Complete or Sparse: A Tale of Two Identifiabilities

Junze Zhou
Cold Spring Harbor Laboratory
juzhou@cshl.edu

David Klindt
Cold Spring Harbor Laboratory
klindt@cshl.edu

Abstract

Understanding when neural networks learn interpretable representations is central to mechanistic interpretability. We study this through identifiability: when can supervised learning recover the true latent variables underlying the data? We introduce two *data generating processes* distinguished by their *causal direction* where either labels cause latents (*generative*) or latents cause labels (*discriminative*). When the representation has at least as many dimensions as the task-relevant latent dimension, we prove that the optimal encoders are *linear* in the latents. Conversely, when dimensions are insufficient, identifiability generically fails and optimal encoders are *nonlinear*. We then show that additional constraints yield recovery of individual latent variables, not just linear mixtures. In the generative setting at exact capacity, non-negative activations and energy efficiency force each neuron encode exactly one latent, offering one possible account of selective *grandmother cell*-like responses. In the discriminative setting, when latents are sparse and the true measurement matrix satisfies the spark condition, the learned encoder inherits *compressed sensing* structure, enabling sparse dictionary learning to perfectly recover individual latents. This constitutes, to our knowledge, the first identifiability result for undercomplete nonlinear independent component analysis. Our framework connects nonlinear ICA, compressed sensing, and neural network interpretability, delineating exactly when representations transition from nonlinear distortions to linear representations and, under additional constraints or steps, to clean, axis-aligned codes. Theoretical predictions are validated on synthetic benchmarks and on activations drawn from vision models, large language models, and macaque inferotemporal cortex, providing suggestive evidence in real systems.

1 Introduction

The discovery of "grandmother cells" [1], neurons that selectively respond to highly specific concepts such as the "Jennifer Aniston neuron" in human medial temporal lobe [2], has challenged traditional theories of neural coding [3]. Similar observations have emerged in modern vision [4] and large language models, exemplified by neurons selectively responding to concepts such as certain despots [5]. Specific concepts, encoded by individual neurons, have even been found across models [6] and brains [7]. The existence of these highly interpretable individual neurons is surprising since dedicated single-neuron representations are generally assumed to be computationally inefficient and fragile [8].

Perhaps more surprising still is the presence of interpretable concepts (*latent variables*) within such nonlinear systems which could, in principle, reformat information in highly unintelligible ways [9]. Recent theoretical advances in identifiability theory, particularly nonlinear Independent Component Analysis (ICA), suggest that neural populations optimally recover the underlying latent variables of the data up to an arbitrary linear transformation [9, 10]. With such an arbitrary linear transformation, individual neurons are almost certainly responsive to linear combinations of latent variables. This means that they exhibit what is called *mixed selectivity* [8] in neuroscience, or, *polysemanticity* [11]

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

in mechanistic interpretability of AI research. This stands in tension to the empirical observations of clean selective (*monosemantic*), interpretable neurons [2, 7, 5, 4, 12].

To resolve this contradiction, we introduce a sparse generative model of latent variables that better aligns theoretical predictions with empirical data. This model captures the intuitive notion that natural stimuli are sparse and conceptually distinct. Specifically, exponential priors naturally emerge in sensory coding [13, 14]. This offers a principled explanation for the success of sparse autoencoders in capturing interpretable neural representations [15, 16, 17]. Combined with biologically plausible assumptions such as non-negative neural activations, we prove mathematically and demonstrate empirically that the statistical optimum of neural coding under this model naturally yields individually interpretable neurons aligned directly with latent variables.

Contributions. **1.** We propose a sparse Data Generating Process (DGP) (sec. 3) that addresses key limitations of existing nonlinear independent component models. **2.** We derive identifiability results for both the overcomplete ($m > n'$) and full-capacity ($m = n'$) settings, proving that optimal encoders recover the latent variables up to a linear transformation (Theorem 4.1). Conversely, we prove that in the undercomplete setting ($m < n'$), the global optimum of the cross-entropy loss is generically nonlinear—linear encoders are provably suboptimal (Theorem 4.2). **3.** Integrating identifiability theory with non-negative and energy constraints, we prove that optimal solutions correspond to axis-aligned, monosemantic neurons (Theorem 4.3). **4.** We analyze the discriminative setting with sufficient capacity ($m \geq n'$) but compressed task-relevant measurements ($n' < n$), and show that the learned representation inherits the null space and spark of the true measurement matrix, so that sparse latents are the unique k-sparse solution of a compressed sensing problem (Theorem 4.4); this provides a sufficient-condition mechanism that may help explain the empirical success of sparse autoencoders.**5.** We validate our theoretical predictions empirically, bridging insights from neuroscience, machine learning, and interpretability research.

2 Related Work

Nonlinear ICA and Neuroscience. Recent advances in nonlinear ICA have established conditions under which latent variables can be identified, up to linear transformations, from observed data [18, 10, 19, 20]. These identifiability theories generally predict neural representations exhibiting *mixed selectivity* [21, 8], thereby posing challenges for the direct interpretation of single-neuron responses [4, 2, 5]. By contrast, nonlinear ICA based on temporal autocorrelations has given stronger identifiability results up to permutation equivalences [22, 23]. Consistent with a one-to-one correspondence, disentangled representations have successfully matched neurons in macaque inferior temporal cortex [12], linking theoretical identifiability with biological observations.

Sparse Coding and Sparse Autoencoders (SAEs). Sparse coding and non-negative matrix factorization [14, 24] have long provided frameworks to explain the emergence of selective and interpretable neural representations. However, theoretical extensions beyond linear mixing remain limited. Notably, the uniqueness of sparse solutions, highlighted by compressed sensing theory [25], has been extended to sparse coding scenarios [26]. Additionally, biological plausibility has been reinforced by imposing non-negativity constraints [27, 28, 29]. Our generative-setting analysis is inspired by Whittington et al. [29], who show that non-negativity, fixed variance, and energy minimization make an already-linear representation axis-aligned. We differ in two respects: (i) we *derive* linearity of the classifier-visible representation from cross-entropy training of a possibly nonlinear encoder, including its capacity boundary, rather than assuming a linear mixing; and (ii) we assume class-conditional exponential latents, which are only conditionally independent and have unbounded support, so our proof relies on conditional moments and a fixed within-class variance instead of bounded, symmetric factors. Recent empirical studies utilizing SAEs have demonstrated that sparse decompositions applied to neural representations yield interpretable sparse codes [30, 31, 16, 17, 15, 32, 33, 34]. Relatedly, studies of emotion and other concept representations in large language models report both semantically coherent linear directions and more localized features [35], illustrating the distinction we draw between identifiability up to a linear transformation and the stronger selection of individual axes. These findings underscore the potential of sparsity as a crucial factor driving interpretability in neural representations, providing practical motivation for the theoretical considerations presented here.

Integration and Theoretical Contribution. Our work synthesizes these perspectives, explicitly bridging sparse coding frameworks with nonlinear identifiability theory. By integrating the uniqueness guarantees of non-negative coding into identifiability theory, we resolve conceptual tensions between sparse selectivity and theoretical optimality [36, 3], thereby offering a rigorous theoretical foundation for interpreting neural representations.

3 Data Generating Processes

We assume *latent variables* $z \in \mathcal{Z}$ that are n-dimensional, and class labels $y \in \mathcal{Y}$ over $|\mathcal{Y}|$ classes. The difference between our two DGPs will be the causal direction between these: either *generative* $Y \to Z$ or *discriminative* $Z \to Y$. The latent variables are transformed into *inputs* $x = g(z)$ by an invertible nonlinear function $g : \mathcal{Z} \to \mathcal{X}$. Inputs are transformed into *activations* by a learned nonlinear function $f : \mathcal{X} \to \mathcal{H}$, where the activation space $\mathcal{H}$ is m-dimensional. The composite function $h : \mathcal{Z} \to \mathcal{H}$ is defined as $h = f \circ g$. We assume h is continuous. Crucially, g is fixed while f must be learned.

The learning task in both settings is *classification*. We assume the true relationship between latents and labels is captured by a log-linear model:

$$p(y|z) = \mathrm{softmax}(W^* z) \tag{1}$$

for some matrix $W^* \in \mathbb{R}^{|\mathcal{Y}| \times n}$, where * denotes quantities from the true data generating process. Softmax logits are invariant to a common row shift, so throughout we assume that W^* is row-centered. See Lemma 8.1 for more details. The *effective dimensionality* $n' = \mathrm{rank}(W^*)$ will play a central role in our analysis.

The learned model predicts class probabilities via a linear readout $U \in \mathbb{R}^{|\mathcal{Y}| \times m}$ applied to the activations:

$$q(y|z) = \mathrm{softmax}(Uh(z)). \tag{2}$$

Training minimizes the cross-entropy loss:

$$\mathcal{L}(h, U) = \mathbb{E}_Z \left[\mathrm{CE}\big(p(\cdot \mid Z), q(\cdot \mid Z)\big) \right]. \tag{3}$$

See complete assumption list in Supplement 8.1. We now describe the two causal settings in detail.

3.1 Generative Setting: From Labels to Latents $Y \to Z$

In this setting, we draw a label first, then a latent conditioned on that label. This is akin to the causal process underlying MNIST: a person thinks of a digit, then draws it [37]. Prior work has assumed hyperspherical latents [19, 20, 38], but this is incompatible with sparsity. We propose a different model: non-negative latents with varying norms.

Specifically, we set $\mathcal{Z} = \mathbb{R}^n_+$ and associate each class y with rate parameters $\lambda_y \in \mathbb{R}^n_+$, giving factorized exponential conditionals:

$$p(z|y) = \prod_{i=1}^{n} \lambda_i(y) \exp\big(-\lambda_i(y) z_i\big) = \exp(\langle w_y^*, z \rangle) \prod_{i=1}^{n} \lambda_i(y) \tag{4}$$

where $w_y^* = -\lambda_y < 0$. The conditional entropy is

$$H(p(z|y)) = \sum_{i=1}^{n} \big[1 - \log \lambda_i(y)\big] = n - \sum_{i=1}^{n} \log \lambda_i(y). \tag{5}$$

We assume constant conditional entropy across classes:

$$\prod_{i=1}^{n} \lambda_i(y) = \prod_{i=1}^{n} \lambda_i(\tilde{y}), \quad \forall\, y, \tilde{y} \in \mathcal{Y}. \tag{6}$$

With a uniform prior $p(y) = 1/|\mathcal{Y}|$, Bayes' rule gives the posterior in the form of (1), where $W^* \in \mathbb{R}^{|\mathcal{Y}| \times n}$ has rows $(w_y^*)^\top$. We assume $\mathrm{rank}(W^*) = n$, requiring $|\mathcal{Y}| > n$. Thus, in the generative setting, the effective dimensionality equals the full latent dimension: $n' = n$.

Additional assumptions for axis-alignment. To obtain the stronger result that neurons align with individual latent dimensions (Section 4.3.1), we restrict to the exact-capacity regime $m = n = n'$ and additionally assume: (i) activations are non-negative, $\mathcal{H} = \mathbb{R}^m_+$ (e.g., real neurons or ReLU networks); (ii) a fixed within-class population variance (e.g., variance normalization); and (iii) an energy constraint favors small activations (e.g., via energy optimization or weight decay during training).

3.2 Discriminative Setting: From Latents to Labels $Z \to Y$

In this setting, we draw a latent first from some prior $p(z)$, and then draw a label conditioned on that latent. This models the common image labeling problem—an image, containing some (imperfect) measurement of latent features, is presented to a human observer who labels it [37].

Here we allow more flexibility: $p(z)$ may be any distribution over $\mathcal{Z} = \mathbb{R}^n$. The conditional over labels is log-linear as in (1): $p(y|z) = \mathrm{softmax}(W^* z)_y$.

The key assumption is that W^* factorizes as

$$W^* = U^* A^* \tag{7}$$

where $A^* \in \mathbb{R}^{n' \times n}$ is a *measurement matrix* and $U^* \in \mathbb{R}^{|\mathcal{Y}| \times n'}$ is a *readout matrix*. This means the label depends on z only through the compressed measurement $A^* z \in \mathbb{R}^{n'}$. When $n' < n$, not all latent dimensions are directly relevant for classification—a key difference from the generative setting.

Additional assumptions for compressed sensing recovery. To obtain the stronger result that latents can be fully recovered up to permutation (Section 4.3.2), we additionally assume: (i) latents are *sparse*, $\|z\|_0 \le k$ for all $z \in \mathrm{supp}(p_Z)$; and (ii) A^* satisfies the spark condition $\mathrm{spark}(A^*) > 2k$: no nonzero $2k$-sparse vector lies in $\mathcal{N}(A^*)$. This ensures all k-sparse latents remain distinguishable after projection, making compressed sensing recovery possible [39].

4 Theory

A central question is why neural representations sometimes align with individual latent concepts (*monosemanticity*) and other times represent linear combinations of them (*polysemanticity*). To answer this, we must first address a fundamental prerequisite: *linearity*. Before a system can be axis-aligned, it must first be linearly identifiable—that is, the encoder f must successfully invert the nonlinear mixing g up to a linear transformation. We show that the answer depends critically on the data generating process and the capacity of the neural representation. First, we establish when linear identifiability holds or fails (shared by both settings). Then, we show how to go beyond linear mixing to recover the exact latent variables up to permutations—via two distinct paths depending on the setting. For readers who prefer intuition over algebra, every step ends with a box that restates the key point. For a complete list of assumptions, see Supplement 8.1.

4.1 Linear Identifiability

With supervised training, let encoder $h = f \circ g$ and linear readout U minimize the cross-entropy

$$\mathrm{CE}(h, U) \;:=\; \mathbb{E}_Z\big[\mathrm{CE}\big(p(\cdot \mid Z), \mathrm{softmax}(Uh(Z))\big)\big]. \tag{8}$$

Recall that the true posterior is $p(y|z) = \mathrm{softmax}(W^* z)$ where $n' = \mathrm{rank}(W^*)$ is the effective latent dimension.[1] Building on prior theory [18, 19, 40], we show that when the model has sufficient capacity, the optimal encoder must be linear.

Theorem 4.1 (Linear identifiability)**.** *Let Assumption 8.1 hold with $m \ge n' = \mathrm{rank}(W^*)$. Every global minimizer (U, h) of the cross-entropy satisfies*

$$h(z) = Az + \nu(z), \qquad UA = W^*$$

where $A \in \mathbb{R}^{m \times n}$ has $\mathrm{rank}(A) = n'$, and $\nu(z) \in \mathcal{N}(U)$ for all z.

[1] By the centering lemma (Supplement 8.1), we assume without loss of generality that W^* is row-centered, so rank equals the dimension of the row space.

Proof in Supplement 8.3.

The term $\nu(z)$ captures the degrees of freedom in the $m - n'$ dimensions orthogonal to the readout. These dimensions do not affect the classification loss and may be nonlinear or arbitrary. The classification-relevant part Az is always linear.

TL;DR: Although the network *could* apply complicated nonlinear transformations, the best it can do for supervised classification is recover the latent variables up to linear mixing. Extra dimensions beyond n' are unconstrained by the loss.

4.2 Non-identifiability under Capacity Constraints

What happens when there are not enough neurons? If $m < n'$, under the support and genericity conditions below, no linear encoder can achieve zero KL divergence, so the optimal encoder is nonlinear.

Theorem 4.2 (Non-identifiability)**.** *Let Assumption 8.1 hold with $m < n' = \operatorname{rank}(W^*)$. Assume that $\operatorname{supp}(p_Z)$ has nonempty interior $\Omega \subseteq \mathbb{R}^n$, that p_Z is continuous and strictly positive on Ω, and that for every candidate $A \in \mathbb{R}^{m\times n}$ under consideration, $\mathcal{N}(A) \cap \Omega \neq \emptyset$. Then the global minimizer of the cross-entropy loss is generically nonlinear.*

Proof in Supplement 8.4.

TL;DR: With fewer neurons than the effective latent dimension, it is not possible to recover latents up to a linear transformation. The encoder must distort the geometry.

4.3 Beyond Linear: Axis Alignment and Sparse Recovery

Linear identifiability recovers z up to an arbitrary linear transformation A. Can we do better? We present two paths to stronger results. In Path A, the representation itself is identifiable up to permutation and component-wise scaling, meaning individual neurons correspond to individual latent variables. In Path B, the representation may remain polysemantic, but each sparse latent is uniquely recoverable from it by sparse decoding.

4.3.1 Path A: Axis-Aligned Neurons (Generative Setting)

In the generative setting with $m = n = n'$, we have linear identifiability. If we additionally impose non-negative activations (e.g., ReLU or biological spikes) and an energy constraint (e.g., weight decay or efficient coding), the encoder is a scaled permutation.

Theorem 4.3 (Axis-aligned neurons)**.** *Following Assumption 8.1 and Assumption 8.2, in the generative setting with $m = n = n' = \operatorname{rank}(W^*)$, let (U, h) be any minimizer of $\mathcal{R}(h) = \mathbb{E}_z\|h(Z)\|_2^2$ among the cross-entropy global minimizers. Then*

$$h(z) = DP_\sigma z,$$

where P_σ is a permutation matrix and $D = \operatorname{diag}(\beta_1, \ldots, \beta_n)$ with $\beta_j > 0$. That is, the representation is identifiable up to permutation and component-wise positive scaling.

Proof in Supplement 8.5.1.

TL;DR: With non-negative activations and energy constraints under exact capacity, each neuron encodes exactly one latent variable. This provides a possible theoretical explanation for "grandmother cells."

4.3.2 Path B: Compressed Sensing Recovery (Discriminative Setting)

In the discriminative setting with $m \geq n'$, the encoder learns $h(z) = Az + \nu(z)$. What can we say about the structure of A? If the latents are sparse and the true projection A^* satisfies the spark condition, then A inherits this structure, enabling sparse recovery of z.

Theorem 4.4 (Compressed sensing recovery)**.** *Following Assumption 8.1 and Assumption 8.3, in the discriminative setting with $W^* = U^* A^*$, assume $m \geq n'$, latents are k-sparse, and* $\operatorname{spark}(A^*) > 2k$*. Let P_U denote the orthogonal projector onto* $\operatorname{Row}(U)$ *and define the classifier-visible representation $\widetilde{h}(z) := P_U h(z) = Az$. Then the learned encoder satisfies $\mathcal{N}(A) = \mathcal{N}(A^*)$. Consequently,* $\operatorname{spark}(A) = \operatorname{spark}(A^*) > 2k$*, and z can be recovered from the unique k-sparse solution of*

$$Az = \widetilde{h}(z).$$

Proof in Supplement 8.5.2.

TL;DR: With sparse latents, networks learn the best possible projector. SAEs may then recover all latent variables—explaining their empirical success in mechanistic interpretability.

4.4 Summary

	Generative ($n' = n$)	**Discriminative** ($n' \leq n$)
$m \geq n'$	Linear identifiability	Linear identifiability
$m < n'$	Nonlinear (non-identifiable)	Nonlinear (non-identifiable)
Stronger result via	Non-negativity + energy	Sparsity + spark
Result	Axis-aligned neurons	Unique k-sparse recovery
Interpretation	Grandmother cells	Sparse autoencoders

Experiments in Section 5 corroborate these predictions.

5 Experiments

5.1 Axis-aligned Neurons

We begin with the *full–capacity* setting ($m = n'$), in which our theory predicts that every neuron should align with a single latent dimension when activations are constrained to be non-negative. We compare two DGPs: a) The hyperspherical latent model used in prior work [19, 38, 40]; and b) Our proposed sparse, non-negative exponential latent model (Section 3). For both DGPs we train identical encoders that differ only in the final activation: a *linear* head (no non-linearity) versus a *non-negative* ReLU activation function. Disentanglement is assessed with two metrics. *Linear alignment* measures recovery of the latent space up to an arbitrary linear transform, whereas the stricter *axis alignment* metric evaluates recovery up to permutation and element-wise scaling using the mean correlation coefficient (MCC) [18]. Each experiment is repeated with ten random seeds. All configurations reach near-perfect *linear* alignment (Fig. 1, top), showing that the encoder behaves approximately linearly across settings—consistent with Theorem 4.1. However, axis alignment tells a different story (bottom panels): the combination of a sparse DGP and a non-negative output layer converges (albeit slowly) to an axis-aligned neural representation, whereas both the hyperspherical DGP and the linear output retain mixed-selectivity—consistent with Theorem 4.3. Although convergence is slower for the non-negative encoder, we show in Section 9.1 that weight decay and alternative activation functions substantially accelerate axis alignment.

5.2 Non-identifiability in the undercomplete regime

We next test the *undercomplete* regime ($m < n'$), in which Theorem 4.2 predicts that the global optimum of the cross-entropy loss is generically nonlinear. We fix $n' = 8$ and sweep the encoder dimension $m \in \{2, 4, 6, 8, 10\}$, straddling the theoretical threshold at $m = n'$. Two encoder architectures are compared: linear encoder $h(z) = Az$ and nonlinear encoder *MLP* of the same input-output dimensions. We repeat the experiment across three latent distributions. Each configuration is evaluated over ten random seeds, and we report the final $\mathrm{KL}(p^* \| q)$ as a function of m in Figure 2.

The results reveal a clear transition at $m = n'$ that is consistent across all three distributions. When $m < n'$, the MLP achieves lower KL than the linear encoder across every setting, consistent with the

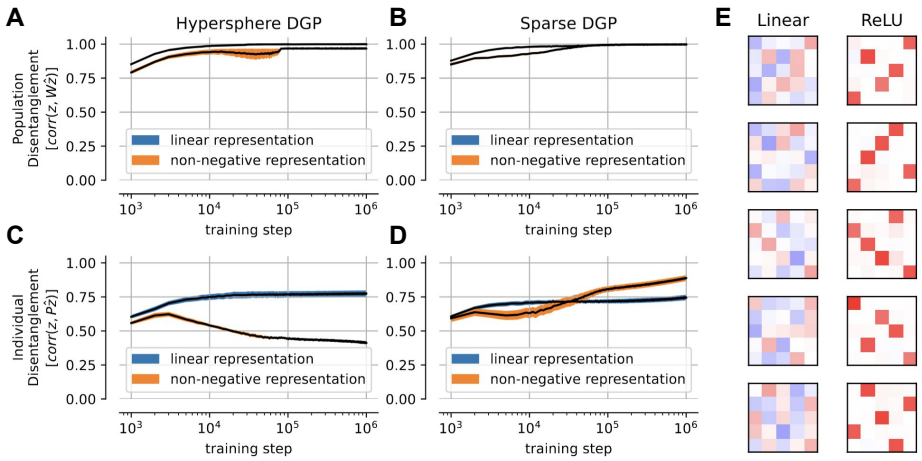

Figure 1: **Linear and axis-aligned latent recovery at exact capacity. A)** Hyperspherical DGP [19, 20], training a neural network with linear representation (i.e., no final activation function, blue) and a non-negative representation (ReLU, orange); Disentanglement between true latent variables (z) and learned neural representations ($\hat{z} = h(z)$), measured up to arbitrary linear transformations (W). **B)** Our proposed sparse, non-negative DGP, same architectures as on the left (with/without activation function), linear disentanglement reported. **C)** Same as A, but disentanglement is measured only up to permutations (P), measuring axis-aligned latent variable recovery. **D)** Same as B but individual, axis-aligned disentanglement measured. **E)** Sample visualizations of five optimization runs, learned linear representation $h(z) = Az$ showing a clear permutation only for the non-negative case. Shaded areas in all plots indicate standard error of the mean over 10 random seeds.

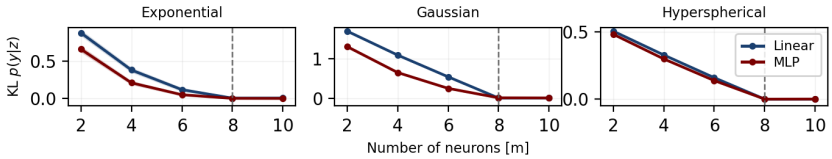

Figure 2: **Phase transition at the capacity threshold.** Final $\mathrm{KL}(p^* \| q)$ is reported as a function of encoder dimension m, for linear (blue) and MLP (red) encoders across three latent distributions. The dashed vertical line marks $m = n'$, the capacity threshold.

suboptimality of linear encoders predicted by Theorem 4.2. Additionally, when $m \geq n'$, both the linear and the MLP encoder converge to near-zero KL divergence. This supports Theorem 4.1 that sufficient capacity ensures the optimal encoder being linear to latent variables, z.

5.3 Compressed Sensing Recovery

We then validate compressed sensing recovery in the discriminative setting, where the labels depend on sparse latents through a low-rank projection $W^* = U^* A^*$. The theory predicts two interlocking results: the learned encoder will precisely recover the null space of A^*, i.e. $\mathcal{N}(A) = \mathcal{N}(A^*)$; and this structural match makes the high-dimensional sparse latent z the unique k-sparse solution given the compressed encoding Az, so that it can be recovered by sparse decoding.

As presented in Figure 3, we construct a discriminative DGP at full capacity $m = n'$ and sweep over three sparsity levels k to probe different regimes while ensuring $\mathrm{spark}(A^*) > 2k$. Three recovery methods are applied to the activation Az: supervised sparse decoding (Lasso), an OLS baseline, and unsupervised dictionary learning with Lasso code inference. Each experiment is repeated with ten random seeds. In Figure 3A, the training dynamics and null-space alignment show that as the KL approaches zero, the row-space alignment $\cos(\theta_{\max})$ between the learned A and the true A^* undergoes an acute transition to 1 (Figure 3A, right), consistent with $\mathcal{N}(A) = \mathcal{N}(A^*)$ at the optimum. The recovery results complete the picture (Figure 3B): supervised Lasso achieves near-perfect R^2

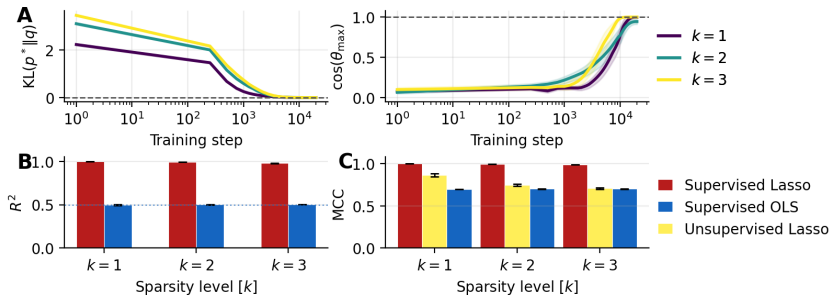

Figure 3: **Compressed Sensing Recovery. A)** Training dynamics for a linear encoder $h(z) = Az$ trained under different sparsity levels. Left: training $\mathrm{KL}(p^*\|q)$. Right: row-space match $\cos(\theta_{\max})$ between the learned A and the true A^*. **B)** R^2 of supervised latent recovery by Lasso and OLS. **C)** MCCs of supervised latent recovery and unsupervised sparse dictionary learning. Shaded areas in **A)** and error bars in **B)** and **C)** indicate standard error of the mean over 10 random seeds.

across all values of k, demonstrating faithful reconstruction of the full n-dimensional latent. By contrast, the OLS baseline plateaus at $R^2 \approx n'/n$, as expected from the rank bound. Also, the MCC of the unsupervised dictionary learning reaches ≈ 0.86 under extreme sparsity (1-sparse) and declines notably as k increases, indicating the significant impact of sparsity on latent variable recovery and offering a possible and principled account of why sparse autoencoders succeed in recovering latent structure from undercomplete neural representations.

5.4 Disentangled Neurons Generalize Across Models and Brains

Our identifiability results predict that, under suitable conditions, individual neurons recover individual latent variables (Theorems 4.3 and 4.4). A direct consequence is that if two systems trained on similar data have each achieved this recovery, well-disentangled units in one system should have a counterpart in the other. We test this prediction across vision models, language models, and macaque cortex. Since true latents are unobservable in real data, we use the Mechanistic Interpretability Score (MIS) of Zimmermann et al. [41] as a proxy for disentanglement—high-MIS neurons exhibit selective, monosemantic responses consistent with having cleanly captured a latent concept. We quantify cross-system preservation via the Maximum Matching Unit (MMU) score: for each neuron, the maximum absolute correlation with any neuron in the comparison system. Our prediction is then simply a positive MIS–MMU correlation.

Vision models. Across four ImageNet-trained networks (VGG-19, ResNet-50, ViT-b/16, DINOv2 ViT-S/14), MIS predicts MMU positively for all twelve directed model pairs (Fig. 4A,C).

Language models. The same pattern, though weaker, holds across GPT-2, Pythia-70m, and TinyLlama, evaluated on the 100,000 most frequent Wikipedia unigrams (Fig. 4B,D).

Biological neurons. In macaque IT cortex [42] (macaque 1 "Chabo": 58 neurons; macaque 2 "Tito": 110 neurons; 3,200 natural images), the directional MIS–MMU correlations are positive. Tito $\rightarrow$ Chabo yields $r = 0.158$ ($p = 0.0986$;Fig. 4E), whereas Chabo $\rightarrow$ Tito yields $r = 0.135$ with less statistical significance ($p = 0.312$), plausibly due to the imbalance candidate pool of MMU.

Across all three settings, the more disentangled a unit, the more likely it is to be found in another system. This is consistent with the theoretical prediction that identifiability drives cross-system convergence on shared latent variables, although the gap between our idealized framework and real neural systems warrants caution.

6 Limitations and Ablations

Our analysis rests on a set of idealized assumptions that rarely hold exactly in real data [20]. We highlight four salient ones. **Constant conditional entropy.** The axis-alignment result (Theorem 4.3) relies on the class-independent entropy condition in eq. (6). Real datasets almost always exhibit some

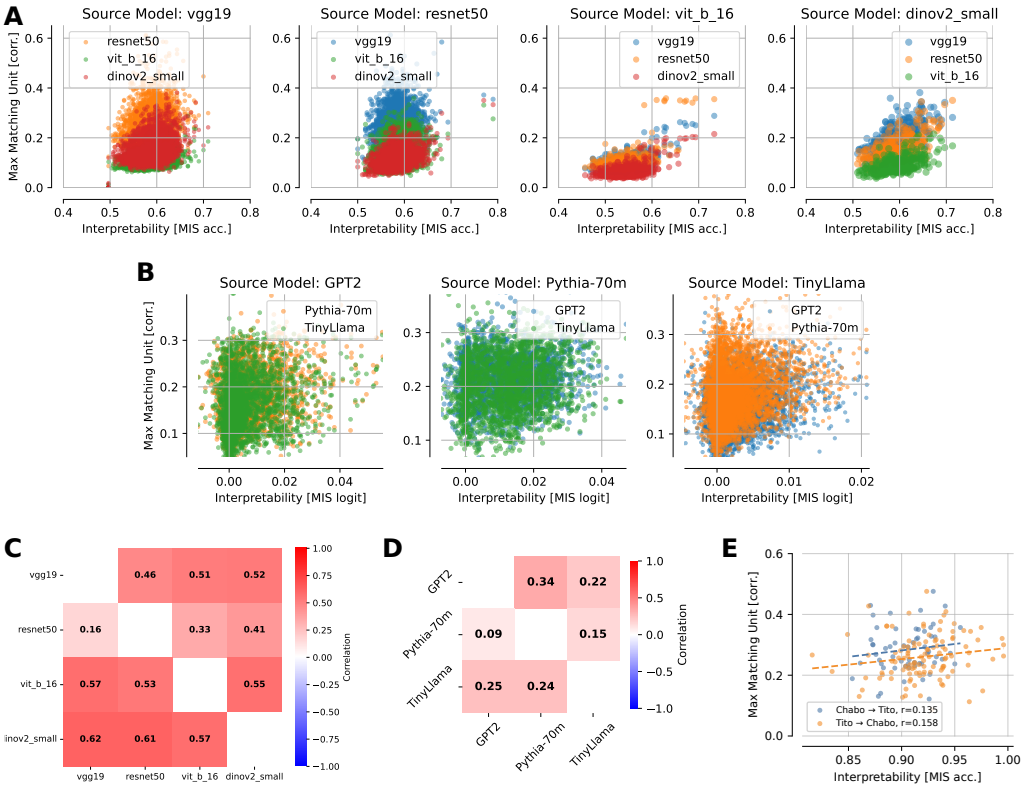

Figure 4: **Disentanglement predicts cross-system preservation. A,B)** Per-neuron interpretability (MIS, Zimmermann et al. 41) versus the Maximum Matching Unit (MMU) score, for ImageNet-trained vision models (A) and LLMs (B; MIS logits reported as accuracy is at ceiling). **C,D)** MIS–MMU correlations for all model pairs in A and B. **E)** Directional MIS–MMU associations in macaque IT cortex [42]. Responses to 3,200 images were recorded from 58 IT neurons in "Chabo" and 110 IT neurons in "Tito", two macaques.

spread in conditional entropies. In preliminary simulations, moderate deviations merely blur axis alignment rather than destroy it, suggesting robustness, but a systematic study is left for future work. **Non-negativity.** Imposing non-negativity is analogous to restricting latents to the positive orthant of a sphere: inner products remain meaningful, but sign information is lost. Alternative sparsity models with Laplace-distributed (*signed*) latents break this symmetry; Theorem 4.3 does not apply directly there. **Uniform class prior.** Assuming $p(y) = 1/|\mathcal{Y}|$ simplifies the inner-product identity. Unequal priors merely add a constant offset $-\log \pi_y$ to each logit row of W^*, and the main results extend verbatim after subtracting row means. Extreme class imbalance could nonetheless interact with optimization in ways not captured here. **DGP.** Finally, the largest gap remains the existence of a *true* sparse DGP [43]–perhaps a philosophical question that no amount of data or theory will settle.

We also conduct ablation tests to assess the robustness of our results to violations of these assumptions. Overall, both linear identifiability and axis alignment are robust to moderate violations of constant conditional entropy, particularly with an affine readout (i.e., a readout with a bias term), and under moderate departures from conditional independence and the exponential conditional distribution. In contrast, violating the non-negative latent-support assumption in the generative setting substantially degrades axis alignment. Full results and experimental details are provided in Supplement 9.1.

7 Discussion

Our framework unifies disparate empirical findings—dedicated, monosemantic units; polysemantic superposition; and sparse-autoencoder recovery—within a single statistical picture by combining nonlinear identifiability theory with two complementary data generating processes. Two ingredients

control the qualitative regime: the *causal direction* of the DGP (labels→latents vs. latents→labels) and the *capacity ratio* between neurons and effective latents (m vs. n').

When capacity is sufficient ($m \geq n'$), the cross-entropy optimum identifies the latents up to an arbitrary linear transformation (Theorem 4.1). Two distinct paths then go beyond this linear ambiguity. In the generative setting, exact capacity, non-negative activations, and an energy constraint yield an axis-aligned optimum in which each neuron encodes a single latent (Theorem 4.3), offering one possible theoretical mechanism for grandmother-cell-like responses. In the discriminative setting, sparsity together with the spark condition on the true measurement matrix endows the encoder with a compressed-sensing geometry, so individual latents can be recovered from polysemantic representations via sparse coding (Theorem 4.4)—an explanation for why sparse autoencoders work. When capacity is insufficient ($m < n'$), no linear encoder is optimal: the bottleneck forces nonlinear distortion and identifiability fails generically (Theorem 4.2).

This perspective also speaks to *efficient coding* [36, 44], which has long struggled to reconcile distributed-coding intuitions with empirically dedicated units such as the frog's fly detector [45] or the mouse's bird detector [46]. The dichotomy is not between efficiency and selectivity but between regimes: under sparse, non-negative latents with sufficient capacity, selectivity *is* the efficient code. Future work should relax the most restrictive assumptions—non-negativity, constant conditional entropy, the uniform class prior—and address signed or manifold-valued latents, narrowing the gap between our idealized framework and real biological and artificial systems.

Code availability. Code for reproducing the experiments and figures in this paper is available at `https://github.com/klindtlab/complete-or-sparse`.

Acknowledgments and Disclosure of Funding

We would like to thank Luc Lehéricy and Katrin Franke for helpful discussions. David Klindt acknowledges the CSHL GPU cluster with assistance from the US National Institutes of Health Grant S10OD028632-01. This work was supported by a Goodfire Research Grant and conducted using Silico, Goodfire's interpretability research agent.

References

[1] Charles G Gross. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518, 2002. doi: 10.1177/1073858402238512. (Cited on page 1.)

[2] Rodrigo Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045): 1102–1107, 2005. (Cited on pages 1, 2, and 2.)

[3] Simon Thorpe. Local vs. Distributed Coding. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, 8(2):3–40, 1989. ISSN 0769-4113. doi: 10.3406/intel.1989.873. URL `https://www.persee.fr/doc/intel_0769-4113_1989_num_8_2_873`. (Cited on pages 1 and 3.)

[4] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020. (Cited on pages 1, 2, and 2.)

[5] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, March 2021. doi: 10.23915/distill.00030. URL `https://distill.pub/2021/multimodal-neurons/`. (Cited on pages 1, 2, and 2.)

[6] Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1934–1943, 2023. (Cited on page 1.)

[7] Meenakshi Khosla, Alex H. Williams, Josh McDermott, and Nancy Kanwisher. Privileged representational axes in biological and artificial neural networks, June 2024. URL `https://www.`

biorxiv.org/content/10.1101/2024.06.20.599957v1. Pages: 2024.06.20.599957 Section: New Results. (Cited on pages 1 and 2.)

[8] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013. (Cited on pages 1, 1, and 2.)

[9] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. (Cited on page 1 and 1.)

[10] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *Journal of Machine Learning Research*, 20(1):2569–2594, 2019. (Cited on pages 1 and 2.)

[11] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The Building Blocks of Interpretability. *Distill*, 3(3):e10, March 2018. ISSN 2476-0757. doi: 10.23915/distill.00010. URL https://distill.pub/2018/building-blocks. (Cited on page 1.)

[12] Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, 12(1):6456, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26751-5. (Cited on page 2 and 2.)

[13] Peter Földiak. Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 64(2):165–170, 1990. (Cited on page 2.)

[14] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. (Cited on page 2 and 2.)

[15] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, and et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity. (Cited on page 2 and 2.)

[16] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, and Amanda Askell. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, page 2, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features. (Cited on page 2 and 2.)

[17] David Klindt, Sophia Sanborn, Francisco Acosta, Frédéric Poitevin, and Nina Miolane. Identifying Interpretable Visual Features in Artificial and Biological Neural Systems, October 2023. URL http://arxiv.org/abs/2310.11431. arXiv:2310.11431 [cs, stat]. (Cited on page 2 and 2.)

[18] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016. (Cited on pages 2, 4, 6, and 24.)

[19] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021. URL https://proceedings.mlr.press/v139/zimmermann21a.html. (Cited on pages 2, 3, 4, 6, 7, and 23.)

[20] Patrik Reizinger, Randall Balestriero, David Klindt, and Wieland Brendel. An empirically grounded identifiability theory will accelerate self-supervised learning research. *arXiv preprint arXiv:2504.13101*, 2025. (Cited on pages 2, 3, 7, 8, and 23.)

[21] Omri Barak, Mattia Rigotti, and Stefano Fusi. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *Journal of Neuroscience*, 33(9):3844–3856, 2013. URL https://www.jneurosci.org/content/33/9/3844.short. Publisher: Soc Neuroscience. (Cited on page 2.)

[22] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pages 460–469. PMLR, 2017. (Cited on page 2.)

[23] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding, March 2021. URL http://arxiv.org/abs/2007.10930. arXiv:2007.10930 [cs, stat]. (Cited on page 2.)

[24] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. (Cited on page 2.)

[25] David L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. URL https://ieeexplore.ieee.org/abstract/document/1614066/?casa_token=vtpGjU5mzFcAAAAA:rU2N5NCWY2K9IaaU0GHdJEuOj8P0dFk39KnF-rchFhrMrAe9T0XiWvCPGgJ5pszVR4-UWxvhvg. Publisher: IEEE. (Cited on page 2.)

[26] Christopher J Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61(11):6290–6297, 2015. (Cited on page 2.)

[27] Cengiz Pehlevan, Sreyas Mohan, and Dmitri B Chklovskii. Blind nonnegative source separation using biological neural networks. *Neural computation*, 29(11):2925–2954, 2017. (Cited on page 2.)

[28] David Lipshutz, Cengiz Pehlevan, and Dmitri B Chklovskii. Biologically plausible single-layer networks for nonnegative independent component analysis. *Biological cybernetics*, 116(5): 557–568, 2022. (Cited on page 2.)

[29] James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy EJ Behrens. Disentanglement with biological constraints: A theory of functional cell types. *arXiv preprint arXiv:2210.01768*, 2022. (Cited on pages 2, 2, and 24.)

[30] Alona Fyshe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. Interpretable Semantic Vectors from a Joint Model of Brain- and Text- Based Meaning. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 489–499, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1046. URL https://aclanthology.org/P14-1046. (Cited on page 2.)

[31] Zeyu Yun, Yubei Chen, Bruno A. Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors, 2021. URL http://arxiv.org/abs/2103.15949. arXiv:2103.15949 [cs]. (Cited on page 2.)

[32] Charles O'Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amortisation gap in sparse autoencoders. *arXiv preprint arXiv:2411.13117*, 2024. (Cited on page 2.)

[33] Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pages 2024–11, 2024. (Cited on page 2.)

[34] David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. From superposition to sparse codes: interpretable representations in neural networks. *arXiv preprint arXiv:2503.01824*, 2025. (Cited on page 2.)

[35] Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026. (Cited on page 2.)

[36] H. B. Barlow. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1(4):371–394, 1972. ISSN 0301-0066. doi: 10.1068/p010371. (Cited on pages 3 and 10.)

[37] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012. (Cited on pages 3 and 4.)

[38] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143*, 2024. (Cited on pages 3 and 6.)

[39] Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9(1):92, 2010. (Cited on page 4.)

[40] Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. *arXiv preprint arXiv:2410.21869*, 2024. (Cited on pages 4, 6, and 23.)

[41] Roland Zimmermann, David Klindt, and Wieland Brendel. Measuring Mechanistic Interpretability at Scale Without Humans. *NeurIPS*, 2024. URL `https://openreview.net/forum?id=M8yBcvRvwn`. (Cited on pages 8, 9, and 28.)

[42] Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015. (Cited on pages 8, 9, 28, and 31.)

[43] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. (Cited on page 9.)

[44] Christian K Machens, Tim Gollisch, Olga Kolesnikova, and Andreas VM Herz. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47(3):447–456, 2005. (Cited on page 10.)

[45] Jerome Y Lettvin, Humberto R Maturana, Warren S McCulloch, and Walter H Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951, 2007. (Cited on page 10.)

[46] Yifeng Zhang, In-Jung Kim, Joshua R Sanes, and Markus Meister. The most numerous ganglion cell type of the mouse retina is a selective feature detector. *Proceedings of the National Academy of Sciences*, 109(36):E2391–E2398, 2012. (Cited on page 10.)

Supplement

8 Proofs

8.1 Assumptions

We delineate the assumptions for our theoretical framework in three parts: General assumptions shared across all settings, assumptions specific to the Generative setting, and assumptions specific to the Discriminative setting.

Assumption 8.1 (General Assumptions)**.**

These assumptions hold for both the Generative and Discriminative settings.

(G1) ***Latent-to-Data Map:*** *The data generation process is given by $x = g(z)$, where $g : \mathcal{Z} \to \mathcal{X}$ is a fixed, continuous, and invertible function. This ensures that the observation x contains all information about the latent variables z.*

(G2) ***Log-linear True Posterior:*** *The true relationship between latents and labels is $p(y|z) = \mathrm{softmax}(W^* z)$ for some matrix $W^* \in \mathbb{R}^{|\mathcal{Y}| \times n}$. The effective dimensionality is $n' = \mathrm{rank}(W^*)$.*

(G3) ***Learned Model Structure:*** *The learner optimizes a continuous encoder $f : \mathcal{X} \to \mathcal{H}$ with hidden dimensionality $\dim(\mathcal{H}) = m$ and a linear readout $U \in \mathbb{R}^{|\mathcal{Y}| \times m}$. The composite representation is $h = f \circ g : \mathcal{Z} \to \mathcal{H}$. The model's predictive distribution is $q(y|z) = \mathrm{softmax}(Uh(z))$.*

(G4) ***Global Optimality:*** *We assume the training procedure finds a global minimizer of the population cross-entropy loss: $\mathcal{L}(h, U) = \mathbb{E}_{Z,Y}[\mathrm{CE}(p(Y|Z), q(Y|Z))]$.*

8.2 General Lemmas

Lemma 8.1 (Centering Invariance)**.** *Let $\bar{W} = W^* - \mathbf{1}\bar{w}^\top$ where $\bar{w} = \frac{1}{|\mathcal{Y}|}\sum_y w_y^*$ is the mean over rows. Then $\mathrm{softmax}(W^* z) = \mathrm{softmax}(\bar{W} z)$ for all z. Consequently, all identifiability results depend only on the centered matrix $\bar{W}$, and we assume without loss of generality that W^* is row-centered.*

Proof. For any constant vector $c \in \mathbb{R}^{|\mathcal{Y}|}$, $\mathrm{softmax}(v + c) = \mathrm{softmax}(v)$ since the shift cancels in the ratio. Setting $c = -(\bar{w}^\top z)\mathbf{1}$, we have $\mathrm{softmax}(W^* z) = \mathrm{softmax}(W^* z - (\bar{w}^\top z)\mathbf{1}) = \mathrm{softmax}(\bar{W} z)$. □

Remark 1. *The shift-invariance of the softmax function means that the true posterior $p(y|z) = \mathrm{softmax}(W^* z)$ is unchanged if we add any constant vector to all rows of W^*. What matters for identifiability is not W^* itself, but rather the space spanned by row differences $\{w_y^* - w_{\tilde{y}}^*\}_{y,\tilde{y} \in \mathcal{Y}}$. However, working with row differences directly is cumbersome. Centering W^* gives us a convenient equivalence: for a row-centered matrix, $\mathrm{rank}(W^*) = \dim(\mathrm{Span}\{w_y^* - w_{\tilde{y}}^*\})$. This allows us to state identifiability results in terms of the simpler quantity $\mathrm{rank}(W^*)$.*

Lemma 8.2 (Log-Ratio Identity)**.** *Assume the true posterior is $p(y|z) = \mathrm{softmax}(W^* z)$ and the learned model is $q(y|z) = \mathrm{softmax}(Uh(z))$. If $p(\cdot \mid z) = q(\cdot \mid z)$ at a given z, then for any pair of classes $y, \tilde{y} \in \mathcal{Y}$:*

$$\langle u_y - u_{\tilde{y}}, h(z) \rangle = \langle w_y^* - w_{\tilde{y}}^*, z \rangle,$$

where u_y and w_y^* are the y-th rows of U and W^* respectively.

Proof. For the given z, $p(y|z) = q(y|z)$ for all y. Writing out the softmax probabilities:

$$\frac{e^{\langle w_y^*, z \rangle}}{\sum_k e^{\langle w_k^*, z \rangle}} = \frac{e^{\langle u_y, h(z) \rangle}}{\sum_k e^{\langle u_k, h(z) \rangle}}.$$

Taking the log-ratio of probabilities for classes y and $\tilde{y}$ eliminates the partition functions:

$$\log \frac{p(y|z)}{p(\tilde{y}|z)} = \log \frac{q(y|z)}{q(\tilde{y}|z)}$$
$$\Rightarrow \langle w_y^* - w_{\tilde{y}}^*, z \rangle = \langle u_y - u_{\tilde{y}}, h(z) \rangle.$$

Consequently, the identity holds throughout the support whenever the two continuous posteriors agree p_Z-almost surely. □

Lemma 8.3 (Linearity from Inner Products)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^n$ be a domain spanning $\mathbb{R}^n$, and let $h : \mathcal{Z} \to \mathbb{R}^m$ be a continuous function with $m \geq n'$. Suppose there exist two families of vectors $\mathcal{B} := \{b_k\}_{k \in \mathcal{K}} \subset \mathbb{R}^n$ and $\mathcal{C} := \{c_k\}_{k \in \mathcal{K}} \subset \mathbb{R}^m$ indexed by a set $\mathcal{K}$, satisfying:*

$$\langle c_k, h(z) \rangle = \langle b_k, z \rangle, \quad \forall z \in \mathcal{Z}, \forall k \in \mathcal{K}. \tag{9}$$

If $\mathrm{Span}(\mathcal{B})$ has dimension n', then there exists a matrix $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = n', \mathrm{Im}(A) \subseteq \mathrm{Span}(\mathcal{C})$ such that:

$$h(z) = Az + \nu(z)$$

where $\nu(z) \in \mathrm{Span}(\mathcal{C})^\perp$ for all $z \in \mathcal{Z}$.

Proof.

Step 1: Dimension bound. We first show $\dim(\mathrm{Span}(\mathcal{C})) \geq n'$. Suppose $\sum_k \alpha_k c_k = 0$ for some coefficients $\{\alpha_k\}$. Taking the inner product with $h(z)$ and applying (9):

$$0 = \Big\langle \sum_k \alpha_k c_k, h(z) \Big\rangle = \sum_k \alpha_k \langle b_k, z \rangle = \Big\langle \sum_k \alpha_k b_k, z \Big\rangle.$$

Since $\mathcal{Z}$ spans $\mathbb{R}^n$, this implies $\sum_k \alpha_k b_k = 0$. Thus, every linear dependence among $\mathcal{C}$ induces a linear dependence among $\mathcal{B}$. This implies $\dim(\mathrm{Span}(\mathcal{C})) \geq \dim(\mathrm{Span}(\mathcal{B})) = n'$.

Step 2: Construction of A. Select indices $\mathcal{I} \subseteq \mathcal{K}$ such that $\{c_i\}_{i \in \mathcal{I}}$ forms a basis for $\mathrm{Span}(\mathcal{C})$. Let C be the matrix with rows $c_i^\top$ for $i \in \mathcal{I}$, and let B be the matrix with corresponding rows $b_i^\top$. The inner product conditions give $Ch(z) = Bz$ for all $z \in \mathcal{Z}$.

Since the rows of C are linearly independent by construction, $CC^\top$ is invertible. Define:

$$A := C^\top (CC^\top)^{-1} B \in \mathbb{R}^{m \times n}.$$

Let $P_C = C^\top (CC^\top)^{-1} C$ be the orthogonal projection onto $\mathrm{Span}(\mathcal{C})$. We decompose:

$$h(z) = P_C h(z) + (I - P_C)h(z) = C^\top (CC^\top)^{-1} Ch(z) + (I - P_C)h(z).$$

Substituting $Ch(z) = Bz$:

$$h(z) = Az + \nu(z), \quad \text{where } \nu(z) := (I - P_C)h(z).$$

By construction, $\mathrm{Im}(A) \subseteq \mathrm{Span}(\mathcal{C})$ and $\nu(z) \in \mathrm{Span}(\mathcal{C})^\perp$.

Step 3: Rank of A. Observe that $CA = CC^\top (CC^\top)^{-1} B = B$. We now show that $\mathrm{rank}(B) = n'$. Since $\{c_i\}_{i \in \mathcal{I}}$ is a basis for $\mathrm{Span}(\mathcal{C})$, for every $c_k \in \mathcal{C}$, there exists a sequence $\{\beta_i\}_{i \in \mathcal{I}}$ such that $c_k = \sum_{i \in \mathcal{I}} \beta_i c_i$. By Equation (9), we have for every $z \in \mathcal{Z}$,

$$\langle b_k, z \rangle = \langle c_k, h(z) \rangle = \Big\langle \sum_{i \in \mathcal{I}} \beta_i c_i, h(z) \Big\rangle = \sum_{i \in \mathcal{I}} \beta_i \langle c_i, h(z) \rangle = \sum_{i \in \mathcal{I}} \beta_i \langle b_i, z \rangle = \Big\langle \sum_{i \in \mathcal{I}} \beta_i b_i, z \Big\rangle.$$

Hence for all $z \in \mathcal{Z}$, $\Big\langle b_k - \sum_{i \in \mathcal{I}} \beta_i b_i, z \Big\rangle = 0$. Again, as $\mathcal{Z}$ spans $\mathbb{R}^n$, we have $b_k = \sum_{i \in \mathcal{I}} \beta_i b_i$. That is, $\forall k \in \mathcal{K}$, b_k is a linear combination of $\{b_i\}_{i \in \mathcal{I}}$, i.e.

$$\mathrm{Span}(\mathcal{B}) = \mathrm{Span}(\{b_i\}_{i \in \mathcal{I}}) = \mathrm{Row}(B).$$

Thus, $\mathrm{rank}(B) = \dim(\mathrm{Span}(\mathcal{B})) = n'$. For the rank of A: since A factors through B, we have $\mathrm{rank}(A) \leq \mathrm{rank}(B) = n'$. Conversely, since C has full row rank, $\mathrm{rank}(B) = \mathrm{rank}(CA) \leq \mathrm{rank}(A)$. Thus $\mathrm{rank}(A) = n'$. □

8.3 Linear Identifiability

Theorem (Linear identifiability)**.** *Let Assumption 8.1 hold with $m \geq n' = \operatorname{rank}(W^*)$. Every global minimizer (U, h) of the cross-entropy satisfies*

$$h(z) = Az + \nu(z), \qquad UA = W^*$$

where $A \in \mathbb{R}^{m \times n}$ has $\operatorname{rank}(A) = n'$, and $\nu(z) \in \mathcal{N}(U)$ for all z.[2]

Remark 2. *The component $\nu(z)$ lies in the null space of the readout matrix U and is therefore unconstrained by the classification loss; it may be nonlinear.*

Proof. By Lemma 8.1, assume without loss of generality that W^* is row-centered, so $n' = \operatorname{rank}(W^*)$.

1. Establishing the inner product identity. The decomposition of the cross-entropy loss in equation (3) gives

$$\mathcal{L}(h, U) = H(Y \mid Z) + \mathbb{E}_Z D_{\mathrm{KL}}(p(\cdot \mid Z) \,\|\, q(\cdot \mid Z)) \,.$$

Because $m \geq n'$, take a rank factorization $W^* = \widetilde{U}\widetilde{A}$ with inner dimension n' and pad it with zeros to obtain $U_0 \in \mathbb{R}^{|\mathcal{Y}| \times m}$ and $A_0 \in \mathbb{R}^{m \times n}$ satisfying $U_0 A_0 = W^*$. By Item G1, the continuous map $h_0(z) = A_0 z$ is realizable by an admissible encoder. Thus $q_0 = p$ is attained. Every global minimizer consequently has zero expected KL divergence, hence $p(\cdot \mid z) = q(\cdot \mid z)$ for p_Z-almost every z.

We then invoke the Log-Ratio Identity (Lemma 8.2). For any pair of classes $y, \tilde{y} \in \mathcal{Y}$:

$$\langle u_y - u_{\tilde{y}}, h(z) \rangle = \langle w_y^* - w_{\tilde{y}}^*, z \rangle, \quad \forall z \in \mathcal{Z}. \tag{10}$$

2. Linearity of the encoder. Define the families $\mathcal{B} := \{w_y^* - w_{\tilde{y}}^*\}_{y,\tilde{y} \in \mathcal{Y}}$ and $\mathcal{C} := \{u_y - u_{\tilde{y}}\}_{y,\tilde{y} \in \mathcal{Y}}$, with the natural indexing $k = (y, \tilde{y})$.

Since W^* is centered with rank n', the difference vectors $\mathcal{B}$ span an n'-dimensional subspace of $\mathbb{R}^n$. The identity (10) gives $\langle c_k, h(z) \rangle = \langle b_k, z \rangle$ for all $z \in \mathcal{Z}$ and all k.

Applying Lemma 8.3 yields:

$$h(z) = Az + \nu(z)$$

where $A \in \mathbb{R}^{m \times n}$ has $\operatorname{rank}(A) = n'$ and $\nu(z) \in \operatorname{Span}(\mathcal{C})^{\perp}$.

3. The residual lies in the null space of U. Since $\mathcal{C} = \{u_y - u_{\tilde{y}}\} \subseteq \operatorname{rowspace}(U)$, we have $\operatorname{Span}(\mathcal{C})^{\perp} \supseteq \mathcal{N}(U)$.

To show equality on $\nu(z)$: from $h(z) = Az + \nu(z)$ and $\nu(z) \perp \operatorname{Span}(\mathcal{C})$, we have for any class pair $(y, \tilde{y})$:

$$\langle u_y - u_{\tilde{y}}, \nu(z) \rangle = 0.$$

This means $(U\nu(z))_y - (U\nu(z))_{\tilde{y}} = 0$ for all $y, \tilde{y}$, so $U\nu(z)$ is a constant vector. Since $U\nu(z) \in \operatorname{rowspace}(U^\top)$ and constant vectors lie in $\operatorname{rowspace}(U^\top)$ only if they are zero (by the centering of U at optimality), we have $U\nu(z) = 0$, i.e., $\nu(z) \in \mathcal{N}(U)$. □

8.4 Non-identifiability

Theorem (Non-identifiability)**.** *Let Assumption 8.1 hold with $m < n' = \operatorname{rank}(W^*)$. Assume that $\operatorname{supp}(p_Z)$ has nonempty interior $\Omega \subseteq \mathbb{R}^n$, that p_Z is continuous and strictly positive on Ω, and that for every candidate $A \in \mathbb{R}^{m \times n}$ under consideration, $\mathcal{N}(A) \cap \Omega \neq \emptyset$. Then the global minimizer of the cross-entropy loss is generically nonlinear.*

Remark 3. *The support condition is automatic in the discriminative setting where p_Z has full support on $\mathbb{R}^n$. In the generative setting where $\operatorname{supp}(p_Z) = \mathbb{R}^n_+$, it amounts to assuming $\mathcal{N}(A)$ is not entirely confined to the boundary $\partial\mathbb{R}^n_+$. This holds for generic A and is needed only to make the variational argument below well-posed.*

[2] This does not mean that f is linear, but that it inverts *g up to a linear transform*: $h(z) = f(g(z)) = Az + \nu(z)$.

Proof. By Lemma 8.1, assume WLOG that W^* is row-centered, so $n' = \operatorname{rank}(W^*)$. We proceed by contradiction: assume some linear encoder $h^*(z) = Az$ with readout $U^* \in \mathbb{R}^{|\mathcal{Y}| \times m}$ achieves the global minimum of $\mathcal{L}(h, U) = \mathbb{E}_Z[\mathrm{CE}(p(\cdot|Z), q_{h(Z)})]$, where $q_v := \operatorname{softmax}(U^* v)$. Since $m < n'$, $\dim \mathcal{N}(A) \geq n - m \geq n - n' + 1 \geq 1$.

For brevity, write $\ell(v; z) := \mathrm{CE}(p(\cdot|z), \operatorname{softmax}(U^* v))$, so that $\mathcal{L}(h^*, U^*) = \mathbb{E}_Z[\ell(h^*(Z); Z)]$. The standard softmax-plus-cross-entropy gradient is

$$\nabla_v \ell(v; z) = (U^*)^\mathsf{T} \big(\operatorname{softmax}(U^* v) - p(\cdot|z) \big). \tag{11}$$

1. Pointwise stationarity on Ω. We first show that at the global minimum,

$$\nabla_v \ell|_{v=h^*(z)} = 0, \quad \forall z \in \Omega. \tag{12}$$

This is the pointwise consequence of global optimality, by the fundamental lemma of the calculus of variations.

Suppose for contradiction that there exists $z_0 \in \Omega$ with $\gamma_0 := \nabla_v \ell|_{v=h^*(z_0)} \neq 0$. Since h^* is continuous and $\nabla_v \ell$ is continuous in both arguments, $z \mapsto \nabla_v \ell|_{v=h^*(z)}$ is continuous. So there exists an open ball $B_\delta(z_0) \subset \Omega$ such that

$$\langle \nabla_v \ell|_{v=h^*(z)}, \gamma_0 \rangle \geq \tfrac{1}{2} \|\gamma_0\|^2 > 0, \quad \forall z \in B_\delta(z_0).$$

Let $\phi : \mathbb{R}^n \to [0, 1]$ be a smooth bump function supported in $B_\delta(z_0)$ with $\phi(z_0) = 1$. For $\epsilon > 0$, define the continuous perturbation

$$\tilde{h}(z) := h^*(z) - \epsilon\, \phi(z)\, \gamma_0.$$

By Taylor's theorem applied to $\ell(\cdot; z)$, which is smooth in its first argument:

$$\begin{aligned} \mathcal{L}(\tilde{h}, U^*) - \mathcal{L}(h^*, U^*) &= -\epsilon \int_{B_\delta(z_0)} \phi(z) \langle \nabla_v \ell|_{h^*(z)}, \gamma_0 \rangle\, p_Z(z)\, dz + O(\epsilon^2) \\ &\leq -\tfrac{\epsilon}{2} \|\gamma_0\|^2 \int_{B_\delta(z_0)} \phi(z)\, p_Z(z)\, dz + O(\epsilon^2). \end{aligned}$$

Since $p_Z > 0$ on $\Omega \supset B_\delta(z_0)$ and ϕ is positive on a set of positive measure, the integral is strictly positive. For ϵ small enough, $\mathcal{L}(\tilde{h}, U^*) < \mathcal{L}(h^*, U^*)$, contradicting global optimality. Hence (12) holds.

TL;DR. Global minimization in expectation forces *pointwise* stationarity in h on the interior of $\operatorname{supp}(p_Z)$. A localized bump perturbation in any direction with nonzero gradient would strictly decrease the integral.

2. Restriction to $\mathcal{N}(A)$. For $z \in \mathcal{N}(A) \cap \Omega$, $h^*(z) = Az = 0$, so $q_{h^*(z)} = \operatorname{softmax}(0) = \frac{1}{|\mathcal{Y}|}\mathbf{1}$. Substituting into (11) and (12):

$$(U^*)^\mathsf{T} \Big(\tfrac{1}{|\mathcal{Y}|}\mathbf{1} - p(\cdot|z) \Big) = 0, \quad \forall z \in \mathcal{N}(A) \cap \Omega.$$

Let u_y denote the rows of U^* and $\bar{u} := \frac{1}{|\mathcal{Y}|} \sum_y u_y$ their mean. The condition rearranges to

$$\sum_y p(y|z)\, u_y = \bar{u}, \quad \forall z \in \mathcal{N}(A) \cap \Omega.$$

Substituting $p(y|z) = e^{\langle w_y^*, z \rangle} / Z(z)$ and clearing the partition function $Z(z) > 0$:

$$\sum_y (u_y - \bar{u})\, e^{\langle w_y^*, z \rangle} = 0, \quad \forall z \in \mathcal{N}(A) \cap \Omega. \tag{13}$$

TL;DR. On $\mathcal{N}(A)$, the encoder collapses to zero, so the predicted distribution is uniform regardless of U^*. Stationarity then balances U^* against the *true* posterior on this subspace.

3. Linear independence of exponentials forces U^* rank-1. Since $\mathcal{N}(A) \cap \Omega$ is open in the subspace $\mathcal{N}(A)$ (intersection of a subspace with an open set in $\mathbb{R}^n$), and $\dim \mathcal{N}(A) \geq 1$, we may pick $z_* \in \mathcal{N}(A) \cap \Omega$ together with a direction $\xi \in \mathcal{N}(A)$ such that:

(i) $\xi \neq 0$ and the line segment $\{z_* + t\xi : t \in I\} \subset \mathcal{N}(A) \cap \Omega$ for some open interval $I \ni 0$ (possible since $\mathcal{N}(A) \cap \Omega$ is open in $\mathcal{N}(A)$);

(ii) the scalars $\lambda_y := \langle w_y^*, \xi \rangle$ are distinct across y.

Condition (ii) holds for almost every $\xi \in \mathcal{N}(A)$: it fails only on a finite union of hyperplanes within $\mathcal{N}(A)$, namely $\{\xi : \langle w_y^* - w_{\tilde{y}}^*, \xi \rangle = 0\}$ for $y \neq \tilde{y}$. (The "generic" qualifier in the theorem statement covers the degenerate cases where W^* has repeated rows or where $\mathcal{N}(A)$ lies entirely in the row-difference annihilator of W^*. The latter cannot occur when $m < n'$ for generic W^*.)

Substituting $z = z_* + t\xi$ into (13) and defining $c_y := e^{\langle w_y^*, z_* \rangle}(u_y - \bar{u}) \in \mathbb{R}^m$:

$$\sum_y c_y \, e^{\lambda_y t} = 0, \quad \forall t \in I.$$

Projecting onto each coordinate $k = 1, \ldots, m$ gives a scalar identity

$$\sum_y (c_y)_k \, e^{\lambda_y t} = 0, \quad \forall t \in I.$$

Since the exponentials $\{e^{\lambda_y t}\}_y$ with distinct λ_y are linearly independent in $C(I)$, all $(c_y)_k$ vanish. Hence $c_y = 0$ for every y, and since $e^{\langle w_y^*, z_* \rangle} > 0$:

$$u_y = \bar{u}, \quad \forall y \in \mathcal{Y}. \tag{14}$$

TL;DR. Sliding z along a one-parameter ray in $\mathcal{N}(A) \cap \Omega$ turns the constraint into a vanishing exponential sum with distinct exponents. Linear independence forces every coefficient to vanish, collapsing U^* to rank one.

4. Contradiction via strict improvement. By (14), $U^* = \mathbf{1}\,\bar{u}^\mathsf{T}$, so for every z:

$$U^* h^*(z) = \mathbf{1}(\bar{u}^\mathsf{T} A z) = (\bar{u}^\mathsf{T} A z)\mathbf{1},$$

which is a constant vector in y. Consequently, $q_{h^*(z)} = \mathrm{softmax}((\bar{u}^\mathsf{T} A z)\mathbf{1}) = \frac{1}{|\mathcal{Y}|}\mathbf{1}$ for all z, and the loss is

$$\mathcal{L}(h^*, U^*) = \mathbb{E}_z \left[-\sum_y p(y|z) \log \tfrac{1}{|\mathcal{Y}|} \right] = \log |\mathcal{Y}|.$$

To derive a contradiction, we exhibit a linear pair $(\tilde{A}, \tilde{U})$ with strictly smaller loss. Define the *deviation-expectation matrix* $\bar{C} \in \mathbb{R}^{|\mathcal{Y}| \times n}$ as follows:

$$\begin{aligned} \bar{C}_{y,j} &= \mathbb{E}_z \left[(p(y|z) - \frac{1}{|\mathcal{Y}|}) z_j \right] \\ &= \mathbb{E}_z \left[p(y|z) z_j \right] - \frac{1}{|\mathcal{Y}|} \mathbb{E}_z z_j \end{aligned}$$

As $\sum_y \bar{C}_{y,j} = \mathbb{E}_z \left[\sum_y p(y|z) z_j \right] - \sum_y \frac{1}{|\mathcal{Y}|} \mathbb{E}_z z_j = \mathbb{E}_z z_j - \mathbb{E}_z z_j = 0$, we have $\bar{C}$ is row centralized. Let $\bar{C} = P \Sigma Q^\top$ be the SVD decomposition of $\bar{C}$. By Eckart–Young–Mirsky theorem, the best m rank approximation of $\bar{C}$ is $\bar{C}_m = P_m \Sigma_m Q_m^\top$. Then we define (A_ϵ, U_ϵ) as:

$$U_\epsilon = \frac{\sqrt{\epsilon}}{\|\bar{C}_m\|_F^{1/2}} P_m \Sigma_m^{1/2}$$

$$A_\epsilon = \frac{\sqrt{\epsilon}}{\|\bar{C}_m\|_F^{1/2}} \Sigma_m^{1/2} Q_m^\top.$$

Denote $V = \bar{C}_m/\|\bar{C}_m\|_F$, thus $\epsilon V = U_\epsilon A_\epsilon$. As $\mathbf{1}^\top \bar{C} = 0 \Leftrightarrow \mathbf{1}^\top P_{[:,j]} = 0$, it directly gives $\mathbf{1}^\top V = 0$. Now we write out the loss function for (A_ϵ, U_ϵ).

$$\begin{aligned} f(\epsilon) &= \mathcal{L}(A_\epsilon, U_\epsilon) \\ &= \mathbb{E}_z \left[-\sum_y p(y|z) \log \operatorname{softmax}(\epsilon V z)_y \right] \\ &= \mathbb{E}_z \left[-\epsilon \sum_y p(y|z)(Vz)_y + \log \sum_y e^{\epsilon (Vz)_y} \right] \end{aligned}$$

Here, notice that when $\epsilon = 0$, we have $q(y|z) = 1/|\mathcal{Y}|$ uniform and $\mathcal{L}(U_0, A_0) = \log |\mathcal{Y}|$. Therefore, to prove contradiction, we need only prove $f'(0) < 0$ equivalently.

$$\begin{aligned} f'(0) &= \frac{\partial}{\partial \epsilon} f \Big|_{\epsilon=0} \\ &= \mathbb{E}_z \left[-\sum_y p(y|z)(Vz)_y + \frac{\sum_y (Vz)_y e^{\epsilon (Vz)_y}}{\sum_y e^{\epsilon (Vz)_y}} \right] \Bigg|_{\epsilon=0} \\ &= \mathbb{E}_z \left[-\sum_y p(y|z)(Vz)_y + \frac{1}{|\mathcal{Y}|} \mathbf{1}^\top V z \right] \\ &= \mathbb{E}_z \left[-\sum_y p(y|z)(Vz)_y \right] \\ &= -\sum_{y,j} V_{y,j} \mathbb{E}_z \left[p(y|z) z_j \right] \\ &= -\sum_{y,j} V_{y,j} \left(\mathbb{E}_z \left[p(y|z) z_j \right] - \frac{1}{|\mathcal{Y}|} \mathbb{E}_z z_j \right) - \sum_{y,j} V_{y,j} \cdot \frac{1}{|\mathcal{Y}|} \mathbb{E}_z z_j \\ &= -\langle V, \bar{C} \rangle_F - \frac{1}{|\mathcal{Y}|} \sum_j \mathbb{E}_z z_j \sum_y V_{y,j} \\ &= -\|\bar{C}_m\|_F < 0 \end{aligned}$$

TL;DR. Once U^* collapses to rank one, the network outputs uniform predictions and incurs the entropy bound $\log |\mathcal{Y}|$. Any linear encoder that "shifts" the latent variable z in the direction of the difference between the true distribution and the uniform distribution, i.e. $p(y|z) - 1/|\mathcal{Y}|$, will achieve strictly lower loss.

This completes the contradiction, so no linear encoder achieves the global minimum.

□

8.5 Permutation Identifiability

8.5.1 Axis-Aligned Neurons (Generative Setting)

Additional Assumptions

Assumption 8.2 (Generative Setting Assumptions)**.**

These assumptions apply to the generative setting ($Y \to Z$) where we seek axis alignment.

(GEN1) ***Non-negative Assumptions:*** *The latent space is $\mathcal{Z} = \mathbb{R}^n_+$. The activation space is $\mathcal{H} = \mathbb{R}^m_+$, e.g., from ReLU networks.*

(GEN2) ***Exponential Conditionals:*** *The true conditional $p(z|y)$ follows a factorized exponential family:*

$$p(z|y) = \prod_{i=1}^{n} \lambda_i(y) \exp\big(-\lambda_i(y) z_i\big).$$

(GEN3) ***Constant Conditional Entropy:*** *The product $\prod_{i=1}^{n} \lambda_i(y)$ is independent of y. Combined with a uniform prior $p(y) = 1/|\mathcal{Y}|$, this yields $p(y|z) = \mathrm{softmax}(W^* z)$ with $n' = n$.*

(GEN4) ***Fixed within-class population variance:*** *For a constant $C > 0$ shared by all admissible representations,*

$$\mathcal{V}_{\mathrm{within}}(h) := \mathbb{E}_Y[\mathrm{Tr}\,\mathrm{Cov}(h(Z) \mid Y)] = \mathbb{E}\big[\|h(Z) - \mathbb{E}[h(Z) \mid Y]\|_2^2\big] = C. \tag{15}$$

(GEN5) ***Energy Constraint:*** *Among global minimizers satisfying* (15)*, the learner selects the solution minimizing $\mathcal{R}(h) = \mathbb{E}_z\|h(Z)\|_2^2$.*

Theorem (Axis-aligned neurons)**.** *Following Assumption 8.1 and Assumption 8.2, in the generative setting with $m = n = n' = \mathrm{rank}(W^*)$, let (U, h) be any minimizer of $\mathcal{R}(h) = \mathbb{E}_z\|h(Z)\|_2^2$ among the cross-entropy global minimizers. Then*

$$h(z) = DP_\sigma z,$$

where P_σ is a permutation matrix and $D = \mathrm{diag}(\beta_1, \ldots, \beta_n)$ with $\beta_j > 0$. That is, the representation is identifiable up to permutation and component-wise positive scaling.

Proof.

We divide the proof into four steps.

1. Pure linear representation under exact capacity. By Theorem 4.1, every global minimizer of the cross-entropy satisfies

$$h(z) = Az + \nu(z), \qquad \nu(z) \in \mathcal{N}(U), \qquad UA = W^*,$$

where $A \in \mathbb{R}^{n\times n}$ has rank n. Since $m = n' = n$,

$$n = \mathrm{rank}(W^*) = \mathrm{rank}(UA) \le \mathrm{rank}(U) \le m = n.$$

Thus U has full column rank and $\mathcal{N}(U) = \{0\}$. Hence $\nu(z) = 0$ for every z, so $h(z) = Az$, with A invertible.

2. The linear map is entrywise non-negative. By GEN1, $Az = h(z) \in \mathbb{R}_+^n$ for every $z \in \mathbb{R}_+^n$. In particular, by evaluating on the coordinate rays, each column Ae_i is non-negative. Therefore

$$A_{ji} \ge 0 \qquad \text{for every } j, i \in [n].$$

3. Attainable lower bound of $\mathcal{R}(h)$. Let $r_i(y) := \lambda_i(y)^{-1} > 0, r(y) := (r_1(y), \ldots, r_n(y))^\top$. Under GEN2,

$$\mathbb{E}[Z \mid Y = y] = r(y), \qquad \mathrm{Cov}(Z \mid Y = y) = \mathrm{diag}\big(r_1(y)^2, \ldots, r_n(y)^2\big).$$

Consequently,

$$\begin{aligned}\mathcal{V}_{\mathrm{within}}(h) &= \mathbb{E}_Y\,\mathrm{Tr}\big(A\,\mathrm{Cov}(Z \mid Y)A^\top\big) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{n} A_{ji}^2 c_i = C, \qquad c_i := \mathbb{E}_Y[r_i(Y)^2] > 0.\end{aligned} \tag{16}$$

The conditional bias–variance identity of GEN5 and (16) give

$$\begin{aligned}\mathcal{R}(A) &= \mathbb{E}_z\|AZ\|_2^2 \\ &= \mathbb{E}_Y\big[\mathrm{Tr}\,\mathrm{Cov}(AZ \mid Y) + \|\mathbb{E}[AZ \mid Y]\|_2^2\big] \\ &= C + \mathbb{E}_Y\|A\mathbb{E}[Z \mid Y]\|_2^2 \\ &= C + \sum_{j=1}^{n}\sum_{i,k} A_{ji}A_{jk}\mathbb{E}_Y[r_i(Y)r_k(Y)] \\ &= 2C + 2\sum_{j=1}^{n}\sum_{1\le i<k\le n} A_{ji}A_{jk}\mathbb{E}_Y[r_i(Y)r_k(Y)].\end{aligned} \tag{17}$$

Because for all $j, i \in [n]$, $A_{ji} \geq 0$ and $r_i(y) > 0$, every term added to $2C$ is non-negative. Hence $\mathcal{R}(A) \geq 2C$.

It remains essential to show that this lower bound is feasible under all constraints. Let

$$\alpha_i := \sqrt{\frac{C}{nc_i}}, \qquad A_0 := \operatorname{diag}(\alpha_1, \ldots, \alpha_n), \qquad U_0 := W^* A_0^{-1}.$$

Then A_0 is non-negative and invertible,

$$\sum_{i=1}^{n} \alpha_i^2 c_i = C, \qquad U_0 A_0 = W^*,$$

so $(h_0(z) = A_0 z, U_0)$ is an admissible global minimizer and $\mathcal{R}(A_0) = 2C$. Hence the minimum activation energy is attained and is equal to $2C$.

4. Equality forces a positive scaled permutation. Let $A_\star$ denote the linear map of any activation-energy minimizer. Equality in (17) and strict positivity of every $r_i(y)$ imply

$$(A_\star)_{ji}(A_\star)_{jk} = 0 \qquad \text{for every } j \in [n] \text{ and } i \neq k.$$

Thus every row of $A_\star$ is at most one-sparse. Since $A_\star$ is an invertible $n \times n$ matrix, every row is nonzero and two rows cannot use the same column. Therefore every row and every column contains exactly one positive entry. For a permutation σ and positive numbers $\beta_j := (A_\star)_{j,\sigma(j)}$,

$$A_\star = DP_\sigma, \qquad D = \operatorname{diag}(\beta_1, \ldots, \beta_n),$$

and hence $h(z) = DP_\sigma z$. □

8.5.2 Compressed Sensing Recovery (Discriminative Setting)

Additional Assumptions

Assumption 8.3 (Discriminative Setting Assumptions)**.**

These assumptions apply to the discriminative setting ($Z \to Y$) where we analyze compressed sensing recovery.

(DIS1) ***Low-Rank Factorization:*** *The true weight matrix factorizes as $W^* = U^* A^*$, where $A^* \in \mathbb{R}^{n' \times n}$ is a measurement matrix and $U^* \in \mathbb{R}^{|\mathcal{Y}| \times n'}$ is a readout matrix. Since $n' = \operatorname{rank}(W^*)$, this implies the task depends only on $A^* z$.*

(DIS2) ***Latent Sparsity:*** *The latent variables are k-sparse: $\|z\|_0 \leq k$ for all $z \in \operatorname{supp}(p_Z)$.*

(DIS3) ***Sparse injectivity:*** *The true operator satisfies $\operatorname{spark}(A^*) > 2k$, equivalently,*

$$\mathcal{N}(A^*) \cap \{v \in \mathbb{R}^n : \|v\|_0 \leq 2k\} = \{0\}.$$

Theorem (Compressed sensing recovery)**.** *Following Assumption 8.1 and Assumption 8.3, in the discriminative setting with $W^* = U^* A^*$, assume $m \geq n'$, latents are k-sparse, and $\operatorname{spark}(A^*) > 2k$. Let P_U denote the orthogonal projector onto $\operatorname{Row}(U)$ and define the classifier-visible representation $\widetilde{h}(z) := P_U h(z) = Az$. Then the learned encoder satisfies $\mathcal{N}(A) = \mathcal{N}(A^*)$. Consequently, $\operatorname{spark}(A) = \operatorname{spark}(A^*) > 2k$, and z can be recovered from the unique k-sparse solution of*

$$Az = \widetilde{h}(z).$$

Proof. By Lemma 8.1, assume without loss of generality that W^* is row-centered.

1. Linear identifiability. Since $W^* = U^* A^*$ with $\operatorname{rank}(A^*) = n'$, we have $\operatorname{rank}(W^*) = n'$. With $m \geq n'$, Theorem 4.1 implies the learned encoder is linear:

$$h(z) = Az + \nu(z)$$

with $\operatorname{rank}(A) = n'$ and $\nu(z) \in \mathcal{N}(U)$.

2. Matching nullspaces. We show $\mathcal{N}(A) = \mathcal{N}(A^*)$. By the Log-Ratio Identity (Lemma 8.2), for any $y, \tilde{y} \in \mathcal{Y}$:

$$\langle u_y - u_{\tilde{y}}, h(z) \rangle = \langle w_y^* - w_{\tilde{y}}^*, z \rangle.$$

Since $h(z) = Az$ and $W^* = U^* A^*$, we can rewrite both sides:

$$\text{LHS} = \langle u_y - u_{\tilde{y}}, Az + \nu(z) \rangle = (u_y - u_{\tilde{y}})^\top Az + (u_y - u_{\tilde{y}})^\top \nu(z) = (u_y - u_{\tilde{y}})^\top Az$$
$$\text{RHS} = \langle w_y^* - w_{\tilde{y}}^*, z \rangle = (w_y^* - w_{\tilde{y}}^*)^\top z = (u_y^* - u_{\tilde{y}}^*)^\top A^* z,$$

where $\nu(z) \in \mathcal{N}(U)$. Now suppose $z \in \mathcal{N}(A)$, i.e., $Az = 0$. Then the LHS vanishes for all class pairs $(y, \tilde{y})$. Hence, we have $U^* A^* z = c\mathbf{1}$. Additionally, since $\mathbf{1}^\top U^* A^* = \mathbf{1}^\top W^* = 0$ and A^* has full row rank, we have $\mathbf{1}^\top U^* = 0$. Thus,

$$0 = \mathbf{1}^\top U^* A^* z = c|\mathcal{Y}|$$

That is, $c = 0$ and $A^* z \in \mathcal{N}(U^*)$. As $n' = \operatorname{rank}(W^*) = \operatorname{rank}(U^* A^*) \leq \operatorname{rank}(U^*) \leq n'$, this means that U^* is full column rank and $\mathcal{N}(U^*) = \{0\}$. Thus $A^* z = 0$ and

$$\forall z \in \mathcal{N}(A), z \in \mathcal{N}(A^*) \Longrightarrow \mathcal{N}(A) \subseteq \mathcal{N}(A^*).$$

Conversely, both $A \in \mathbb{R}^{m \times n}$ and $A^* \in \mathbb{R}^{n' \times n}$ have rank n', so:

$$\dim(\mathcal{N}(A)) = n - n' = \dim(\mathcal{N}(A^*)).$$

Combined with the inclusion above, we conclude $\mathcal{N}(A) = \mathcal{N}(A^*)$.

3. Factorization $A = BA^*$. Since A and A^* have the same nullspace of rank n', their row spaces are identical. This means there exists an left invertible matrix $B \in \mathbb{R}^{m \times n'}$ such that:

$$A = BA^*.$$

To see this explicitly: let $A^* = QR$ be a QR decomposition where $Q \in \mathbb{R}^{n' \times n'}$ is orthogonal and $R \in \mathbb{R}^{n' \times n}$ has the same row space as A^*. Since A has the same row space, we can write $A = \tilde{B}R$ for some full-rank $\tilde{B} \in \mathbb{R}^{m \times n'}$. Then $A = \tilde{B}Q^\top QR = (\tilde{B}Q^\top)A^*$, giving $B = \tilde{B}Q^\top$.

4. Sparse recovery. For any two distinct k-sparse vectors z_1, z_2 satisfying $h(z_1) = h(z_2)$, consider their difference $\Delta = z_1 - z_2$, which is at most $2k$-sparse. We have:

$$h(z_1) = h(z_2) \iff Az_1 = Az_2 \iff A\Delta = 0 \iff BA^*\Delta = 0.$$

Since B is invertible, $BA^*\Delta = 0$ if and only if $A^*\Delta = 0$. As $\operatorname{spark}(A^*) > 2k$, no nonzero $2k$-sparse vector lies in $\mathcal{N}(A^*)$. Therefore $\Delta = 0$, i.e., $z_1 = z_2$.

Thus h is injective on k-sparse vectors. Given $(U, h(z))$, the latent z can be recovered via standard sparse coding algorithms (e.g., Basis Pursuit, Lasso, or OMP), which find the sparsest solution to $Az = \widetilde{h}(z)$. □

Corollary (Basis Pursuit recovery)**.** *Following Theorem 4.4, suppose additionally that A^* satisfies the Null Space Property (NSP) of order k. Then A satisfies the same property as $\mathcal{N}(A) = \mathcal{N}(A^*)$. Consequently, every k-sparse latent z is the unique solution of*

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \textit{subject to} \quad Az = \widetilde{h}(z).$$

Proof. The Null Space Property depends only on the null space. It is therefore inherited by A from A^*, and the conclusion follows from the standard equivalence between NSP of order k and uniform exact Basis-Pursuit recovery of all k-sparse vectors. □

9 Experimental Details

Hardware & Software. All simulations were executed on a single *NVIDIA Tesla V100* (32 GB) GPU. The software environment comprised Python `3.12.3`, PyTorch `2.2.2` compiled against CUDA `12.1` and cuDNN `8902`, and scikit-learn `1.6.1`.

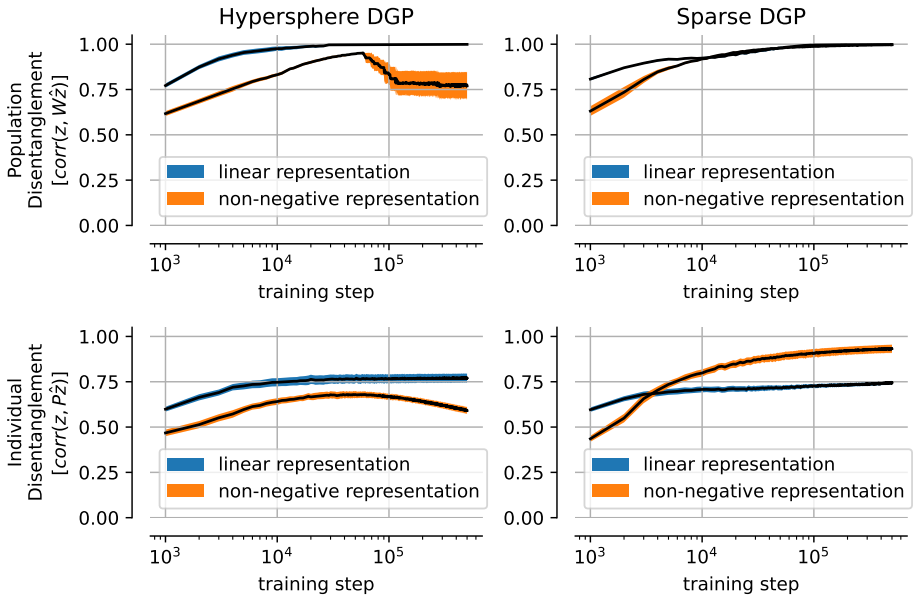

Figure 5: **Empirical Results.** *Same as in main text but with exponential instead of ReLU nonlinearity, which is easier to optimize in this setting.* **Top left)** DGP with latents on a hypersphere [19, 20], training a neural network with a linear representation (i.e., no final activation function, blue) and a non-negative representation (i.e., scale normalization and exponential nonlinearity as final activation function, orange); Disentanglement is measured up to arbitrary linear transformations. **Bottom left)** same as top left, but disentanglement is measured only up to permutations, measuring axis-aligned latent variable recovery. **Top right)** Our proposed sparse DGP with exponential distributions over non-negative latent variables, same architectures as on the left (with/without activation function), linear disentanglement reported. **Bottom right)** same as top right but individual, axis-aligned disentanglement measured. Shaded areas in all plots indicate standard error of the mean over 10 random seeds.

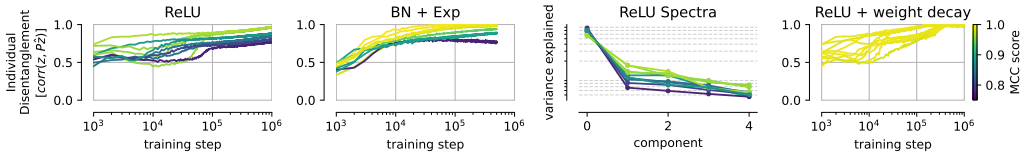

Figure 6: **Optimization Dynamics. A)** MCC curves over 10 random seeds. **B)** Same as A, but with batch normalisation and exponential nonlinearity. **C)** ReLU models with better final performance have a flatter power spectrum. **D)** Adding weight decay 0.03 improves convergence for ReLU model.

Random Seeds. Every figure and table reports the mean ± s.e.m. over ten independent runs. We deterministically set `seed`$= 42 + i$ with $i \in \{0, \ldots, 9\}$, forwarding the value to NumPy, PyTorch (CPU and CUDA), and Python's `random` module.

Hyper-parameters. All optimiser and model settings were adopted verbatim from the study by Reizinger et al. [40] (see their Sec. 4.2 and App. B). Re-using this validated grid eliminates the risk of inadvertently tuning hyper-parameters to our synthetic benchmarks.

9.1 Axis-aligned Neurons ($m = n'$)

For the *full-capacity* experiments we considered five latent variables. In the sparse condition each component followed an exponential distribution with rate $\lambda = 10$, whereas in the hyperspherical condition latents were drawn from a von Mises–Fisher distribution with concentration $\kappa = 150$ and

subsequently normalised to unit length. Latents were assigned to one of $C = 100$ class prototypes, obtained once per seed by sampling C random unit vectors and performing nearest-neighbour assignment. Every configuration used $N = 10\,000$ latent–label pairs.

A fixed, invertible forward map g—implemented as an MLP with layer widths $5 \to 64 \to 64 \to 64 \to 5$, *SiLU* activations, and orthonormal weight initialisation—generated the observations. The encoder f mirrored this architecture. Its output layer was either left *linear* or constrained to be *non-negative*. In the main text we achieved non-negativity with a ReLU; in Fig. 5 of the supplement we instead applied an element-wise exponential followed by scale-only batch normalisation (denoted *exp+BN*), which improves gradient flow in this toy setting. A linear classifier $W : \mathbb{R}^5 \to \mathbb{R}^{100}$ was trained jointly with f under a cross-entropy objective.

Optimisation used Adam with default β-parameters $(0.9, 0.999)$, $\epsilon = 10^{-8}$, a learning-rate of 3×10^{-4}, batch size 512, and no weight decay unless explicitly noted. Training proceeded for $50\,000$ gradient steps (approximately 500 epochs).

We quantified *linear alignment* by the R^2 of an ordinary-least-squares fit $z \leftarrow h(z)$, averaged across dimensions. *Axis alignment* was measured by the mean-correlation coefficient (MCC) of Hyvärinen and Morioka [18], obtained after Hungarian matching between recovered axes and ground-truth latents.

Optimization Dynamics. As shown in Figure 1, although the non-negative encoder eventually resolves the axes, its convergence is noticeably slower. We wanted to know what causes the very slow convergence, beyond the fact that it has to converge to a permutation. In Fig. 6B we tried using a different activation function (exponential) and scale (batch) normalization to alleviate potential gradient issues in the ReLU activation. While this seemed to improve convergence in many cases, it is not ideal because, in contrast to ReLU, this is not a common activation function so that insights in this setting are less likely to bear on real systems. In a different line of attack, we observed in Fig. 6C that the models which converged to a perfect solution tended to have a flatter power spectrum. This motivated the hypothesis that some penalty on the activation scale might be important–which is also an important assumption in prior work to obtain axis alignment [29]. Again, in the spirit of using relevant building blocks, we tried a weight decay of 0.03 which did indeed lead to faster and more reliable convergence across all seeds (Fig. 6D).

To probe the optimisation dynamics, we varied only three factors relative to the base setting: the output activation (ReLU versus exp+BN), the L_2 weight-decay coefficient $\lambda_{\mathrm{WD}} \in \{0, 0.01, 0.03\}$, and—when studying gradient-flow effects—replacement of ReLU by exp+BN. Power-spectrum slopes (panel C) were computed from the singular values of the centred activation matrix $\hat{Z} = h(G(Z)) - \mathbb{E}[\, h(G(Z))\,]$ evaluated on the validation set.

Individual trajectories in Fig. 7A–D show that the sparse-ReLU encoder maintains high linear alignment but can approach axis alignment slowly and with substantial initialization dependence. Without weight decay, only $3/10$ runs reached MCC ≥ 0.90 within 10^6 updates, and final MCC had median 0.878 [IQR $0.816, 0.895$]. At weight decay 0.003, 0.01, and 0.03, the corresponding success counts were $7/10$, $9/10$, and $10/10$, and median final MCC was 0.946, 0.993, and 0.987, respectively. Among runs that reached the threshold, median time-to-threshold decreased from $640{,}000$ updates without weight decay to $569{,}000$, $297{,}000$, and $133{,}500$ updates at weight decay 0.003, 0.01, and 0.03, respectively. Thus, weight decay substantially improves the finite-budget reliability and speed of reaching highly axis-aligned solutions in this experiment, which suggests an optimization option for practice.

Assumption Ablations. We next perturb the assumptions underlying Theorem 4.3 one family at a time. Based on the previous results, we use weight decay 0.03 as the stable reference configuration for these ablations except for the energy-regularization panel. Each condition is trained for 10^6 updates using ten paired initialization seeds same as the optimization dynamics analysis. Within each comparison, the class-scale vectors, nonlinear mixing function, training-batch seed, and evaluation set are held fixed.

The reference configuration attains linear alignment 0.995 ± 0.000 and MCC 0.994 ± 0.000 (Fig. 8). Violating constant conditional entropy primarily affects the stronger axis-alignment criterion. With a bias-free readout, increasing the across-class standard deviation of the conditional-entropy term to 2 nats reduces MCC to 0.662 ± 0.009, while linear alignment remains 0.960 ± 0.001. Allowing

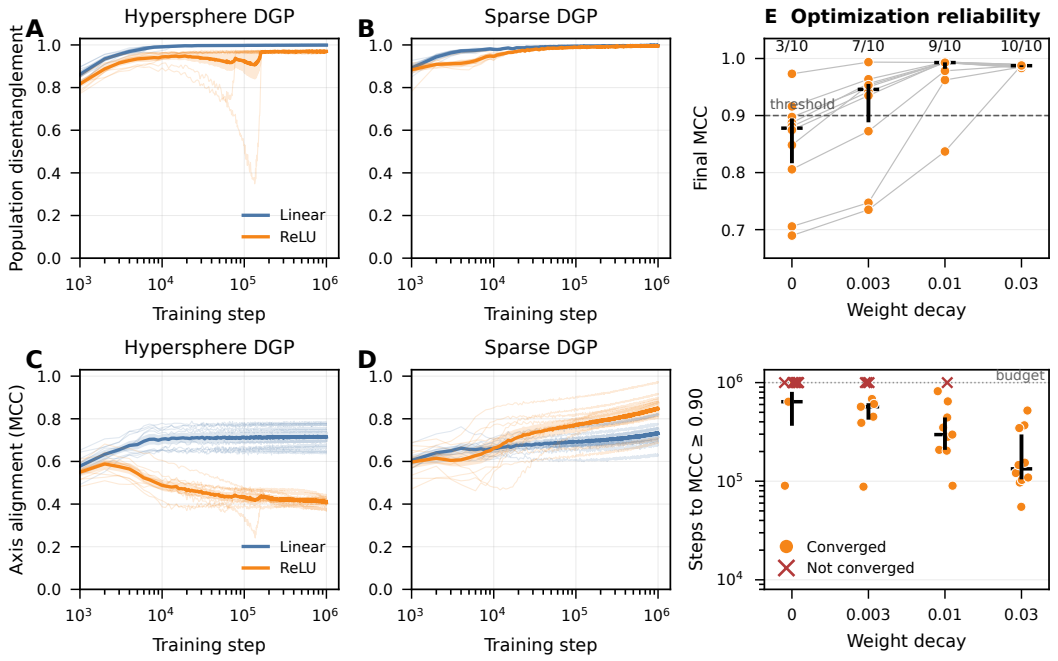

Figure 7: **Seed-level optimization dynamics and finite-budget reliability. A–D)** Exact-capacity experiments under a hyperspherical DGP (A,C) and the class-conditional sparse exponential DGP (B,D), using a linear output (blue) or a non-negative ReLU output (orange). A,B show alignment up to an unrestricted linear map, whereas C,D show permutation-and-scaling axis alignment measured by MCC. Thin lines denote individual initialization seeds, thick lines denote their mean, and shaded bands denote $\pm$ s.e.m. over ten seeds. All A–D runs use zero weight decay. **E)** Paired weight-decay sweep for the sparse-ReLU condition. The upper panel shows final MCC; orange points denote individual runs, gray lines connect the same initialization across weight-decay values, and the counts report the number of runs reaching MCC ≥ 0.90. The lower panel shows the first update at which this threshold was reached. Black bars denote the median and interquartile range. Red crosses denote runs that did not reach the threshold within the 10^6-update budget.

the theoretically matched affine readout largely absorbs the resulting class-dependent logit offsets, restoring linear alignment to 0.990 ± 0.001 and MCC to 0.962 ± 0.002.

Moderate conditional dependence is comparatively well tolerated. At Gaussian copula correlation 0.5, linear alignment and MCC remain 0.971 and 0.968, respectively. Mean-matched Gamma and Weibull conditionals likewise retain MCC 0.990 and 0.984, whereas the matched log-normal condition reduces MCC to 0.668. These results indicate that exponentiality and conditional independence are useful sufficient conditions but that modest departures from either assumption need not eliminate axis alignment in this finite-sample setting.

Signed latent support produces a sharper degradation: MCC falls to 0.770 when latent signs are flipped with probability 0.1, and to 0.049 at probability 0.5. By contrast, LeakyReLU and even a linear output attain MCC 0.987 and 0.968, respectively, under the regularized reference configuration. Removing weight decay leaves linear alignment near 0.999 but reduces MCC to 0.829 ± 0.029, compared with MCC 0.994 at weight decay 0.03.

These ablations should not be interpreted as tests of whether each assumption is individually necessary. The assumptions of Theorem 4.3 jointly provide a sufficient population-level guarantee, whereas the experiments measure the solutions selected by a particular finite-sample optimization procedure, which is a more practical scenario. Accordingly, persistence of axis alignment after an assumption is relaxed shows empirical robustness in this setup. Conversely, degradation identifies practically important boundaries.

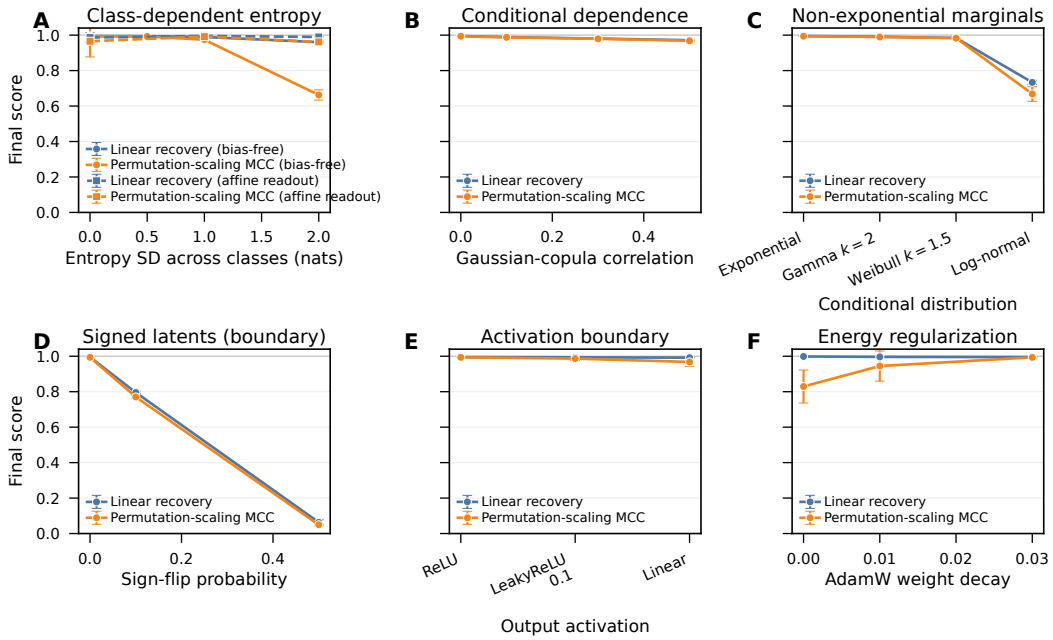

Figure 8: **Sensitivity to the assumptions of the axis-alignment result.** Each point shows mean ± s.e.m. over ten paired initialization seeds after 10^6 updates. Blue denotes held-out alignment up to an unrestricted linear map, and orange denotes permutation-and-scaling MCC. **A)** Class-dependent conditional entropy, with both bias-free and affine-readout controls. **B)** Conditional dependence introduced through a Gaussian copula while preserving the exponential marginals. **C)** Alternative class-conditional marginal families matched in conditional mean. **D)** Signed latent support introduced by independently flipping latent signs. **E)** Relaxations of output non-negativity using LeakyReLU and a linear output. **F)** AdamW weight decay. All other data-generation, training, and evaluation randomness is shared across paired conditions.

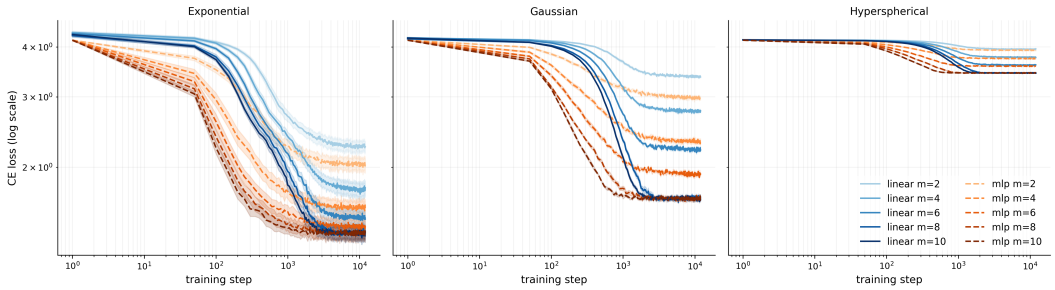

Figure 9: **Training Dynamics under undercomplete regime.** CE loss as a function of gradient update step for linear (solid blue) and MLP (dashed orange) encoders, across three latent distributions with encoder dimensions $m \in \{2, 4, 6, 8, 10\}$ (lighter shades indicate smaller m). Shaded areas indicate standard error of the mean over 10 random seeds.

9.2 Non-identifiability in the Undercomplete Regime ($m < n'$)

For the *undercomplete* experiments we fixed the latent and effective dimensions at $n = n' = 8$ and the number of classes at $|\mathcal{Y}| = 64$. The teacher weight matrix $W^* \in \mathbb{R}^{|\mathcal{Y}| \times n}$ was constructed via the low-rank factorisation $W^* = U^* A^*$, where U^* and A^* were built from the leading columns of independent QR decompositions of random Gaussian matrices with flat (unit) singular values, so that $\mathrm{rank}(W^*) = n'$ by construction. Soft labels $p^*(y \mid z) = \mathrm{softmax}(W^* z)$ were computed from latents drawn from three distributions: exponential $\mathrm{Exp}(1)$, standard Gaussian, and uniform hyperspherical.

Two encoder architectures sharing the same input/output dimensions were compared: a bias-free linear encoder $h(z) = Az$ and an MLP with three hidden layers of width 128, ReLU activations, and a linear output projection to $\mathbb{R}^m$ with no final nonlinearity. In both cases a bias-free linear readout $U \in \mathbb{R}^{|\mathcal{Y}| \times m}$ was trained jointly with the encoder under soft cross-entropy. The encoder dimension m was swept over $\{2, 4, 6, 8, 10\}$, straddling the theoretical threshold at $m = n' = 8$. Optimisation used Adam ($\beta = (0.9, 0.999)$, $\varepsilon = 10^{-8}$) with learning rate 5×10^{-4}, batch size 512, and 12,000 gradient steps without weight decay. After training, $\mathrm{KL}(p^* \| q)$ was evaluated by sampling with $N = 10{,}000$. Every configuration was repeated over ten random seeds ($42 + i$, $i \in \{0, \ldots, 9\}$), with model initialisations re-seeded deterministically as a function of all hyperparameters to ensure full reproducibility. Figure 2 reports the mean $\pm$ s.e.m. across seeds.

We also report the training dynamics in Figure 9. Through three latent distributions (Exponential, Gaussian and Hyperspherical), the MLP consistently attains lower final CE than the linear encoder at every $m < n' = 8$, with the gap closing as m approaches the capacity threshold; at $m \geq n'$, all architectures converge to comparable values, consistent with the Theorem 4.2 prediction.

9.3 Compressed Sensing Recovery

For the compressed sensing experiments we set the full latent dimension to $n = 16$, the effective dimension to $n' = 8$, and the number of classes to $|\mathcal{Y}| = 64$. The measurement matrix $A^* \in \mathbb{R}^{n' \times n}$ was drawn as a random Gaussian matrix with ℓ_2-normalised columns. The readout $U^* \in \mathbb{R}^{|\mathcal{Y}| \times n'}$ was drawn from an independent Gaussian, and the teacher was set to $W^* = U^* A^*$ with $\mathrm{rank}(W^*) = n'$ by construction. Latents z were k-sparse non-negative: the support of size k was drawn uniformly at random and the nonzero entries sampled from $\mathrm{Exp}(1)$, with all other entries set to zero. The sparsity level was swept over $k \in \{1, 2, 3\}$. Since $\mathrm{spark}(A^*) \leq n' + 1$, the spark condition $\mathrm{spark}(A^*) > 2k$ of Theorem 4.4 requires $2k \leq n'$, which holds for all three values. We verified this condition by an exhaustive check that every set of $2k$ columns of A^* is linearly independent.

A bias-free linear encoder $h(z) = Az$ with $A \in \mathbb{R}^{m \times n}$, $m = n' = 8$, was trained jointly with a bias-free linear readout $U \in \mathbb{R}^{|\mathcal{Y}| \times m}$ under soft cross-entropy against teacher labels $p^*(y \mid z) = \mathrm{softmax}(W^* z)$. Optimisation used Adam ($\beta = (0.9, 0.999)$, $\varepsilon = 10^{-8}$) with initial learning rate 10^{-3} and a cosine annealing schedule over 20,000 gradient steps at batch size 512 without weight decay. Throughout training we logged $\mathrm{KL}(p^* \| q)$ and the row-space alignment $\cos(\theta_{\max}) = \sigma_{\min}(Q_A^\top Q_{A^*})$, where Q_A and Q_{A^*} are orthonormal bases of $\mathrm{row}(A)$ and $\mathrm{row}(A^*)$ obtained via QR decomposition to measure the similarity between the $\mathcal{N}(A)$ and $\mathcal{N}(A^*)$.

Three recovery methods were applied to the compressed activations Az: supervised Lasso (ℓ_1 regression with $\alpha = 5 \times 10^{-3}$ and non-negativity constraint, solved per sample), dense linear OLS (fitted on 5,000 training samples and applied to 1,000 test samples), and unsupervised dictionary learning followed by sparse coding (sklearn `DictionaryLearning` with n components, $\alpha = 5 \times 10^{-3}$, coordinate-descent fitting, and trained on $h(z_{\mathrm{train}}) = Az_{\mathrm{train}}$ without access to z). Recovery quality was assessed by R^2 for supervised methods and by the mean correlation coefficient (MCC) with Hungarian matching between recovered and ground-truth latent dimensions for all methods. Each configuration was repeated over ten independent seeds ($42 + i$, $i \in \{0, \ldots, 9\}$). Results are reported in Figure 3.

9.4 Real-data Analyses

Vision models. We examined four ImageNet-trained networks: VGG-19 BN, ResNet-50, ViT-b/16, and DINOv2 ViT-S/14. All images from the 50 000-sample ImageNet-1K validation set were resized to 224×224 pixels and normalised according to the respective preprocessing pipelines. We extracted activations from layer `fc6` (4096 units) in VGG-19, the global-average-pooled penultimate layer (2048 units) in ResNet-50, and the class token (768 and 384 units, respectively) in the two Vision Transformers.

Large-language models. For GPT-2 small (124 M parameters), Pythia-70 M, and TinyLlama-1.1 B we generated single-token sentences containing each of the 100 000 most frequent Wikipedia unigrams. We recorded the last-token MLP activations from the final transformer block and computed semantic similarity using GPT-2-small embeddings (768-d), thereby avoiding circularity in the MIS calculation.

Biological data. Neuronal firing rates were taken from the inferior-temporal cortex recordings of Majaj et al. [42]. Following the original preprocessing (square-root transformation and per-neuron normalisation), we analysed 58 neurons from subject 0 and 110 neurons from subject 1, each viewing 3 200 natural images.

Evaluation metrics.

Neuron interpretability was quantified with the Mechanistic Interpretability Score (MIS) of Zimmermann et al. [41], using 100 reference items and 10 queries per reference. MIS asks whether a unit's activation pattern is coherent enough to be predicted on unseen inputs. For unit i, each of N tasks shows high- and low-activating reference stimuli together with two held-out query stimuli, one high- and one low-activating for that unit. An automated observer compares the queries with the two reference sets and chooses which query is the high-activating one. If $p_{i,t}$ denotes its probability of the correct choice in task t, then

$$\mathrm{MIS}_i = \frac{1}{N} \sum_{t=1}^{N} p_{i,t},$$

so chance level is 0.5 and 1 indicates a perfectly recognizable activation pattern. For language models, stimulus similarity is computed from GPT-2-small embeddings, independently of the evaluated unit.

To assess representational overlap between two models we computed the Maximum Matching Unit (MMU), which asks whether a unit has a response-matched counterpart in another system. Let $a_i^{(S)}$ denote the responses of unit i in source system S to the shared stimulus set, and $a_j^{(T)}$ those of unit j in target system T with d_T units. Then

$$\mathrm{MMU}_{S \to T}(i) = \max_{j \in [d_T]} \left| \mathrm{Corr}\big(a_i^{(S)}, a_j^{(T)}\big) \right|,$$

where Corr is the Pearson correlation across stimuli. A high MMU indicates a preserved response profile. MMU is directional, and its value tends to increase with the number d_T of candidate units. For each directed pair $S \to T$, we report the Pearson correlation, across the units of S, between MIS_i and $\mathrm{MMU}_{S \to T}(i)$. Correlations reported in Fig. 4C–E were tested two-sided against the null hypothesis $\rho = 0$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The five contributions stated in the abstract and Section 1 are substantiated by Theorem 4.1 to Theorem 4.4 in Section 4 and the corresponding experiments in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 explicitly discusses four limitations of the work, including three assumptions proposed, and one open question of real data generating process.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: General assumptions are formally stated in Assumption 8.1, with additional assumptions for two DGPs formally stated in Assumption 8.2 and Assumption 8.3 respectively. Complete proofs of Theorem 4.1 to Theorem 4.4 are provided in Section 8.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 9 provides all information needed to reproduce the results, including architectures, hyperparameters, optimizer settings, and the deterministic random seed protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for reproducing all experiments and figures is publicly available at `https://github.com/klindtlab/complete-or-sparse`, with a README, a reproducibility guide, and a pinned requirements file. Synthetic data are generated by the released code from the DGP specifications in Section 3 and Supplement 9. The real-data analyses use publicly available pretrained models (vision and language) and the publicly released macaque IT recordings of Majaj et al. [42]; access and preprocessing are described in the repository and in Supplement 9.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 9 fully specify all training and evaluation details. For each experiment in Section 5, there is a corresponding subsection in Section 9 reporting implementation details including data generating process parameters, network architectures, optimizer settings, training durations, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All synthetic experiments are repeated over ten independent random seeds and results are reported with mean $\pm$ one standard error. (shaded areas or error bars in the figures)

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 9 lists complete information about hardware & software. All experiments were small scale and can be reproduced on a single V100.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work is a theoretical research that does not involve human subjects, sensitive data, or applications with foreseeable potential for harm, and conforms with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This paper presents purely theoretical results on neural network identifiability with no direct path to harmful applications; potential positive impacts (advancing mechanistic interpretability) are noted in Section 7.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release pre-trained large-scale generative models or scraped datasets, and therefore poses no risks requiring safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets are credited via citations.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets, benchmarks, or model checkpoints are introduced as standalone assets. The codes are released at `https://github.com/klindtlab/complete-or-sparse`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper involves no crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects research was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs do not form any original or non-standard component of the core methodology, and does not affect the scientific originality of the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.