) = O(d)$, where $O(d)$ denotes the *orthogonal group* on $\mathbb{R}^d$. This ambiguity arises from a symmetry of Gaussian distributions: if $Q = \mathcal{N}(\mathbf{0}, \mathsf{I}_d)$, then the pushforward distribution $Q' = E_\tau Q$ also equals $\mathcal{N}(\mathbf{0}, \mathsf{I}_d)$ for any $\tau \in O(d)$. To avoid this problem, the seminal work of Comon (1994) shows that this symmetry disappears when we enforce non-Gaussianity. To state their result in our framework, we define the concept predicate $\mathsf{NonGaussian}$, which holds for Q if and only if $Q \sim \lambda$ and at most one component of Q is Gaussian. Letting $\mathcal{Q}^d_{\text{ind-ng}} := \mathcal{P}(\mathbb{R}^d; \mathsf{Ind}, \mathsf{UnitVar}, \mathsf{NonGaussian})$, the next result is a minor adaptation of Theorem 11 in Comon (1994):

Claim 5.3. $\mathcal{T}(\mathcal{Q}^d_{\text{ind-ng}}) \cap O(d) = \text{Relabel}(d)$.

See App. D.3 for the proof. Gathering these assumptions, we define the model class for linear ICA as

$$\mathcal{M}_{\text{lin-ica}}(d, p) := \mathcal{Q}^d_{\text{ind-ng}} \times \mathcal{K}(\mathbb{R}^d \to \mathbb{R}^p; \mathsf{LinearInj}).$$

By construction, $\mathcal{Q}^d_{\text{ind-ng}}$ is a full support measure class, and $\mathcal{K} = \mathcal{K}(\mathcal{C} \to \mathcal{Z}; \mathsf{LinearInj})$ is CBR, so we can apply Cor. 4.1. Simple set manipulation shows that $\mathcal{T}(\mathcal{Q}^d_{\text{ind-ng}}) \cap \mathcal{T}(\mathcal{K}) = \text{Relabel}(d)$, which gives our version of Corollary 13 in Comon (1994):

Claim 5.4. *Let $\mathcal{M} = \mathcal{M}_{\text{lin-ica}}(d, p)$. Then $\mathcal{M}$ is identifiable up to $\sim_{\text{lbl}}$.*

6 METHODOLOGICAL IMPLICATIONS

By design, a latent concept generative model (LC-GM) is a type of generative model over the feature space $\mathcal{Z}$. Thus, to learn an LC-GM, one may apply many of the same tools as for learning other kinds of generative models. Here, we briefly discuss one natural workflow for algorithm development suggested by our framework, and how this workflow relates to a common model class for sparse autoencoders:

$$\mathcal{M}_{\text{sae}}(d, p; s) := \mathcal{P}(\Sigma^d_s) \times \mathcal{K}(\Sigma^d_s \to \mathbb{R}^p; \mathsf{Linear}),$$

a relaxation of the model class in Eq. (1).[5] As shorthand, we let $\mathcal{M}_{\text{sae}} = \mathcal{M}_{\text{sae}}(d, p; s)$. In this section, we are slightly more informal, e.g. we assume all distributions have strictly positive densities.

6.1 Architecture Design

Our framework suggests a *decomposed* parameterization of the model class $\mathcal{M}$, similar to deep latent variable models (Diederik and Max, 2019).

In particular, let $\Theta = \Omega \times \Psi$ for parameter spaces Ω and Ψ, where each $\omega \in \Omega$ yields a concept distribution Q_ω (i.e., a learnable latent "prior"), and each $\psi \in \Psi$ yields a mixing kernel K_ψ (i.e., a stochastic decoder). Then, each $\theta = (\omega, \psi)$ yields a latent concept generative model $\mathsf{M}_\theta := (Q_\omega, K_\psi)$, with the joint distribution $J_\theta(c, z) := K_\psi(z \mid c) \cdot Q_\omega(c)$ and the induced feature distribution $P_\theta := (K_\psi)_\sharp(Q_\omega)$.

One may use architectural design to exactly enforce the constraints, i.e., $Q_\omega \in \mathcal{Q}$ and $K_\psi \in \mathcal{K}$, or may softly enforce these constraints through regularization. For the $\mathcal{M}_{\text{sae}}$ example, we require Q_ω to output

[5] Note that $\mathcal{M}_{\text{sae}}(d, p; s)$ does not satisfy the same identifiability guarantees as $\mathcal{M}_{\text{spk}}(d, p; s, \sigma)$. However, one may perform post-hoc identifiability checks; see App. E.1.

s-sparse vectors, and we require K_ψ to be linear. The linearity constraint can be easily enforced by taking $\Psi = \mathbb{R}^{p \times d}$ and letting $K_\psi : c \mapsto \mathsf{G}_\psi \cdot c$ be a linear layer. Meanwhile, the sparsity constraint can be encoded into the architecture via a discrete operation, as in TopK SAEs (Makhzani and Frey, 2013; Gao et al., 2024), or can be softly enforced by ℓ_1-regularization, as in traditional SAEs (Elhage et al., 2022).

6.2 Training

After designing the architecture, one needs to train $\theta = (\omega, \psi)$ on a consistent loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, i.e., a loss function such that the population loss

$$\mathcal{L}(\theta, P^*) := \mathbb{E}_{\mathbf{Z} \sim P^*}[\ell(\theta, \mathbf{Z})]$$

is minimized if and only if $P_\theta = P^*$, assuming the model class is well-specified.[6] In this approach, the choice of ℓ is constrained by the nature of Q_ω and K_ψ. For example, if Q_ω provides evaluation access, and $K_\psi = E_{g_\psi}$ for some diffeomorphism g_ψ with a tractable inverse $h_\psi := g_\psi^{-1}$, then the negative log likelihood $\ell(\theta, z) := -\log P_\theta(z)$ can be evaluated using the change of measure formula with $c = h_\psi(z)$:

$$-\log P_\theta(z) = -\log Q_\omega(c) + \log |\mathsf{J}_{g_\psi}(c)|,$$

where the Jacobian $\mathsf{J}_{g_\psi}(z)$ can be computed via autodifferentiation. However, if Q_ω and K_ψ are implicit models, then other choices of ℓ, e.g. an f-GAN divergence (Nowozin et al., 2016), may be necessary.

6.3 Concept Extraction as Inference

As described in Sec. 3, a latent concept generative model $\mathsf{M}_{\theta^*} = (Q_{\omega^*}, K_{\psi^*})$ induces a concept extractor

$$H_{\theta^*}(c \mid z) = \frac{K_{\psi^*}(z \mid c) \cdot Q_{\omega^*}(c)}{P_{\theta^*}(z)}.$$

Again, in the simplest setting where $K_\psi = E_{g_\psi}$ for some injective g_ψ, the concept extractor has a simple form: namely, $H_{\theta^*} = E_{h_{\psi^*}}$. For example, in $\mathcal{M}_{\text{spk}}$, the exact solution c such that $\mathsf{G}_{\psi^*} \cdot c = z$ can be computed via orthogonal matching pursuit under additional conditions on G_{ψ^*} (Tropp, 2004).

In the more general stochastic mixing setting, concept extraction requires *posterior inference*, and the form of Q_ω and K_ψ dictate the available inference procedures. For example, if Q_ω and K_ψ both provide evaluation access, then $H_{\theta^*}(c \mid z) \propto K_{\psi^*}(z \mid c) \cdot Q_{\omega^*}(c)$ can be treated as an energy-based model, allowing for the use of inference procedures such as Hamiltonian Monte Carlo (Neal, 2011). However, if Q_ω and K_ψ are implicit models, then one must turn to simulation-based inference (Cranmer et al., 2020).

If concept extraction (rather than generation) is the main goal, one may instead wish to *amortize* posterior inference by learning a parameterized concept extractor, as is done for traditional SAEs. In App. E.2, we further discuss this option, and App. E.3 discusses connections between our framework and the *projection encoders* framework of Hindupur et al. (2025).

7 DISCUSSION

In this paper, we have introduced a unified framework which formalizes the task of concept extraction from the lens of identifiability theory. Our two main theorems, Thm. 4.1 and Thm. 4.2, give powerful tools for establishing identifiability results, with our two applications in Sec. 5 showing how these theorems can be applied. In App. F, we cover a third application to a model class with stochastic mixing, which is a major extension beyond similar previous work (Xi and Bloem-Reddy, 2023). We expect these contributions to pave the way for a number of directions:

Bridging the theory-practice gap. By focusing on concept extraction as our motivating application, we have hopefully made identifiability theory accessible to a broader audience that greatly wants and needs these tools (Cui et al., 2025; Méloux et al., 2025; Song et al., 2025). Such theory can usefully inform method development and usage, but several issues remain, e.g. reasoning about identifiability in misspecified model classes (where $P^{\mathsf{M}} \neq P^*$ for any $\mathsf{M} \in \mathcal{M}$) and reasoning about identifiability under additional constraints on the concept extractor H^{M} (as brought up in App. E.3).

Extensions to weak supervision. In this paper, we have considered concept extraction from the perspective of unsupervised representation learning. However, several works on identifiability can be interpreted as providing *weak supervision* on the concepts, e.g. auxiliary variables in nonlinear ICA (Hyvarinen et al., 2019), interventions in causal representation learning (Zhang et al., 2023; Buchholz et al., 2023; Varici et al., 2025), or even downstream labels in traditional representation learning (Roeder et al., 2021; Reizinger et al., 2025; Zhai et al., 2025). Often, such supervision allows one to consider much more general latent concept generative models. Thus, it would be very useful to extend the framework here to incorporate weak supervision.

[6] We briefly discuss misspecification in Sec. 7.

Acknowledgements

We would like to thank Jerry Huang, Che-Ping Tsai, and Jason Hartford for insightful discussions on the overall framework, and Yizhou Lu and Soheun Yi for careful proofreading and discussions about methodology. We thank our anonymous reviewers for the encouragement to add Sec. 6 and the comparisons in App. E.3. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) via HR0011-25-3-0239, FA8750-23-2-1015, ONR via N00014-23-1-2368, and NSF via IIS-1909816.

References

Aharon, Michal et al. (2006a). "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation". In: *IEEE Transactions on Signal Processing*.

– (2006b). "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them". In: *Linear Algebra and Its Applications*.

Alemi, Alexander A et al. (2017). "Deep variational information bottleneck". In: *The International Conference on Learning Representations*.

Arora, Sanjeev et al. (2014). "New algorithms for learning incoherent and overcomplete dictionaries". In: *Conference on Learning Theory*.

Ashuach, Tomer et al. (2025). "CRISP: Persistent Concept Unlearning via Sparse Autoencoders". In: *arXiv preprint arXiv:2508.13650*.

Bengio, Yoshua et al. (2013). "Representation learning: A review and new perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Blackwell, David (1953). "Equivalent comparisons of experiments". In: *The Annals of Mathematical Statistics*.

Buchholz, Simon et al. (2023). "Learning linear causal representations from interventions under general nonlinear mixing". In: *Advances in Neural Information Processing Systems*.

Chen, Ricky et al. (2018). "Isolating sources of disentanglement in variational autoencoders". In: *Advances in Neural Information Processing Systems*.

Comon, Pierre (1994). "Independent component analysis, a new concept?" In: *Signal Processing*.

Cranmer, Kyle et al. (2020). "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences*.

Cui, Jingyi et al. (2025). "On the theoretical understanding of identifiable sparse autoencoders and beyond". In: *arXiv preprint arXiv:2506.15963*.

Cunningham, Hoagy et al. (2023). "Sparse autoencoders find highly interpretable features in language models". In: *arXiv preprint arXiv:2309.08600*.

Diederik, P Kingma and Welling Max (2019). "An introduction to variational autoencoders". In: *Foundations and Trends® in Machine Learning*.

Elhage, Nelson et al. (2022). "Toy models of superposition". In: *arXiv preprint arXiv:2209.10652*.

Gao, Leo et al. (2024). "Scaling and evaluating sparse autoencoders". In: *arXiv preprint arXiv:2406.04093*.

Higgins, Irina et al. (2017). "Beta-VAE: Learning basic visual concepts with a constrained variational framework". In: *The International Conference on Learning Representations*.

Hindupur, Sai Sumedh R et al. (2025). "Projecting assumptions: The duality between sparse autoencoders and concept geometry". In: *Advances in Neural Information Processing Systems*.

Hyvarinen, Aapo et al. (2019). "Nonlinear ICA using auxiliary variables and generalized contrastive learning". In: *The International Conference on Artificial Intelligence and Statistics*.

Hyvärinen, Aapo and Petteri Pajunen (1999). "Nonlinear independent component analysis: Existence and uniqueness results". In: *Neural Networks*.

Jiang, Yibo et al. (2024). "On the origins of linear representations in large language models". In: *The International Conference on Machine Learning*.

Leask, Patrick et al. (2025). "Sparse autoencoders do not find canonical units of analysis". In: *The International Conference on Learning Representations*.

Locatello, Francesco et al. (2019). "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *The International Conference on Machine Learning*.

Makhzani, Alireza and Brendan Frey (2013). "K-sparse autoencoders". In: *arXiv preprint arXiv:1312.5663*.

Méloux, Maxime et al. (2025). "Everything, everywhere, all at once: Is mechanistic interpretability identifiable?" In: *The International Conference on Learning Representations*.

Neal, Radford M (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*.

Nowozin, Sebastian et al. (2016). "f-GAN: Training generative neural samplers using variational divergence minimization". In: *Advances in Neural Information Processing Systems*.

Olshausen, Bruno A and David J Field (1997). "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research*.

Park, Kiho et al. (2024). "The linear representation hypothesis and the geometry of large language models". In: *The International Conference on Machine Learning*.

Paulo, Gonçalo and Nora Belrose (2025). "Sparse autoencoders trained on the same data learn different features". In: *arXiv preprint arXiv:2501.16615*.

Reizinger, Patrik et al. (2025). "Cross-entropy is all you need to invert the data generating process". In: *The International Conference on Learning Representations*.

Roeder, Geoffrey et al. (2021). "On linear identifiability of learned representations". In: *The International Conference on Machine Learning*.

Song, Xiangchen et al. (2025). "Position: Mechanistic interpretability should prioritize feature consistency in SAEs". In: *arXiv preprint arXiv:2505.20254*.

Sucholutsky, Ilia et al. (2023). "Getting aligned on representational alignment". In: *arXiv preprint arXiv:2310.13018*.

Tropp, Joel A (2004). "Greed is good: Algorithmic results for sparse approximation". In: *IEEE Transactions on Information theory*.

Turner, Alexander et al. (2023). "Activation addition: Steering language models without optimization". In: *CoRR*.

Varici, Burak et al. (2025). "Score-based causal representation learning: Linear and general transformations". In: *The Journal of Machine Learning Research*.

Xi, Quanhan and Benjamin Bloem-Reddy (2023). "Indeterminacy in generative models: Characterization and strong identifiability". In: *The International Conference on Artificial Intelligence and Statistics*.

Yakowitz, Sidney J and John D Spragins (1968). "On the identifiability of finite mixtures". In: *The Annals of Mathematical Statistics*.

Zhai, Runtian et al. (2025). "Contextures: Representations from contexts". In: *International Conference on Machine Learning*. PMLR.

Zhang, Jiaqi et al. (2023). "Identifiability guarantees for causal disentanglement from soft interventions". In: *Advances in Neural Information Processing Systems*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents of Appendix

A Unifying Framework for Unsupervised Concept Extraction

A BACKGROUND

In Sec. 2, we described our notation for distributions and Markov kernels. This section reintroduces that notation and provides more explicit definitions for the unfamiliar reader.

A.1 Borel Spaces and Distributions

Let $\mathcal{U}$ be a topological space. The ***Borel space*** associated with $\mathcal{U}$ is the pair $(\mathcal{U}, \mathcal{B})$, where $\mathcal{B}$ is the ***Borel σ-algebra*** of $\mathcal{U}$, i.e., the σ-algebra generated by the open sets of $\mathcal{U}$. We define a ***Polish Borel space*** as a Polish space (i.e., a separable completely metrizable topological space) endowed with its Borel σ-algebra. Ignoring its topological data, such a space is commonly called a ***standard Borel space***, see Kechris (1995, Definition 12.5). Here, we include *Polish* in the name to indicate that the topological structure is kept.

For any measurable space $\mathcal{V}$, we let $\mathcal{P}(\mathcal{V})$ denote the set of probability distributions on $\mathcal{V}$. We say that $Q \in \mathcal{P}(\mathcal{V})$ is ***absolutely continuous*** with respect to a measure μ on $\mathcal{V}$ if $\mu(A) = 0$ implies $Q(A) = 0$ for any $\mathcal{V}$-measurable set A.

A.2 Markov Kernels

A ***Markov kernel*** from $\mathcal{U}$ to $\mathcal{V}$ is a measurable function from $\mathcal{U}$ to $\mathcal{P}(\mathcal{V})$. We denote the set of Markov kernels from $\mathcal{U}$ to $\mathcal{V}$ as $\mathcal{K}(\mathcal{U} \to \mathcal{V})$, and given a Markov kernel $K \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$, we let $K(\cdot \mid u)$ denote its output for $u \in \mathcal{U}$. Given a measure μ on $\mathcal{U}$, we say that two kernels K and K' are ***equal μ-almost everywhere*** if $\mu(N) = 0$ for the set $N = \{u \in \mathcal{U} : K(\cdot \mid u) \neq K'(\cdot \mid u)\}$. We write $K \sim_\mu K'$ to denote that K and K' are equal μ-almost everywhere.

Pushforwards and Posterior Kernels For a Markov kernel $K \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$, we define an associated operator $K_\sharp \colon \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{V})$, which maps the distribution $Q \in \mathcal{P}(\mathcal{U})$ to the distribution $P = K_\sharp(Q)$ such that

$$P(\cdot) = \int_{\mathcal{U}} K(\cdot \mid u) \cdot Q(du). \tag{2}$$

Given a Markov kernel $K \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$ and a distribution $Q \in \mathcal{P}(\mathcal{U})$, a ***posterior kernel*** of K with respect to Q is a Markov kernel $L \in \mathcal{K}(\mathcal{V} \to \mathcal{U})$ such that, for all measurable sets $A \subseteq \mathcal{U}$ and $B \subseteq \mathcal{V}$,

$$\int_A K(B \mid du) \cdot Q(du) = \int_B L(A \mid dv) \cdot P(dv), \tag{3}$$

where $P = K_\sharp(Q)$. Assuming that $\mathcal{U}$ and $\mathcal{V}$ are standard Borel spaces, a posterior kernel always exists, and is unique P-almost everywhere (for reference, see Example 11.6 in Fritz (2020)). We let $K_Q^\dagger$ denote any such posterior kernel.

Composition Given two Markov kernels $K \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$ and $L \in \mathcal{K}(\mathcal{V} \to \mathcal{W})$, the ***composition*** of these Markov kernels, denoted LK, is a Markov kernel from $\mathcal{U}$ to $\mathcal{W}$, where for all $u \in \mathcal{U}$,

$$(LK)(\cdot \mid u) = \int_{\mathcal{V}} L(\cdot \mid v) \cdot K(dv \mid u)$$

A basic but important fact is that the pushforward and composition operators can be interchanged, i.e., $(LK)_\sharp = L_\sharp \circ K_\sharp$. This property is actually a special case of the associativity of Markov kernels, since $Q \in \mathcal{P}(\mathcal{V})$ can be viewed as a Markov kernel from the empty set to $\mathcal{V}$, in which case $K_\sharp(Q) = KQ$. For a general proof of associativity, see Proposition 3.2 of Panangaden (1999).

Composition Notation As defined in Sec. 2, for two sets of Markov kernels $\mathcal{K} \subseteq \mathcal{K}(\mathcal{U} \to \mathcal{V})$ and $\mathcal{L} \subseteq \mathcal{K}(\mathcal{V} \to \mathcal{W})$, we let

$$\mathcal{L} \circ \mathcal{K} := \{LK : L \in \mathcal{L}, K \in \mathcal{K}\} \subseteq \mathcal{K}(\mathcal{U} \to \mathcal{W}).$$

For a set of probability distributions $\mathcal{P} \subseteq \mathcal{P}(\mathcal{U})$, we let

$$\mathcal{K} \circ \mathcal{P} := \{KP : K \in \mathcal{K}, P \in \mathcal{P}\} \subseteq \mathcal{P}(\mathcal{V}).$$

For a set of functions $\mathcal{F} \subseteq \mathcal{F}(\mathcal{U} \to \mathcal{V})$, we let

$$\mathcal{L} \circ \mathcal{F} := \{LE_f : L \in \mathcal{L}, f \in \mathcal{F}\} \ \subseteq \mathcal{K}(\mathcal{U} \to \mathcal{W}).$$

and

$$\mathcal{F} \circ \mathcal{P} := \{E_f P : f \in \mathcal{F}, P \in \mathcal{P}\} \subseteq \mathcal{P}(\mathcal{V}).$$

A.3 Technical Results for Composition

Normal composition (on the left) respects almost-everywhere equality, i.e., if $K \sim_\mu K'$, then $LK \sim_\mu LK'$. In several proofs, we need to *pre-compose* two kernels which are equal almost everywhere with respect to some measure. Thus, we will show that pre-composition also respects a.e.-equality in the necessary sense. First, we prove two auxiliary claims:

Claim A.1. *Let $\mathcal{U}$ and $\mathcal{V}$ be standard Borel spaces. For any measurable functions $f, g \colon \mathcal{U} \to \mathcal{V}$, the equalizer set*

$$\text{Equalizer}(f, g) := \{u \in \mathcal{U} : f(u) = g(u)\}$$

is measurable.

Proof. Defining $h \colon \mathcal{U} \to \mathcal{V} \times \mathcal{V}$ with $h \colon u \mapsto (f(u), g(u))$, we have that $\text{Equalizer}(f, g) = h^{-1}[\Delta_\mathcal{V}]$, where $\Delta_\mathcal{V} = \{(v, v) : v \in \mathcal{V}\}$. As noted by Kallenberg, 2011, the diagonal $\Delta_\mathcal{V}$ is measurable in $\mathcal{V} \times \mathcal{V}$ when $\mathcal{V}$ is Borel, and h is measurable by construction. Thus, $h^{-1}[\Delta_\mathcal{V}]$ is a measurable set. □

Claim A.2. *Take $R \in \mathcal{P}(\mathcal{U})$ and $J \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$, and let $R' = JR \in \mathcal{P}(\mathcal{V})$. Let $\mathcal{V}_1 \subseteq \mathcal{V}$ be a measurable set with $R'(\mathcal{V}_1) = 1$. Define*

$$\mathcal{U}_1 = \{u \in \mathcal{U} : J(\mathcal{V}_1 \mid u) = 1\}.$$

Then $\mathcal{U}_1$ is measurable and $R(\mathcal{U}_1) = 1$.

Proof. Define the measurable function $f \colon \mathcal{U} \to \mathbb{R}$ with $f(u) \colon u \mapsto J(\mathcal{V}_1 \mid u)$, and let $g \colon \mathcal{U} \to \mathbb{R}$ with $g(u) = 1$. Then $\mathcal{U}_1$ is the equalizer set of f and g, so it is measurable by Claim A.1. Then, by definition of the pushforward,

$$R'(\mathcal{V}_1) = (JR)(\mathcal{V}_1) = \int_\mathcal{U} f(u) \cdot R(du) = 1.$$

By linearity of the Lebesgue integral, we have $\int_\mathcal{U} (1 - f(u)) \cdot R(du) = 0$. Since $1 - f(u) \geq 0$, we must have $1 - f(u) = 0$ almost surely on R. Thus, $f(u) = J(\mathcal{V}_1 \mid u) = 1$ almost surely on R; i.e., $R(\mathcal{U}_1) = 1$. □

Claim A.3 (Left composition respects a.e.-equality)**.** *Take $R \in \mathcal{P}(\mathcal{U})$, $J \in \mathcal{K}(\mathcal{U} \to \mathcal{V})$, and $L_1, L_2 \in \mathcal{K}(\mathcal{V} \to \mathcal{W})$.*

Let $R' = JR \in \mathcal{P}(\mathcal{V})$. If $L_1 \sim_{R'} L_2$, then $L_1 J \sim_R L_2 J$.

Proof. Define the sets

$$\begin{aligned}\mathcal{V}_{\text{eq}} &:= \{v \in \mathcal{V} : L_1(\cdot \mid v) = L_2(\cdot \mid v)\},\\ \mathcal{U}_{\text{eq}} &:= \{u \in \mathcal{U} : (L_1 J)(\cdot \mid u) = (L_2 J)(\cdot \mid u)\}, \text{ and}\\ \mathcal{U}_1 &:= \{u \in \mathcal{U} : J(\mathcal{V}_{\text{eq}} \mid u) = 1\}.\end{aligned}$$

We briefly check these sets are measurable:

- $\mathcal{V}_{\text{eq}}$ is the equalizer set of the measurable functions $f, g\colon \mathcal{V} \to \mathcal{P}(\mathcal{W})$ with $f\colon v \mapsto L_1(\cdot \mid v)$ and $g\colon v \mapsto L_1(\cdot \mid v)$, and is thus measurable by Claim A.1.
- $\mathcal{U}_{\text{eq}}$ is the equalizer set of the measurable functions $f, g\colon \mathcal{U} \to \mathcal{P}(\mathcal{W})$ with $f\colon u \mapsto (L_1 J)(\cdot \mid u)$ and $g\colon u \mapsto (L_2 J)(\cdot \mid u)$, and is thus measurable by Claim A.1.
- $\mathcal{U}_1$ is the equalizer set of the measurable functions $f, g\colon \mathcal{U} \to \mathbb{R}$ with $f\colon u \mapsto J(\mathcal{V}_{\text{eq}} \mid u)$ and $g\colon u \mapsto 1$, and is thus measurable by Claim A.1.

Notably, we use that $\mathcal{P}(\mathcal{W})$ is a Polish Borel space, which follows from the fact that $\mathcal{P}(\mathcal{W})$ is a Polish space (under the topology of weak convergence) for any Polish space $\mathcal{W}$, see e.g. Bogachev and Ruas (2007, Theorem 8.9.5).

By the assumption that $L_1 \sim_{R'} L_2$, we have $R'(\mathcal{V}_{\text{eq}}) = 1$. By Claim A.2, we have $R(\mathcal{U}_1) = 1$.

Finally, we aim to show that $\mathcal{U}_1 \subseteq \mathcal{U}_{\text{eq}}$. For $u \in \mathcal{U}_1$, we have

$$\begin{aligned}(L_1 J)(\cdot \mid u) &= \int_{\mathcal{V}} L_1(\cdot \mid v) \cdot J(dv \mid u) && \text{(By definition)}\\ &= \int_{\mathcal{V}_{\text{eq}}} L_1(\cdot \mid v) \cdot J(dv \mid u) && \text{(Since } J(\mathcal{V}_1 \mid u) = 1 \text{ for } u \in \mathcal{U}_1\text{)}\\ &= \int_{\mathcal{V}_{\text{eq}}} L_2(\cdot \mid v) \cdot J(dv \mid u) && \text{(Since } L_1(\cdot \mid v) = L_2(\cdot \mid v) \text{ for } v \in \mathcal{V}_{\text{eq}}\text{)}\\ &= \int_{\mathcal{V}} L_2(\cdot \mid v) \cdot J(dv \mid u) && \text{(Since } J(\mathcal{V}_1 \mid u) = 1 \text{ for } u \in \mathcal{U}_1\text{)}\\ &= (L_2 J)(\cdot \mid u) && \text{(By definition)}\end{aligned}$$

Thus $\mathcal{U}_1 \subseteq \mathcal{U}_{\text{eq}}$, hence $R(\mathcal{U}_{\text{eq}}) \geq R(\mathcal{U}_1) = 1$. Hence, by the definition of $\mathcal{U}_{\text{eq}}$, we have $L_1 J \sim_R L_2 J$. □

B ADDITIONAL EXAMPLES

B.1 Feature Equivalence Does Not Imply Blackwell Comparability

In Sec. 3.3, we mentioned that a minor modification of Ex. 3.2 shows that two models M and M' may be feature equivalent, but may not be comparable under the Blackwell relation, i.e., we may have neither $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$ nor $\mathsf{M}' \preceq_{\text{Bl}} \mathsf{M}$. We now furnish the promised example.

Example B.1 (Feature equivalence)**.** *Consider the spaces* $\mathcal{C} = \{a, b\}$, $\mathcal{C}' = \{c, d\}$, *and* $\mathcal{Z} = \{u, v\}$.

Let $\mathsf{M} = (Q, K)$ *for*

$$Q = \begin{matrix} a \\ b \end{matrix}\begin{pmatrix} 3/4 \\ 1/4 \end{pmatrix} \quad K = \begin{matrix} u \\ v \end{matrix}\begin{pmatrix} 2/3 & 0 \\ 1/3 & 1 \end{pmatrix}\begin{matrix} \\ \\ \end{matrix}\begin{matrix} a & b \end{matrix}.$$

Meanwhile, let $\mathsf{M}' = (Q', K')$ *for*

$$Q' = \begin{matrix} c \\ d \end{matrix}\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \quad K' = \begin{matrix} u \\ v \end{matrix}\begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}\begin{matrix} \\ \\ \end{matrix}\begin{matrix} c & d \end{matrix}.$$

Then $P^{\mathsf{M}} = P^{\mathsf{M}'} = \frac{1}{2}(\delta_u + \delta_v)$. *We do not have* $K \sim_Q K'T$ *for any Markov kernel* $T \in \mathcal{K}(\mathcal{C} \to \mathcal{C}')$. *In particular, since* Q *has full support, this condition becomes an equality, i.e.,* $K = K'T$. *This equality induces a well-posed system of linear equations:*

$$\begin{bmatrix} 2/3 & 0 \\ 1/3 & 1 \end{bmatrix} = \begin{bmatrix} T_{ac} & T_{bc} \\ T_{ad} & T_{bd} \end{bmatrix} \cdot \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix}. \tag{4}$$

The only solution to this system is

$$\begin{bmatrix} T_{ac} & T_{bc} \\ T_{ad} & T_{bd} \end{bmatrix} = \begin{bmatrix} 5/6 & -1/2 \\ 1/6 & 3/2 \end{bmatrix},$$

which is not a valid Markov kernel (since, e.g., $T_{bc} < 0$*).*

Similarly, we do not have $K' \sim_{Q'} KT'$ *for any Markov kernel* $T' \in \mathcal{K}(\mathcal{C}' \to \mathcal{C})$. *Setting up a similar system of linear equations, the only solution is*

$$\begin{bmatrix} T'_{ca} & T'_{da} \\ T'_{cb} & T'_{db} \end{bmatrix} = \begin{bmatrix} 9/8 & 3/8 \\ -1/8 & 5/8 \end{bmatrix},$$

which is again not a valid Markov kernel.

B.2 A Blackwell Irreducible Class of Kernels

In Sec. 4.1, we mentioned that the Blackwell reducibility condition rules out cases like Ex. 3.2. We now describe this statement in more detail.

To adapt Ex. 3.2 to the case where $\mathcal{C} = \mathcal{C}'$, we let $\mathcal{C} = \{a, b\}$ and $\mathcal{Z} = \{u, v\}$. We consider the kernel class $\mathcal{K} = \{K, K'\} \subseteq \mathcal{K}_{\text{sep}}(\mathcal{C} \to \mathcal{Z})$, where

$$K = \begin{matrix} \\ u \\ v \end{matrix} \begin{matrix} \begin{matrix} a & b \end{matrix} \\ \begin{pmatrix} 2/3 & 0 \\ 1/3 & 1 \end{pmatrix} \end{matrix} \quad \text{and} \quad K' = \begin{matrix} \\ u \\ v \end{matrix} \begin{matrix} \begin{matrix} a & b \end{matrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{matrix}. \tag{5}$$

Then, we can show the following.

Claim B.1. *The above kernel class* $\mathcal{K}$ *is not Blackwell reducible.*

Proof. Suppose there exists a Polish Borel space $\mathcal{I}$, a base kernel $B \in \mathcal{K}(\mathcal{I} \to \mathcal{Z})$, and functions $\phi, \phi' \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{Z})$ such that $K = BE_\phi$ and $K' = BE_{\phi'}$. We will show that the base kernel B cannot be measure-separating.

Let $R = E_\phi Q$ and $R' = E_{\phi'} Q'$ for

$$Q = \begin{matrix} a \\ b \end{matrix} \begin{pmatrix} 3/4 \\ 1/4 \end{pmatrix} \quad \text{and} \quad Q' = \begin{matrix} a \\ b \end{matrix} \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}. \tag{6}$$

Then we have $R = \frac{3}{4}\delta_{\phi(a)} + \frac{1}{4}\delta_{\phi(b)}$ and $R' = \frac{1}{2}\delta_{\phi'(a)} + \frac{1}{2}\delta_{\phi'(b)}$. Since ϕ and ϕ' are injective, we must have $R \neq R'$, since (for example), R has lower entropy than R'. However, we have $BR = KQ = \frac{1}{2}(\delta_u + \delta_v)$ and $BR' = K'Q' = \frac{1}{2}(\delta_u + \delta_v)$, i.e., $BR = BR'$ for $R \neq R'$, so B is not measure-separating. □

C DEFEFFED PROOFS (MAIN RESULTS)

C.1 Proofs from Sec. 3.3

As mentioned in Sec. 3.3, the Blackwell coarsening relation is transitive, justifying the use of the ordering symbol $\preceq_{\text{Bl}}$. We now state this formally and give a short proof.

Proposition C.1 (Transitivity of the Blackwell coarsening relation)**.** *Let* $\mathsf{M} = (Q, K)$, $\mathsf{M}' = (Q', K')$, *and* $\mathsf{M}'' = (Q'', K'')$, *where* $\mathsf{M} \in \mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$, $\mathsf{M}' \in \mathcal{M}_{\text{all}}(\mathcal{C}', \mathcal{Z})$, *and* $\mathsf{M}'' \in \mathcal{M}_{\text{all}}(\mathcal{C}'', \mathcal{Z})$.

If $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$ *and* $\mathsf{M}' \preceq_{\text{Bl}} \mathsf{M}''$, *then* $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}''$. *In particular, if* M *is Blackwell coarser than* M' *under* T, *and* M' *is Blackwell coarser than* M'' *under* T', *then* M *is Blackwell coarser than* M'' *under* $T'' = T'T$.

Proof. From the Blackwell coarsening relation, we have (1) $Q' = TQ$, (2) $K \sim_Q K'T$, (3) $Q'' = T'Q'$, and (4) $K' \sim_{Q'} K''T'$. We must check the measure refinement and kernel coarsening conditions.

Measure refinement: The proof follows from substitution and functoriality:

$$\begin{aligned} Q'' &= T'Q' && \text{(Fact (3))} \\ &= T'(TQ) && \text{(Fact (1))} \\ &= (T'T)(Q) && \text{(Functoriality)} \\ &= T''Q, && \text{(Since } T'' := T'T\text{)} \end{aligned}$$

as desired.

Kernel coarsening: By Claim A.3, we have that (4) $K' \sim_{Q'} K''T'$ implies that (5) $K'T \sim_Q (K''T')T$.[7] The proof follows from transitivity of $\sim_Q$ and functoriality:

$$\begin{aligned} K &\sim_Q K'T && \text{(Fact (1))} \\ &\sim_Q (K''T')T && \text{(Fact (5) and transitivity of } \sim_Q\text{)} \\ &= K''(T'T) && \text{(By functoriality)} \\ &= K''T'', && \text{(Since } T'' := T'T\text{)} \end{aligned}$$

as desired. □

Our first result in Sec. 3.3 shows that Blackwell coarsening is sufficient for feature equivalence. We restate this result here and provide the proof.

Proposition (Restatement of Prop. 3.1)**.** *If* $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$, *then* M *and* M' *are feature equivalent. In particular, if* $Q' = TQ$ *and* $K \sim_Q K'T$, *then* $KQ = K'Q'$.

Proof. Since $K \sim_Q K'T$, we have $KQ = (K'T)Q$. Since $Q' = TQ$, we have $K'Q' = K'(TQ)$.

Thus, $P^{\mathsf{M}} = P^{\mathsf{M}'}$ directly follows from functoriality, since $(K'T)Q = K'(TQ)$. □

Our second result in Sec. 3.3 gives a result in the reverse direction: under an injectivity assumption, feature equivalence and the kernel coarsening condition imply the measure refinement condition. Again, we restate this result and provide the proof.

Lemma (Restatement of Lemma 3.1)**.** *Assume that* K' *is measure-separating, and that* $K \sim_Q K'T$ *for some* $T \in \mathcal{K}(\mathcal{C} \to \mathcal{C}')$.

If $P^{\mathsf{M}} = P^{\mathsf{M}'}$, *then* $Q' = TQ$, *i.e.,* $\mathsf{M} \preceq_{\text{Bl}} \mathsf{M}'$.

Proof. By substitution and functoriality,

$$\begin{aligned} KQ &= (K'T)(Q) && \text{(Since } K \sim_Q K'T\text{)} \\ &= K'(TQ). && \text{(Functoriality)} \end{aligned}$$

By assumption, $KQ = K'Q'$, hence the above gives that $K'(TQ) = K'(Q')$. By the assumption that K' is measure-separating, $Q' = TQ$. □

[7] In particular, we use Claim A.3 with $R = Q$, $J = T$, $L_1 = K'$, $L_2 = K''T'$, and $R' = Q' = TQ$.

C.2 Proofs from Sec. 4.1

Our first result in Sec. 4.1 shows that feature equivalance and Blackwell equivalence coincide when we only consider mixing kernels from a Blackwell reducible class $\mathcal{K}$. We now prove this key result.

Theorem (Restatement of Thm. 4.1)**.** *Let $\mathcal{M} = \mathcal{P}(\mathcal{C}) \times \mathcal{K}$ for some BR class of kernels $\mathcal{K}$, and let $\mathsf{M}, \mathsf{M}' \in \mathcal{M}$ with $\mathsf{M} = (Q, K)$ and $\mathsf{M}' = (Q', K')$. If $P^{\mathsf{M}} = P^{\mathsf{M}'}$, then $\mathsf{M} =_{\text{Bl}} \mathsf{M}'$.*

If we also assume $\mathcal{K}$ is a CBR class of kernels and $\text{supp}(Q) = \text{supp}(Q') = \mathcal{C}$, then M is a Blackwell coarsening of M' under E_τ for some $\tau \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{C})$.

Proof. By Blackwell reducibility, $K = BE_\phi$ and $K' = BE_{\phi'}$ for some base kernel $B \in \mathcal{K}_{\text{sep}}(\mathcal{I} \to \mathcal{Z}; \mathcal{R})$ and injective functions ϕ and ϕ'. Thus, by functoriality, $P^{\mathsf{M}} = B(E_\phi Q)$ and $P^{\mathsf{M}'} = B(E_{\phi'} Q')$. By definition, $\mathcal{R} = \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{I}) \circ \mathcal{P}(\mathcal{C})$, so $E_\phi, E_{\phi'} \in \mathcal{R}$. Since B is measure-separating on $\mathcal{R}$, and $P^{\mathsf{M}} = P^{\mathsf{M}'}$, we have $E_\phi Q = E_{\phi'} Q' = R$ for some $R \in \mathcal{P}(\mathcal{I})$.

To address domain issues, we must define the restrictions of several objects:

- **Sets:** Let $\mathcal{I}_0 = \text{supp}(R)$, let $\mathcal{C}_0 = \phi^{-1}[\mathcal{I}_0]$, and let $\mathcal{C}_0' = (\phi')^{-1}[\mathcal{I}_0]$.
- **Functions:** Let $\phi_0 \colon \mathcal{C}_0 \to \mathcal{I}_0$ denote the restriction of ϕ to $\mathcal{C}_0$, and let $\phi_0' \colon \mathcal{C}_0' \to \mathcal{I}_0$ denote the restriction of ϕ' to $\mathcal{C}_0'$.
- **Kernels:** Let $K_0 \in \mathcal{K}(\mathcal{C}_0 \to \mathcal{Z})$ denote the restriction of K to $\mathcal{C}_0$, and let $K_0' \in \mathcal{K}(\mathcal{C}_0' \to \mathcal{Z})$ denote the restriction of K' to $\mathcal{C}_0'$. Note that $K_0 = BE_{\phi_0}$ and $K_0' = BE_{\phi_0'}$.

By construction, ϕ_0 and ϕ_0' have matching images, so $\tau_0 := (\phi_0')^{-1} \circ \phi_0$ is a well-defined function from $\mathcal{C}_0$ to $\mathcal{C}_0'$, and τ_0 is injective. Let τ be any measurable extension of τ_0 onto all of $\mathcal{C}$. We now show that M is Blackwell coarser than M' under E_τ.

Kernel Coarsening By definition of τ_0, we have $E_{\phi_0'} E_{\tau_0} = E_{\phi_0}$. Thus, by functoriality, $K_0' E_{\tau_0} = (BE_{\phi_0'}) E_{\tau_0} = B(E_{\phi_0'} E_{\tau_0}) = BE_{\phi_0} = K_0$, i.e., $K_0 = K_0' E_{\tau_0}$. Since τ only extends τ_0 on a Q-null set, this gives $K \sim_Q K' E_\tau$.

Measure refinement By Blackwell reducibility, K' is measure-separating, and we have just shown that $K \sim_Q K' E_\tau$. Thus, by Lemma 3.1, we have $Q' = E_\tau Q$.

Repeating the proof by swapping M and M' shows that M' is Blackwell coarser than M under any extension τ' of $\tau_0' := \phi_0^{-1} \circ \phi_0'$. Hence, M and M' are Blackwell equivalent.

Special case of full supports and continuous Blackwell reducibility By Claim A.2, since $R(\mathcal{I}_0) = 1$, we have that $Q(\mathcal{C}_0) = 1$. By definition, $\mathcal{I}_0$ is a closed set. Thus, if ϕ is continuous, then $\mathcal{C}_0 = \phi^{-1}[\mathcal{I}_0]$ is a closed set. Hence, by definition of $\text{supp}(Q)$, we have $\text{supp}(Q) \subseteq \mathcal{C}_0$. Similarly, $\text{supp}(Q') \subseteq \mathcal{C}_0'$.

Under the assumption that $\text{supp}(Q) = \text{supp}(Q') = \mathcal{C}$, we thus obtain $\mathcal{C}_0 = \mathcal{C}_0' = \mathcal{C}$. Hence, we avoid all domain issues: τ_0 is already a well-defined injective function from $\mathcal{C}$ to $\mathcal{C}$. □

C.2.1 Measure-separating base kernel under additive noise

To prove that the base kernel in Ex. 4.2 is measure-separating, we use elementary properties of characteristic functions, which we now review; see Lukacs (1970) for reference. Given a probability distribution R over $\mathcal{U} = \mathbb{R}^d$, we recall that the ***characteristic function*** of R is defined as $\phi_R(t) := \mathbb{E}_{\mathbf{U} \sim R}[e^{it^\top \mathbf{U}}]$, and that the map $R \mapsto \phi_R$ is injective, i.e., characteristic functions are unique. Further, given two distributions R and N over $\mathcal{U}$, we have $\phi_{N * R}(t) = \phi_N(t) \cdot \phi_R(t)$.

Claim C.1. *Let $N_\varepsilon \in \mathcal{P}(\mathcal{Z})$ such that ϕ_{N_ε} is nowhere zero. Then the base kernel $B \colon z \mapsto N_\varepsilon * \delta_z$ is measure-separating.*

Proof. Consider $R, R' \in \mathcal{P}(\mathcal{Z})$. Let $Q = BR = N_\varepsilon * R$ and $Q' = BR' = N_\varepsilon * R'$. Then we have $\phi_Q(t) = \phi_{N_\varepsilon}(t) \cdot \phi_R(t)$ and $\phi_{Q'}(t) = \phi_{N_\varepsilon}(t) \cdot \phi_{R'}(t)$.

Suppose $Q = Q'$. Then $\phi_Q = \phi_{Q'}$, and since ϕ_{N_ε} is nowhere zero, we must have $\phi_R = \phi_{R'}$. Since characteristic functions are unique, this entails $R = R'$, i.e., B is measure-separating, as desired. □

The fact that the base kernel in Ex. 4.2 is measure-separating follows from Claim C.1, since the characteristic function of $N_\varepsilon = \mathcal{N}(0, \Sigma)$ is $\phi_{N_\varepsilon}(t) = \exp(-t^\top \Sigma t/2)$, which is nowhere zero.

C.2.2 Measure-separating base kernel in Gaussian mixture

Recall that, in Ex. 4.3, we have the concept $\mathcal{C} = [d]$, the index space $\mathcal{I} = \mathbb{R}^p \times \mathcal{S}^p_{\succeq 0}$, and the feature space $\mathcal{Z} = \mathbb{R}^p$. We defined the base kernel $B : (\mu, \Sigma) \mapsto \mathcal{N}(\mu, \Sigma)$, and we wished to show that B is measure-separating over measures in $\mathcal{R} := \{E_\phi Q : \phi \in \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{I}), Q \in \mathcal{P}(\mathcal{C})\}$. In particular, $\mathcal{R}$ contains all distributions of the form $R = \sum_{i=1}^r w_i \delta_{\mu_i, \Sigma_i}$, where $r \leq d$ and $(\mu_i, \Sigma_i) \neq (\mu_j, \Sigma_j)$ for $i \neq j$.

Consider $R, R' \in \mathcal{R}$ with $R = \sum_{i=1}^r w_i \delta_{\mu_i, \Sigma_i}$ and $R' = \sum_{i=1}^{r'} w'_i \delta_{\mu'_i, \Sigma'_i}$. Suppose $Q = BR$ and $Q' = BR'$ with $Q = Q'$. Then, by linearity of the pushforward operator, $B_\sharp(R - R') = 0$, where $R - R' \in \mathcal{R} - \mathcal{R}$ is a ***signed measure*** on $\mathcal{I}$, and $\mathcal{R} - \mathcal{R} := \{R - R' : R, R' \in \mathcal{R}\}$. Thus, if $\text{nullspace}(B_\sharp|_{\mathcal{R}-\mathcal{R}}) = \{0\}$, we can conclude that $R - R' = 0$, i.e., $R = R'$. This connection is essentially the same one described in the first theorem of Yakowitz and Spragins (1968), where B is treated as an indexed family of distributions (rather than as a Markov kernel), and the result is stated in terms of linear independence (rather than in terms of the nullspace). In particular, Proposition 2 of Yakowitz and Spragins (1968) can be viewed as showing the more general result that $\text{nullspace}(B_\sharp|_{\mathcal{S}}) = \{0\}$, where $\mathcal{S}$ is the set of signed measures with finite (discrete) support (this result is more general since $\mathcal{R} - \mathcal{R} \subset \mathcal{S}$).

C.3 Proofs from Sec. 4.2

In Sec. 4.2, we mentioned that any group G of functions from $\mathcal{C}$ to $\mathcal{C}$ defines an equivalence relation $\sim_G$ over $\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$. We now formally state and prove this result.

Proposition C.2. *Let G be a group of functions on $\mathcal{C}$, and define the binary relation Rel_G on $\mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z})$ as follows: $\text{Rel}_G(\mathsf{M}, \mathsf{M}')$ holds if and only M is Blackwell coarser than M' under some $\tau \in G$.*

Then, Rel_G is an equivalence relation.

Proof. We check the three axioms of an equivalence relation:

Reflexivity: Since G is a group, it contains the identity map $\text{id}_\mathcal{C} : c \mapsto c$. Clearly, M is a Blackwell coarsening of itself under the identity map, so $\text{Rel}_G(\mathsf{M}, \mathsf{M})$ holds.

Symmetry: Let $\mathsf{M} = (Q, K)$ be Blackwell coarser than $\mathsf{M}' = (Q', K')$ under $\tau \in G$, i.e., $Q' = E_\tau Q$ and $K \sim_Q K' E_\tau$. Since G is a group, we have $\tau^{-1} \in G$.

We now show that M' is Blackwell coarser than M under τ^{-1}.

- *Measure refinement:* By substitution and functoriality, we have $E_{\tau^{-1}}(Q') = (E_{\tau^{-1}} E_\tau)(Q)$. Since τ is injective, this gives $E_{\tau^{-1}} Q' = Q$, as desired.
- *Kernel coarsening:* By Claim A.3, $K \sim_Q K' E_\tau$ implies that $K E_{\tau^{-1}} \sim_{Q'} (K' E_\tau) E_{\tau^{-1}}$.[8] By functoriality and injectivity of τ, we obtain $K E_{\tau^{-1}} \sim_{Q'} K'$, as desired.

Transitivity: Let M be a Blackwell coarsening of M' under E_τ and let M' be a Blackwell coarsening of M'' under $E_{\tau'}$. By transitivity of Blackwell coarsening (see Prop. C.1), M is a Blackwell coarsening of M'' under $E_{\tau''} = E_\tau E_{\tau'}$. Since G is a group of functions, $\tau'' \in G$, hence $\text{Rel}_G(\mathsf{M}, \mathsf{M}'')$ holds. □

[8] In particular, we use Claim A.3 with $R = Q'$, $J = E_{\tau^{-1}}$, $L_1 = K$, $L_2 = K' E_\tau$, and $R' = Q = E_{\tau^{-1}} Q'$.

At the end of Sec. 4.2, we mentioned that the valid transition set $\mathcal{T}(\mathcal{K})$ takes a special form when $\mathcal{K}$ contains only injective deterministic functions from some set $\mathcal{G}$. As in Xi and Bloem-Reddy (2023), we need to assume that $\mathcal{G}$ contains only *image equivalent functions*, i.e., $g[\mathcal{C}] = g'[\mathcal{C}]$ for any $g, g' \in \mathcal{G}$, so that $g^{-1} \circ g'$ is well-defined. We now prove this claim.

Proposition C.3 (Valid kernel transitions in the deterministic case)**.** *Let $\mathcal{G} \subseteq \mathcal{F}_{\text{inj}}(\mathcal{C} \to \mathcal{Z})$ contain image equivalent functions, and assume $\mathcal{K} = \{E_g : g \in \mathcal{G}\}$. Then $\mathcal{T}(\mathcal{K}) = \{\tau : \tau \sim_\mu g^{-1} \circ g' \text{ for some } g, g' \in \mathcal{G}\}$.*

Proof. Let $\mathcal{T}(\mathcal{G}) = \{\tau : \tau \sim_\mu g^{-1} \circ g' \text{ for some } g, g' \in \mathcal{G}\}$. We first prove the $\mathcal{T}(\mathcal{K}) \subseteq \mathcal{T}(\mathcal{G})$, then that $\mathcal{T}(\mathcal{G}) \subseteq \mathcal{T}(\mathcal{K})$.

First direction: Let $\tau \in \mathcal{T}(\mathcal{K})$. In particular, there exist functions $g, g' \in \mathcal{G}$ such that $K \sim_\mu K' E_\tau$ for the mixing kernels $K = E_g$ and $K' = E_{g'}$. It is straightforward that $K \sim_\mu K' E_\tau$ implies that $g \sim_\mu g' \circ \tau$. Composing both sides with $(g')^{-1}$, we obtain $(g')^{-1} \circ g \sim_\mu \tau$, i.e., $\tau \in \mathcal{T}(\mathcal{G})$.

Second direction: Let $\tau \in \mathcal{T}(\mathcal{G})$. In particular, there exists functions $g, g' \in \mathcal{G}$ such that $\tau \sim_\mu g^{-1} \circ g'$. Composing both sides with g, we obtain $g \circ \tau \sim_\mu g'$. Thus, $K' \sim_\mu K E_\tau$ for $K = E_g$ and $K' = E_{g'}$. Hence, $\tau \in \mathcal{T}_{K'}(\mathcal{K})$, and since $\mathcal{T}_{K'}(\mathcal{K}) \subseteq \mathcal{T}(\mathcal{K})$, we have $\tau \in \mathcal{T}(\mathcal{K})$. □

We note that, to drop the image equivalence assumption, one would need to define the set

$$\mathcal{T}(\mathcal{G}) = \{\tau : \tau \sim_\mu g^{-1} \circ g' \text{ for some } g, g' \in \mathcal{G} \text{ such that } \operatorname{supp}(g_\sharp(\mu)) = \operatorname{supp}(g'_\sharp(\mu))\}$$

and carefully handle domain issues, similar to the proof of Thm. 4.1.

D DEFERRED PROOFS (APPLICATIONS)

D.1 Definition of the Canonical Stratified Measure

For $J \subseteq [d]$, let $S_J := \operatorname{span}(\{e_i : i \in J\})$. For $|J| = k$, we let λ_J denote the k-dimensional Lebesgue measure on S_J, i.e., $\lambda_J = (\iota_J)_\sharp \lambda^k$, where $\iota_J \colon \mathbb{R}^k \to S_J$ is the canonical isomorphism between $\mathbb{R}^k$ and S_J, and λ^k is Lebesgue measure on $\mathbb{R}^k$. Then, we define the ***canonical stratified measure*** λ_s^d referred to in Sec. 5.1 as

$$\lambda_s^d := \sum_{\substack{J \subseteq [d] \\ |J| \le s}} \lambda_J.$$

D.2 Proofs from Sec. 5.1

Before Claim 5.1, we cited the fact that any matrix with spark greater than $2s$ is injective on the set Σ_s^d. This classical result can be found in many places, e.g. Theorem 2.4 of Elad (2010). For completeness and consistency with our notation, we now formally state and prove this result.

Claim D.1. *Let $\mathsf{G} \in \mathbb{R}^{p \times d}$ with $\operatorname{spk}(\mathsf{G}) > 2s$, and let $\tau \colon c \mapsto \mathsf{G} \cdot c$. Then G is injective on Σ_s^d.*

Proof. Suppose $c, c' \in \Sigma_s^d$ and $\mathsf{G} \cdot c = \mathsf{G} \cdot c'$. Let $\Delta = c - c'$, so that $\mathsf{G} \cdot \Delta = 0$. Since $\|c\|_0 \le s$ and $\|c'\|_0 \le s$, we have $\|\Delta\|_0 \le 2s$. Thus, $\mathsf{G} \cdot \Delta$ is a linear combination of at most $2s$ columns of G. Since $\operatorname{spk}(\mathsf{G}) > 2s$, these columns are linearly independent, so $\mathsf{G} \cdot \Delta = 0$ implies that $\Delta = 0$. This shows that $c = c'$, as desired. □

Our main claim in Sec. 5.1 characterizes the valid transition set $\mathcal{T}(\mathcal{K}_{\text{spk}})$ for the set of mixing functions $\mathcal{K}_{\text{spk}} = \mathcal{K}(\Sigma_s^d \to \mathbb{R}^p; \mathsf{Linear}, \mathsf{Spark}_\sigma)$. This characterization is the main difficulty in proving identifiability of $\mathcal{M}_{\text{spk}}$, and a version of this characterization appears in the proof of Theorem 3 from Aharon et al. (2006).

Here, we hope to provide more intuition for the combinatorial nature of this proof. For $s \le d$, we define $\mathcal{J}(d, s)$ as the collection of all size-s subsets of $[d]$, i.e.,

$$\mathcal{J}(d, s) := \{J \subseteq [d] : |J| \le s\}.$$

For $i \in [d]$, we define $\mathcal{J}_i(d,s)$ as the collection of all size-s subsets of $[d]$ which contain i, i.e.,

$$\mathcal{J}_i := \{J \in \mathcal{J} : i \in J\}.$$

Claim D.2. *Let $\mathsf{T} \in \mathbb{R}^{d\times d}$ be an invertible matrix. For $s \le d-1$, assume $\mathsf{T}(\Sigma_s^d) \subseteq \Sigma_s^d$. Then $\mathsf{T} \in \mathrm{Relabel}(d) \circ \mathrm{Scale}(d)$.*

Proof. For $i \in [d]$, let e_i denote the i^{th} canonical basis vector. For $J \subseteq [d]$, let S_J denote the coordinate-aligned subspace spanned by the corresponding basis vectors, i.e., $S_J := \mathrm{span}(\{e_i : i \in J\})$, and let $W_J := T[S_J]$. We note that, for $s \le d-1$, any s-dimensional subspace S of Σ_s^d is coordinate-aligned, i.e., if $\dim(S) = s$, then $S = S_J$ for some $J \in \mathcal{J}(d,s)$.

Induced mapping on size-s coordinate-aligned subspaces: Let $J \in \mathcal{J}(d,s)$. Since T is invertible, we have $\dim(W_J) = s$, and since $\mathsf{T}(\Sigma_s^d) \subseteq \Sigma_s^d$, we have $W_J = S_{J'}$ for some $J' \in \mathcal{J}$. Let $\phi\colon \mathcal{J}(d,s) \to \mathcal{J}(d,s)$ be the map which sends $J \in \mathcal{J}(d,s)$ to $J' \in \mathcal{J}(d,s)$ such that $W_J = S_{J'}$.

Induced mapping on axes: Let $i \in [d]$. Since T is invertible, we have $\dim(W_{\{i\}}) = 1$. Note that $S_{\{i\}} = \bigcap_{J\in\mathcal{J}_i(d,s)} S_J$, so by definition of the image and ϕ, we have $W_{\{i\}} = \bigcap_{J\in\mathcal{J}_i(d,s)} S_{\phi(J)}$, i.e., $W_{\{i\}}$ must be a coordinate-aligned subspace. Thus, since $\dim(W_i) = 1$, we must have $W_{\{i\}} = S_j$ for some $j \in [d]$.

Conclusion: The induced mapping on axes shows that, for each $i \in [d]$, there is one nonzero entry in the i^{th} column of T. Since T is invertible, it must be a generalized permutation matrix, i.e., $\mathsf{T} \in \mathrm{Relabel}(d) \circ \mathrm{Scale}(d)$. □

Using this result, the characterization of $\mathcal{T}(\mathcal{K}_{\text{spk}})$ becomes straightforward:

Claim (Restatement of Claim 5.1)**.** *Let $K, K' \in \mathcal{K}_{\text{spk}}$ defined in Eq. (1), and assume $\sigma > 2s$. If $K \sim_{\lambda_s^d} K'E_\tau$ for some $\tau \in \mathcal{F}_{\text{inj}}(\Sigma_s^d \to \Sigma_s^d)$, then $\tau \in \mathrm{Relabel}(d) \circ \mathrm{Scale}(d)$.*

More concisely, $\mathcal{T}(\mathcal{K}_{\text{spk}}) = \mathrm{Relabel}(d) \circ \mathrm{Scale}(d)$.

Proof. First, we unpack definitions. Since $K, K' \in \mathcal{K}(\Sigma_s^d \to \mathbb{R}^p; \mathsf{Linear}, \mathsf{Spark}_\sigma)$, there exist matrices $\mathsf{G}, \mathsf{G}' \in \mathbb{R}^{p\times d}$, with $\mathrm{spk}(\mathsf{G}) \ge \sigma > 2s$ and $\mathrm{spk}(\mathsf{G}') \ge \sigma > 2s$, such that $K\colon c \mapsto \delta_{\mathsf{G}\cdot c}$ and $K'\colon c \mapsto \delta_{\mathsf{G}'\cdot c}$.

Since G and G' are both injective on Σ_s^d and linear, and τ is injective, τ must also be linear. Thus, we represent τ by an invertible matrix $\mathsf{T} \in \mathbb{R}^{d\times d}$ such that $\tau\colon c \mapsto \mathsf{T}\cdot c$. Moreover, since $\tau \in \mathcal{F}_{\text{inj}}(\Sigma_s^d \to \Sigma_s^d)$, T must be sparsity-preserving, i.e., $\|\mathsf{T}\cdot c\|_0 \le s$ for any c such that $\|c\|_0 \le s$.

Applying Claim D.2 on T, we obtain the result. □

D.3 Proofs from Sec. 5.2

In Sec. 5.2, we recounted the classic result that $\mathcal{T}(\mathcal{K}) \cap \mathcal{T}(\mathcal{Q}_{\text{ind}}^d) = O(d)$ when $\mathcal{K} = \mathcal{K}(\mathbb{R}^d \to \mathbb{R}^p; \mathsf{Linear})$, where $O(d)$ denotes the orthogonal group on $\mathbb{R}^d$. For completeness, we give a brief proof.

Claim D.3. *For $d \le p$, let $\mathcal{Q}_{\text{ind}}^d = \mathcal{P}(\mathbb{R}^d; \mathsf{Ind}, \mathsf{UnitVar})$ and let $\mathcal{K} = \mathcal{K}(\mathbb{R}^d \to \mathbb{R}^p; \mathsf{LinearInj})$. Then $\mathcal{T}(\mathcal{K}) \cap \mathcal{T}(\mathcal{Q}_{\text{ind}}^d) = O(d)$.*

Proof. First, we show that $\mathcal{T}(\mathcal{K}) \cap \mathcal{T}(\mathcal{Q}_{\text{ind}}) \subseteq O(d)$. Recall that $\mathcal{T}(\mathcal{K}) = \mathrm{GL}(d)$.

Let $\tau\colon \mathbf{c} \to \mathsf{T}\cdot\mathbf{c}$ for some invertible matrix $\mathsf{T} \in \mathbb{R}^{d\times d}$, so that $\tau \in \mathcal{T}(\mathcal{K})$. Further, assume $Q, Q' \in \mathcal{Q}_{\text{ind}}$ with $Q' = \tau_\sharp(Q)$, so that $\tau \in \mathcal{T}(\mathcal{Q}_{\text{ind}})$.

By definition of the $\mathsf{UnitVar}$ concept predicate, we have $\mathsf{Cov}(Q) = \mathsf{Cov}(Q') = \mathsf{I}_d$. By linearity of expectation, we have $\mathsf{Cov}(Q') = \mathsf{T}\cdot\mathsf{Cov}(Q)\cdot\mathsf{T}^\top$. Thus, we obtain $\mathsf{T}\cdot\mathsf{T}^\top = \mathsf{I}_d$, i.e., $\mathsf{T} \in O(d)$.

Now, we show that $O(d) \subseteq \mathcal{T}(\mathcal{K}) \cap \mathcal{T}(\mathcal{Q}_{\text{ind}})$. Let $\tau \in O(d) \subseteq \mathcal{T}(\mathcal{K})$. Then, for $Q = Q' = \mathcal{N}(\mathbf{0}, \mathsf{I}_d)$, we have $Q' = \tau_\sharp(Q)$, i.e., $\tau \in \mathcal{T}(\mathcal{Q}_{\text{ind}})$, completing the proof. □

Our main claim in Sec. 5.2 is a minor adaptation of Theorem 11 in Comon (1994), which shows that any orthogonal transformation of a product distribution with non-Gaussian components will no longer be a product distribution, unless the transformation is a generalized permutation.[9] The proof is a corollary of Darmois's theorem, which we restate here for convenience:

Theorem (Darmois's theorem)**.** *Let $(C_j)_{j=1}^d$ be a collection of independent random variables. Let $C_1' = \sum_{j=1}^d a_j \cdot C_j$, and let $C_2' = \sum_{j=1}^d b_j \cdot C_j$. Finally, let $\mathcal{I} = \{j \in [d] : a_j \cdot b_j \neq 0\}$.*

If C_1' and C_2' are independent, then all $(C_j)_{j \in \mathcal{I}}$ have Gaussian distributions (possibly with zero variance).

As a corollary of Darmois's theorem, we obtain the following:

Corollary D.1. *Let $Q, Q' \in \mathcal{P}(\mathbb{R}^d; \mathsf{Ind})$. Suppose $Q' = \tau_\sharp(Q)$ for $\tau \colon c \mapsto \mathsf{T} \cdot c$, where T has at least two nonzero entries in the k^{th} column, i.e., $\mathsf{T}_{uk} \neq 0$ and $\mathsf{T}_{vk} \neq 0$ for $u \neq v$. Then the k^{th} component of Q has a Gaussian distribution.*

Proof. Define random variables $C \sim Q$ and let $C' = \mathsf{T} \cdot C$, so that $C' \sim Q'$. We have $C_u' = \sum_{j=1}^d \mathsf{T}_{uj} \cdot C_j$ and $C_v' = \sum_{j=1}^d \mathsf{T}_{vj} \cdot C_j$. By the assumption that $Q' \in \mathcal{P}(\mathbb{R}^d; \mathsf{Ind})$, C_u' and C_v' are independent. By the premise of the corollary, $\mathsf{T}_{uk} \cdot \mathsf{T}_{vk} \neq 0$. Hence, by Darmois's theorem, C_k has a Gaussian distribution. □

Applying this corollary and using properties of orthogonal matrices, we obtain the main result:

Claim (Restatement of Claim 5.3)**.** $\mathcal{T}(\mathcal{Q}_{\text{ind-ng}}^d) \cap O(d) = \text{Relabel}(d)$.

Proof. The direction $\text{Relabel}(d) \subseteq \mathcal{T}(\mathcal{Q}_{\text{ind-ng}}) \cap O(d)$ is straightforward, since non-Gaussianity and independence are unaffected by relabeling.

To show that $\mathcal{T}(\mathcal{Q}_{\text{ind-ng}}) \cap O(d) \subseteq \text{Relabel}(d)$, let $\tau \colon c \mapsto \mathsf{T} \cdot c$ for some orthogonal matrix T which is not a permutation matrix. Then there must be at least two columns $j, k \in [d]$ such that T has at least two nonzero entries in each of these columns. Next, suppose that $Q' = E_\tau Q$ for $Q, Q' \in \mathcal{P}(\mathbb{R}^d; \mathsf{Ind})$. By Cor. D.1, the j^{th} and k^{th} component of Q must be Gaussian. Since this applies for any Q and Q', we have $\tau \notin \mathcal{T}(\mathcal{Q}_{\text{ind-ng}})$. Thus, $\mathcal{T}(\mathcal{Q}_{\text{ind-ng}}) \cap O(d) \subseteq \text{Relabel}(d)$. □

E METHODOLOGICAL IMPLICATIONS FOR SPARSE AUTOENCODERS

E.1 Post-hoc Checks

Compared to the model class $\mathcal{M}_{\text{spk}}(d, p; s, \sigma)$, the model class $\mathcal{M}_{\text{sae}}(d, p; s)$ lacks the spark constraint over the mixing kernel K. As discussed in Sec. 5.1, the spark constraint ensures injectivity of the mixing kernel, which is required for the class of mixing kernels $\mathcal{K}_{\text{spk}}$ to be Blackwell reducible, and thus for the proof of Claim 5.2. This raises an important question: if one learns a model $\mathsf{M} \in \mathcal{M}_{\text{sae}}(d, p; s)$, what can be concluded about the identifiability of this model?

As a partial answer to this question, we suggest performing a post-hoc check to determine whether $\mathsf{M} \in \mathcal{M}_{\text{spk}}(d, p; s, \sigma)$. In particular, after training a sparse autoencoder, one may check whether the matrix G associated with its decoder obeys the desired spark constraint, i.e., whether $\text{spk}(\mathsf{G}) > 2s$. If this condition holds, then one can make the following conclusion from Claim 5.2: the model M is identifiable *within* the model class $\mathcal{M}_{\text{spk}}(d, p; s, \sigma)$. This conclusion reflects a generally useful strategy: for computational reasons, one may avoid enforcing some constraints during optimization (e.g. the spark constraint), and instead rely on post-hoc checks to determine identifiability guarantees. However, one must be mathematically careful in such cases: identifiability guarantees must be stated *with respect to a particular model class* (*c.f.* the definition of identifiability in Sec. 4.2). Finally, we note that for such checks, computing the spark of a matrix is NP-hard, but there are computationally tractable lower bounds (e.g. based on mutual incoherence) (Elad, 2010).

[9] As will be evident from the proof, one Gaussian component is allowed, and deterministic components need to be ruled out; these are built into the definition of the $\mathsf{NonGaussian}$ mixing predicate.

E.2 Amortized Posterior Inference

When concept extraction needs to be performed efficiently, the posterior inference methods described in Sec. 6.3 may be too slow, e.g. even for a deterministic decoder $g_\psi \colon c \mapsto \mathsf{G}_\psi \cdot c$, exactly computing the inverse $g_\psi^{-1}(z)$ via matching pursuit may be too expensive. Hence, one may instead wish to learn a parameterized concept extractor (i.e., an encoder) H_ϕ for some parameters $\phi \in \Phi$, e.g. by training on a loss function that enforces $H_\phi \approx H_{\theta^*}$.

Here, we discuss two options for such amortization: *post-hoc amortization*, where ϕ is optimized after training the generative model $\mathsf{M}_{\theta^*} = (Q_{\omega^*}, K_{\psi^*})$, and *end-to-end amortization*, where ϕ is optimized at the same time as $\theta = (\omega, \psi)$. As a reminder, for a model $\mathsf{M}_\theta = (Q_\omega, K_\psi)$, we let $J_\theta(c, z) = K_\psi(z \mid c) \cdot Q_\omega(c)$ denote its associated joint distribution over concepts and features, and we let $P_\theta(z) = \int K_\psi(z \mid dc) \cdot Q_\omega(dc)$ denote its associated marginal distribution over features. Further, we let $H_\theta(c \mid z) = \frac{K_\psi(z|c) \cdot Q_\omega(c)}{P_\theta(z)}$ denote the induced posterior. Finally, we will see that many approaches can be interpreted as performing an amortized version of *maximum a posteriori* (MAP) inference, so we define the ***induced MAP function*** $\overline{h}_\theta \colon \mathcal{Z} \to \mathcal{C}$ as follows:

$$\overline{h}_\theta(z) := \arg\max_{c \in \mathcal{C}} H_\theta(c \mid z) = \arg\max_{c \in \mathcal{C}} \log K_\psi(z \mid c) + \log Q_\omega(c),$$

where for simplicity we assume there is a unique maximizer for all $z \in \mathcal{Z}$.

SAE Running Example To connect the maximum likelihood perspective (common in generative modeling) with the regularized reconstruction perspective (common for sparse autoencoders), we follow Geadah et al. (2024), who (in our terminology) consider mixing kernels of the form $K_\psi(\cdot \mid c) = \mathcal{N}(\mathsf{G}_\psi \cdot c, \sigma^2 I)$, and who consider a fixed concept distribution Q, e.g. a product of Laplace distributions $Q(c_i) \propto \exp(-\alpha|c_i|)$ over each component i of $\mathbf{C}$. Under these choices, one obtains

$$\log K_\psi(z \mid c) \simeq -\frac{1}{2\sigma^2}\|z - \mathsf{G}_\psi \cdot c\|_2^2 \quad \text{and} \quad \log Q(c) \simeq -\alpha\|c\|_1, \tag{7}$$

where $\simeq$ denotes equality up to an absolute constant.

E.2.1 Post-hoc amortization

Assuming evaluation access to H_ϕ, a natural choice is the cross entropy loss

$$\mathcal{L}_{\mathrm{CE}}(\phi; \theta^*) := \mathbb{E}_{(\mathbf{C},\mathbf{Z}) \sim J_{\theta^*}} \left[-\log H_\phi(\mathbf{C} \mid \mathbf{Z}) \right],$$

which by well-known theory is optimized when $H_\phi(c \mid z) = H_{\theta^*}(c \mid z)$ almost everywhere on P_{θ^*} (assuming $\{H_\phi\}_{\phi \in \Phi}$ has universal approximation capability). This approach would resemble recent *neural posterior estimation* approaches in simulation-based inference (Lueckmann et al., 2017; Deistler et al., 2025) and *inference compilation* approaches in probabilistic programming (Le et al., 2017), with roots in the "wake-sleep" algorithm (Hinton et al., 1995). As an illuminating example, one may consider optimizing over concept extractors of the form $H_\phi(\cdot \mid z) = \mathcal{N}(h_\phi(z), \sigma^2 I)$, giving

$$\mathcal{L}_{\mathrm{CE}}(\phi; \theta^*) \simeq \mathbb{E}_{(\mathbf{C},\mathbf{Z}) \sim J_{\theta^*}} \left[-\frac{1}{2\sigma^2}\|\mathbf{C} - h_\phi(\mathbf{Z})\|_2^2 \right],$$

a least-squares objective for reconstructing the sampled concept $\mathbf{C}$ which generated the associated $\mathbf{Z}$.

Amortized MAP inference As described in Geadah et al. (2024), the use of a *deterministic* encoder can be thought of as performing maximum a posteriori (MAP) inference. Hence, to learn an amortized MAP function h_λ (with the aim that $h_\phi \approx \overline{h}_\theta$), one may minimize the objective

$$\mathcal{L}_{\mathrm{MAP}}(\phi; \theta^*) := \mathbb{E}_{\mathbf{Z} \sim P^*} \left[-\log K_{\psi^*}(\mathbf{Z} \mid h_\phi(\mathbf{Z})) - \log Q_{\omega^*}(h_\phi(\mathbf{Z})) \right].$$

Thus, assuming the log-likelihoods in Eq. (7), one obtains

$$\mathcal{L}_{\mathrm{MAP}}(\phi; \theta^*) \simeq \mathbb{E}_{\mathbf{Z} \sim P^*} \left[\frac{1}{2\sigma^2}\|\mathbf{Z} - \mathsf{G}_{\psi^*} \cdot h_\phi(\mathbf{Z})\|_2^2 + \alpha\|h_\phi(\mathbf{Z})\|_1 \right],$$

which is precisely the standard SAE objective, but with the decoder parameters ψ frozen to $\psi = \psi^*$.

E.2.2 End-to-end amortization

As a final option, the concept extractor H_ϕ could be trained in parallel with the generative model M_θ. When the prior is learnable, such end-to-end training could be performed using a general form of the evidence lower bound (ELBO)[10], as found in Tomczak and Welling (2018):

$$\mathbb{E}_{\mathbf{Z}\sim P^*}[-\log P^*(\mathbf{Z})] \le \mathcal{L}_{\text{GELBO}}(\phi,\psi,\omega), \text{ where}$$
$$\mathcal{L}_{\text{GELBO}}(\phi,\psi,\omega) := \mathbb{E}_{\mathbf{Z}\sim P^*,\mathbf{C}\sim H_\psi(\cdot\mid\mathbf{Z})}\left[-\log K_\psi(\mathbf{Z}\mid\mathbf{C}) - \log Q_\omega(\mathbf{C}) + \log H_\phi(\mathbf{C}\mid\mathbf{Z})\right]$$

When the prior is fixed to Q (e.g. a Laplace distribution as describe above), this bound reduces to the more standard ELBO:

$$\mathcal{L}_{\text{ELBO}}(\phi,\psi) := \mathbb{E}_{\mathbf{Z}\sim P^*,\mathbf{C}\sim H_\psi(\cdot\mid\mathbf{Z})}\left[\log K_\psi(\mathbf{Z}\mid\mathbf{C}) + \log\frac{Q(\mathbf{C})}{H_\phi(\mathbf{C}\mid\mathbf{Z})}\right].$$

To connect our framework with traditional SAEs (which use deterministic encoders), one may take the regularization perspective described in Ghosh et al. (n.d.). Alternatively, to continue with the maximum likelihood perspective, one may consider the ***deterministic variational family*** $H_\phi = E_{h_\phi}$. Formally, this choice cannot be interpreted as a density, but can be considered as a certain limit of the variational distributions $H_\phi^\beta(c\mid z) = \mathcal{N}(h_\phi(z),\beta^{-1}I)$ as $\beta\to\infty$[11], yielding

$$\mathcal{L}_{\text{ELBO-lim}}(\phi,\psi) := \mathbb{E}_{\mathbf{Z}\sim P^*}\left[-\log K_\psi(\mathbf{Z}\mid h_\phi(\mathbf{Z}))\right].$$

Under the likelihood in Eq. (7), we thus have

$$\mathcal{L}_{\text{ELBO-lim}}(\phi,\psi) \simeq \mathbb{E}_{\mathbf{Z}\sim P^*}\left[\frac{1}{2\sigma^2}\|\mathbf{Z} - \mathsf{G}_\psi\cdot h_\phi(\mathbf{Z})\|_2^2\right],$$

the classic reconstruction loss for SAEs. Hence, the final distinction between existing SAE approaches is their choice of the deterministic variational family: many SAEs restrict to *projection encoders* (as described below), whereas matching pursuit SAE (Costa et al., 2025) and other methods (Tolooshams and Ba, 2021; Xiao et al., 2025) use variational family which more closely resemble classical methods for computing g_ψ, such as matching pursuit or the iterative shrinkage and thresholding algorithm (Daubechies et al., 2004; Gregor and LeCun, 2010).

E.3 Comparison to the Projective Assumptions Framework

Hindupur et al. (2025) discuss several SAE architectures, including ReLU SAEs (Cunningham et al., 2023), TopK SAEs (Makhzani and Frey, 2013), JumpReLU SAEs (Rajamanoharan et al., 2024; Lieberum et al., 2024), and their newly-introduced SparseMax SAEs. They show that each of these architectures uses a ***constraint set*** $\mathcal{S}\subseteq\mathbb{R}^d$ and a ***projection encoder*** $h\colon\mathcal{Z}\to\mathcal{S}$ of the form

$$h_\phi(z) = \Pi_{\mathcal{S}}(\mathsf{W}_\phi\cdot z + b_\phi) \tag{8}$$

for some matrix $\mathsf{W}_\phi\in\mathbb{R}^{d\times p}$ and some vector $b_\phi\in\mathbb{R}^d$, where $\Pi_{\mathcal{S}}$ denotes the projection operator onto $\mathcal{S}$.[12] As noted in Hindupur et al. (2025), the form of h_ϕ entails implicit assumptions over (1) the geometry of the concept space $\mathcal{C}$ and (2) the functional relationship between concepts and features. Our framework proposes to make these assumptions more transparent by (1) explicitly defining the concept space $\mathcal{C}$ and (2) explicitly stating assumptions on the model class $\mathcal{M}$.

One straightforward implication of assuming a projection encoder is that the concept space $\mathcal{C}$ is constrained by the constraint set $\mathcal{S}$ of the projection operator. For example, since ReLU SAEs project onto the nonnegative orthant (i.e., $\text{ReLU}(v) = \Pi_{\mathcal{S}}(v)$ for $\mathcal{S} = \mathbb{R}^d_{\ge 0}$), they impose $\mathcal{C}\subseteq\mathbb{R}^d_{\ge 0}$; since TopK SAEs project onto the the set of nonnegative sparse vectors, they impose $\mathcal{C}\subseteq\mathbb{R}^d_{\ge 0}\cap\Sigma^d_s$; and so on, as summarized in Table 1.

[10] For consistency with other loss functions, we present the ELBO as an upper bound on the *negative* log likelihoood.

[11] This reduction can also be seen more intuitively, though less rigorously: under $H_\phi = E_{h_\phi}$, the expectation is only evaluated at $\mathbf{C} = h_\phi(\mathbf{Z})$, where $H_\phi(\mathbf{C}\mid\mathbf{Z}) = \infty$ and the second term vanishes.

[12] Strictly speaking, the projection operator $\Pi_{\mathcal{S}}(v) = \arg\min_{v'\in\mathcal{S}}\|v - v'\|_2^2$ is only well-defined when $\mathcal{S}$ is closed and convex, so that the minimizer exists and is unique. This point was not explicitly mentioned in Hindupur et al. (2025), but applies to the constraint sets they considered.

Model	$\mathcal{S}$
ReLU SAE	$\mathbb{R}^d_{\geq 0}$
TopK	$\mathbb{R}^d_{\geq 0} \cap \Sigma^d_s$
Heaviside	$\{0,1\}^d$
JumpReLU	$\mathbb{R}^d_{\geq 0}$
SparseMax	Δ^{d-1}

Table 1: Constraint sets used by different SAE architectures.

However, the assumption of a projection encoder imposes a less straightforward assumption on the model class $\mathcal{M}$. In general, this assumption cannot be decomposed into separate assumptions on the concept distribution Q and the mixing kernel K. Instead, the implicit model class used by an SAE with a linear decoder and the projection encoder in Eq. (8) is as follows:

$$\mathcal{M}_{\text{proj-enc}}(d, p; \mathcal{S}) = \mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z}; \mathsf{ProjEnc}_{\mathcal{S}}),$$

where $\mathsf{ProjEnc}_{\mathcal{S}} \colon \mathcal{M}_{\text{all}}(\mathcal{C}, \mathcal{Z}) \to \{\textsc{True}, \textsc{False}\}$ is a *model predicate*, which is $\textsc{True}$ for $\mathsf{M} = (Q, K)$ if and only if the following conditions hold:

1. $K \colon c \mapsto \mathsf{G} \cdot c$ for some matrix $\mathsf{G} \in \mathbb{R}^{p \times d}$.
2. G is injective on $\operatorname{supp}(Q)$, with a left-inverse of the form $h(z) = \Pi_{\mathcal{C}}(\mathsf{W} \cdot z + b)$

In particular, the second condition involves *both* the concept distribution Q and the mixing kernel K. Due to this lack of product structure on the model class, Thm. 4.2 cannot be applied to these settings, and we propose that the use of projection encoders should be treated as an *approximation* to the true left-inverse h (as discussed in Sec. 6, rather than as additional assumption. More generally, this discrepancy reflects well-known dichotomies, e.g. the dichotomy between *generative* and *discrimative* models (Ng and Jordan, 2001), and the dichotomy between causal and anticausal learning (Kügelgen et al., 2020). Indeed, in the general case of a stochastic mixing kernel (decoder), we have

$$H^{\mathsf{M}}(c \mid z) = \frac{K(z \mid c) \cdot Q(c)}{\int_{\mathcal{C}} K(z \mid dc) \cdot Q(c)}$$

by Bayes' rule, indicating that assumptions directly on H^{M} should typically be expected to involve assumptions on *both* Q and K. Given our focus on the generative perspective, we do not explore this issue further here, but note it as an interesting direction for future elaboration upon the framework presented here.

F ADDITIONAL APPLICATION: FINITE MIXTURE MODELS

As a final example, we consider a model class with stochastic mixing functions. Broadly speaking, a *finite mixture model* is a latent generative concept model where $\mathcal{C} = [d]$. We endow $\mathcal{C}$ with the standard counting measure μ, and we employ the concept predicate $\mathsf{MeasCls}_\mu$, so that $\mathsf{MeasCls}_\mu(Q)$ holds if and only if $Q(\{i\}) > 0$ for all $i \in [d]$. Other than this constraint, we make no restrictions on the concept distribution.

For simplicity, we consider the case of *Gaussian* mixture models. In particular, we define the mixing predicate $\mathsf{DistinctGaussian}$, where $\mathsf{DistinctGaussian}(K)$ holds if and only $K(\cdot \mid c)$ is a multivariate Gaussian distribution for all $c \in \mathcal{C}$, and $K(\cdot \mid c) \neq K(\cdot \mid c')$ for any $c \neq c'$.

Putting these constraints together, we obtain the model class

$$\mathcal{M}_{\text{gmm}}(d, p) := \mathcal{P}([d]; \mathsf{MeasCls}_\mu) \times \mathcal{K}([d] \to \mathbb{R}^p; \mathsf{DistinctGaussian}).$$

As discussed in Ex. 4.3, $\mathcal{K}([d] \to \mathbb{R}^p; \mathsf{DistinctGaussian})$ is Blackwell reducible, so we can apply Cor. 4.1, which gives the result almost immediately. Here, we use $\sim_{\text{perm}}$ as shorthand for the equivalence relation $\sim_G$ with $G = \operatorname{Perm}(d)$, where $\operatorname{Perm}(d)$ denotes the set of all bijective functions $\tau \colon [d] \to [d]$.

Claim F.1. *$\mathcal{M}_{\text{gmm}}$ is identifiable up to $\sim_{\text{perm}}$.*

Proof. Let $\mathcal{Q} = \mathcal{P}([d]; \mathsf{MeasCls}_\mu)$. Then $\mathcal{T}(\mathcal{Q}) = \text{Perm}(d)$, so we immediately have the result. □

As this result shows, the main obstacle to showing identifiability of certain classes of finite mixture models is showing that the kernel class $\mathcal{K}$ is Blackwell reducible. Indeed, showing this property is the essence of several other classic identifiability results for finite mixture models, e.g. Yakowitz and Spragins (1968).

Additional References

Aharon, Michal et al. (2006). "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them". In: *Linear Algebra and Its Applications*.

Bogachev, Vladimir Igorevich and Maria Aparecida Soares Ruas (2007). *Measure theory*. Springer.

Comon, Pierre (1994). "Independent component analysis, a new concept?" In: *Signal Processing*.

Costa, Valérie et al. (2025). "From flat to hierarchical: Extracting sparse representations with matching pursuit". In: *Advances in Neural Information Processing Systems*.

Cunningham, Hoagy et al. (2023). "Sparse autoencoders find highly interpretable features in language models". In: *arXiv preprint arXiv:2309.08600*.

Daubechies, Ingrid et al. (2004). "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint". In: *Communications on Pure and Applied Mathematics*.

Deistler, Michael et al. (2025). "Simulation-based inference: A practical guide". In: *arXiv preprint arXiv:2508.12939*.

Elad, Michael (2010). *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Science & Business Media.

Fritz, Tobias (2020). "A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics". In: *Advances in Mathematics*.

Geadah, Victor et al. (2024). "Sparse-coding variational autoencoders". In: *Neural Computation*.

Ghosh, Partha et al. (n.d.). "From variational to deterministic autoencoders". In: *International Conference on Learning Representations*.

Gregor, Karol and Yann LeCun (2010). "Learning fast approximations of sparse coding". In: *International Conference on Machine Learning*.

Hindupur, Sai Sumedh R et al. (2025). "Projecting assumptions: The duality between sparse autoencoders and concept geometry". In: *Advances in Neural Information Processing Systems*.

Hinton, Geoffrey E et al. (1995). "The wake-sleep algorithm for unsupervised neural networks". In: *Science*.

Kallenberg, Olav (2011). "Invariant Palm and related disintegrations via skew factorization". In: *Probability Theory and Related Fields*.

Kechris, Alexander (1995). *Classical Descriptive Set Theory*. Springer Science & Business Media.

Kügelgen, Julius et al. (2020). "Semi-supervised learning, causality, and the conditional cluster assumption". In: *Conference on Uncertainty in Artificial Intelligence*.

Le, Tuan Anh et al. (2017). "Inference compilation and universal probabilistic programming". In: *Artificial Intelligence and Statistics*. PMLR.

Lieberum, Tom et al. (2024). "Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2". In: *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*.

Lueckmann, Jan-Matthis et al. (2017). "Flexible statistical inference for mechanistic models of neural dynamics". In: *Advances in Neural Information Processing Systems*.

Lukacs, Eugene (1970). *Characteristic functions*.

Makhzani, Alireza and Brendan Frey (2013). "K-sparse autoencoders". In: *arXiv preprint arXiv:1312.5663*.

Ng, Andrew and Michael Jordan (2001). "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". In: *Advances in Neural Information Processing Systems*.

Panangaden, Prakash (1999). "The category of Markov kernels". In: *Electronic Notes in Theoretical Computer Science*.

Rajamanoharan, Senthooran et al. (2024). "Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders". In: *arXiv preprint arXiv:2407.14435*.

Tolooshams, Bahareh and Demba Ba (2021). "Stable and interpretable unrolled dictionary learning". In: *arXiv preprint arXiv:2106.00058*.

Tomczak, Jakub and Max Welling (2018). "VAE with a VampPrior". In: *International Conference on Artificial Intelligence and Statistics*.

Xi, Quanhan and Benjamin Bloem-Reddy (2023). "Indeterminacy in generative models: Characterization and strong identifiability". In: *The International Conference on Artificial Intelligence and Statistics*.

Xiao, Pan et al. (2025). "SC-VAE: Sparse coding-based variational autoencoder with learned ISTA". In: *Pattern Recognition*.

Yakowitz, Sidney J and John D Spragins (1968). "On the identifiability of finite mixtures". In: *The Annals of Mathematical Statistics*.