Linear Identifiability Beyond Diversity: An Empirical Study of Effective Complexity

Thomas Trevisan

I. Abstract

Recent work on identifiable representation learning has shown that, for a broad class of discriminative models with inner-product structure, any two models that induce the same conditional distribution must learn representations that are equivalent up to linear transformations. This guarantee holds under a condition known as diversity, which requires that the encoder and unembedding jointly exploit the full dimensionality of the representation space. When the diversity condition is violated, this equivalence is restricted to a lower-dimensional linear subspace whose dimension defines the model's effective complexity. While these results are asymptotic and distributional in nature, it remains unclear how strongly they manifest in finite-sample, overparameterized neural networks trained with stochastic optimization.

In this work, we empirically study linear identifiability and effective complexity in a controlled synthetic setting. We construct a ground-truth teacher model that defines a conditional distribution over class labels via a nonlinear encoder and a linear unembedding, and train a collection of student networks—varying in architecture, initialization, regularization, and representation dimensionality—to match the teacher distribution by minimizing the KL divergence. After training, we analyze the learned representations using principal component analysis, canonical correlation analysis, and reconstruction-error metrics to assess alignment across models and concentration onto low-dimensional linear subspaces. Our experiments demonstrate that, across architectures and optimization choices, student models consistently recover a shared low-dimensional predictive subspace that is linearly aligned with the teacher's, while exhibiting substantial freedom in orthogonal directions.

II. Introduction

Neural networks trained on the same predictive task often learn internal representations that exhibit strikingly similar geometric structure. In particular, embeddings produced by independently trained models—despite differences in architecture, initialization, or optimization—are frequently observed to be related by simple linear transformations. Such linear correspondences have been documented across a wide range of settings, including supervised classification, contrastive learning, and next-token prediction. These empirical regularities motivate a precise question: when multiple models induce the same conditional distribution, do they recover the same low-dimensional predictive subspace implied by effective complexity?

Recent advances in identifiability theory provide a principled framework for understanding these observations. In particular, Roeder et al. (2021)[1] study discriminative models of the form

$$p(y|x) = \frac{exp(f(x)^\top g(y))}{\sum_{y'} exp(f(x)^\top g(y'))}$$

where $f(x) \in \mathcal{R}^d$ denotes a learned representation of the input and $g(y) \in \mathcal{R}^d$ denotes a learned representation of the output. They show that, under a condition known as diversity, any two models inducing the same conditional distribution must have representations that are equivalent up to an invertible linear transformation. Concretely, if two models (f, g) and $(\tilde{f}, \tilde{g})$ satisfy $p_{f,g}(y|x) = p_{\tilde{f},\tilde{g}}(y|x)$ satisfy and the diversity condition holds, then there exists an invertible matrix $\mathcal{A}$ such that $f(x) = \mathcal{A}\tilde{f}(x)$, with a corresponding linear relationship between g and $\tilde{g}$ that preserves the inner products defining the model. This result provides a theoretical explanation for the reproducibility of linear structure across independently trained models, and formalizes the sense in which such structure is an intrinsic consequence of the modeling framework rather than an artifact of optimization.

However, the diversity condition is often violated in modern neural networks. In practice, models are frequently overparameterized relative to the complexity of the task, and the learned representations may occupy only a strict subspace of the ambient embedding dimension. To address this regime, Marconato et al.[2] introduce an extension of linear identifiability theory based on the notion of effective complexity. Their key observation is that the conditional distribution depends only on inner products of the form $f(x)^\top g_0(y)$, where $g_0(y) = g(y) - g(y_0)$ for a fixed reference output. Let $G = span(g_0(y) : y \in \mathcal{Y}$ denote the subspace spanned by output differences. Then only the projection of $f(x)$ onto G can influence the likelihood, since for any decomposition $f(x) = P_G f(x) + P_{G^\perp} f(x)$,

$$f(x)^\top g_0(y) = (\mathcal{P}_G f(x)^\top) g_0(y)$$

The effective complexity k is defined as $\dim(G)$ (equivalently, the dimension of the corresponding predictive subspace for f). When diversity holds, $G = R^d$ and effective complexity equals the embedding dimension. When diversity fails, $\dim(G) < d$, and models inducing the same conditional distribution are required to coincide up to an invertible linear transformation only on this lower-dimensional(k) predictive subspace. Components orthogonal to it do not affect the likelihood and may vary arbitrarily without changing the distribution, and therefore cannot be identified from data.

In this work, we empirically investigate these theoretical predictions in a controlled synthetic setting designed to closely mirror the assumptions of identifiable discriminative models. We construct a ground-truth teacher network that defines a conditional distribution over class labels via a nonlinear encoder and a linear unembedding, and train multiple student networks—with varying architectures, initialization schemes, regularization, and representation dimensionalities—to match the teacher distribution by minimizing Kullback–Leibler divergence. Since all students are optimized toward the same conditional distribution, identifiability theory predicts that their learned representations should agree with the teacher's up to linear transformations, potentially restricted to a shared low-dimensional predictive subspace determined by the teacher's effective complexity when identifiability does not hold. Using principal component analysis, canonical correlation analysis, and reconstruction-based metrics, we quantify the degree of linear alignment and subspace concentration across independently trained models. This framework enables us to disentangle constraints imposed by the data distribution from those arising through optimization and inductive biases, and to empirically probe how effective complexity manifests in practice.

III. Methods

We design a controlled synthetic framework to empirically study effective complexity in identifiable discriminative models. Our goal is to construct a setting in which the effective complexity of the task is fixed and known by design, and to analyze whether overparameterized neural networks trained to match the same conditional distribution recover representations that concentrate onto the corresponding low-dimensional predictive subspace. To this end, we adopt a teacher–student paradigm in which a ground-truth model defines a conditional distribution with a prescribed effective complexity, and multiple student models are trained to approximate this distribution under varying architectural and optimization choices.

A. Model family and predictive subspaces

We consider discriminative models defining conditional distributions of the form

$$p(y|x) = \frac{exp(f(x)^\top g(y))}{\sum_y' exp(f(x)^\top g(y'))}$$

where $x \in \mathcal{R}^n$, denotes the input, $y \in \mathcal{Y}$ is a discrete label, $f : \mathcal{R}^n \to \mathcal{R}^d$ is an encoder producing a representation $f(x)$ and $g : \mathcal{Y} \to \mathcal{R}^d$ maps labels to embedding vectors. In the supervised classification setting considered here, $g(y)$ corresponds to a column of a matrix $W \in \mathcal{R}^{d\times|\mathcal{Y}|}$, and the model is equivalent to a softmax classifier with learned representations.

Crucially, the conditional distribution depends only on inner products between encoder outputs and unembedding vectors. As a result, only those components of the representation space that lie in the intersection of the image of the encoder and the column space of W can influence predictions. This observation underlies the notion of effective complexity: the dimensionality of the linear subspace of representations that actually contributes to the conditional distribution.

B. Teacher model and controlled effective complexity

To generate data, we construct a fixed teacher model that defines a ground-truth conditional distribution with explicitly controlled effective complexity. Inputs $x \in \mathcal{R}^3$ are sampled independently from a zero-mean Gaussian distribution with diagonal covariance and mapped through a multilayer perceptron encoder $f : \mathcal{R}^3 \to \mathcal{R}^d$, producing teacher embeddings $f(x)$. The teacher unembedding is defined by a matrix $W \in \mathcal{R}^{d\times|\mathcal{Y}|}$ whose columns correspond to class embeddings $g(y)$. We construct W by sampling a random matrix and orthonormalizing its columns via QR decomposition, yielding mutually orthogonal class vectors. This design explicitly controls the dimensionality of the linear subspace spanned by the output embeddings.

Identifiability theory shows that the conditional distribution depends only on relative unembedding vectors $g_0(y) = g(y) - g(y_0)$ where y_0 is a fixed reference class. This subtraction removes one degree of freedom corresponding to uniform shifts of all logits, reflecting the invariance of the softmax to adding the same constant to every class score. Consequently, although W has $|\mathcal{Y}|$ columns, the subspace $\mathcal{G} = span\{g_0(y) : y \in \mathcal{Y}\}|$ has dimension at most $|\mathcal{Y} - 1|$. By construction, our orthogonal W ensures that these $\mathcal{Y} - 1$ relative vectors are linearly independent whenever $d \geq \mathcal{Y} - 1$. We next consider the dimensionality of the encoder image $Im(f)$. Inputs are drawn from a continuous Gaussian distribution, and the encoder is an MLP with randomly initialized weights drawn from continuous distributions. Under these conditions, the probability that the set of embeddings $\{f(x)\}$ lies in a strict linear subspace of $\mathcal{R}^d$ is zero. Therefore, with overwhelming probability and sufficiently many samples, the encoder utilizes the full ambient space, i.e., $dim(Im(f)) = d$ The conditional distribution is defined as

$$p(y|x) = softmax(f(x)^\top W),$$

and depends only on inner products between $f(x)$ and vectors in $\mathcal{G}$. Since

$$f(x)^\top g_0(y) = (P_G f(x))^\top g_0(y)$$

only the projection of $f(x)$ onto $\mathcal{G}$ can influence the likelihood.

It follows that the effective complexity of the teacher model—the dimensionality of the predictive subspace—is:

$$k = min\{dim(Im(f)), |\mathcal{Y}| - 1\}.$$

Because $dim(Im(f)) = d$ with overwhelming probability in our construction, effective complexity is equal to $|\mathcal{Y}| - 1$ whenever $d \geq \mathcal{Y} - 1$. This setup allows us to fix the task's effective complexity by choosing the number of classes, while freely varying the ambient embedding dimension.

The teacher assigns soft labels according to:

$$p(y|x) = softmax(f(x)^\top W)$$

yielding a dataset of input–distribution pairs $(x, p(\cdot|x))$. Using soft targets ensures that the full conditional distribution, rather than only the most likely class, is matched during training. The dataset is split into training, validation, and test sets, and teacher embeddings on the test set are stored for subsequent geometric analysis.

C. Student models and overparameterization

We train a collection of student models to approximate the teacher distribution. Each student consists of an encoder f_θ and an unembedding matrix W_θ , inducing a conditional distribution

$$p_\theta(y|x) = softmax(f_\theta(x)^\top W_\theta)$$

Student encoders are implemented as either standard multilayer perceptrons or residual MLP variants, and may have embedding dimensionality strictly larger than that of the teacher.

All students are trained by minimizing the expected Kullback–Leibler divergence between the teacher and student distributions:

$$\mathcal{L}(\theta) = \mathrm{E}_x\Big[KL(p^*(\cdot|x) \parallel p_\theta(\cdot|x))\Big]$$

Optimization is performed using stochastic gradient methods with minibatching. We vary initialization seeds, optimizer choices, and regularization penalties in order to assess how optimization and inductive biases influence the structure of learned representations beyond what is required by distributional equivalence.

Because all student models are trained to approximate the same conditional distribution, theory predicts that they must agree with the teacher on the predictive subspace associated with the effective complexity.

D. Extraction of representations

After training, all models are evaluated on the held-out test set. For each input x, we extract the encoder output $f(x)$, yielding collections of teacher and student embeddings. These embeddings constitute the primary object of analysis.

To facilitate geometric comparisons, embeddings are centered and rescaled using simple normalization procedures. No alignment, projection, or linear transformation is applied prior to analysis; all structure is inferred directly from the learned representations.

E. Estimating effective dimensionality via PCA

To assess whether student representations concentrate onto a low-dimensional linear subspace matching the teacher's effective complexity, we apply principal component analysis (PCA) to the embeddings of each model. We compute the cumulative variance explained as a function of the number of retained principal components.

If embeddings are well-approximated by a k-dimensional linear subspace, the first k components should account for the majority of the total variance, with additional components contributing progressively less. Deviations from this behavior indicate spreading of representations beyond the predictive subspace.

Comparing variance-explained curves across student models allows us to evaluate how closely learned representations adhere to the effective complexity imposed by the teacher.

F. Linear alignment and shared predictive subspaces

To evaluate whether student models recover representations that are linearly related to the teacher's on the predictive subspace, we analyze alignment using canonical correlation analysis (CCA).

CCA identifies pairs of linear projections of teacher and student embeddings that maximize correlation, providing a direct measure of shared linear structure.

We further analyze residual embeddings after removing these leading canonical directions to assess whether remaining components carry predictive information or correspond to unconstrained degrees of freedom.

IV. Results

We evaluate whether overparameterized student models trained to match the same conditional distribution recover representations consistent with the effective complexity of the teacher model. We perform experiments across multiple architectures, numbers of classes, and teacher embedding dimensionalities. All reported results are averaged over five independent runs with different random seeds, and we report mean ± one standard deviation. The results reported in this section correspond to experiments with teacher embedding dimension $d = 3$ and $C = 3$ classes. The student embedding dimensions are $d' = 10$. Identifiability theory therefore predicts that representations should concentrate on a 2-dimensional predictive subspace, and that models matching the same conditional distribution should be linearly equivalent on this subspace.

A. Students accurately match the teacher distribution

Across all architectures and model sizes, student networks achieve low KL divergence with respect to the teacher distribution on the held-out test set, as displayed in table I. This indicates that all student models successfully approximate the same conditional distribution, validating the use of identifiability and effective-complexity predictions in the subsequent analysis. Performance is stable across random initializations, with small variance across seeds.

Architecture	Loss
Small MLP	0.00001
Medium MLP	0.00006
Large MLP	0.00004
Small residual MLP	0.00006
Medium residual MLP	0.00004
Large residual MLP	0.00003

TABLE I

KL DIVERGENCE VALUE FOR STUDENT AND TEACHER'S PROBABILITY DISTRIBUTION.

B. Learned representations concentrate around a low-dimensional subspace

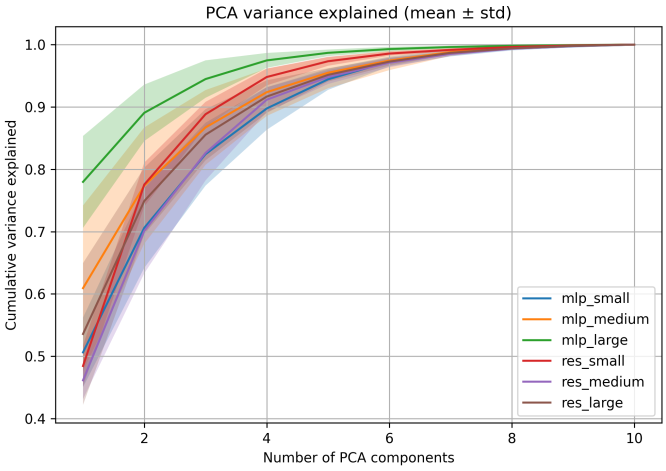

Fig. 1. Variance explained over PCA components

Figure 1 shows the cumulative variance explained by PCA applied to student embeddings. For all architectures, a small number of principal components explain a large fraction of the total variance. In particular, the first 2 components account for the majority of variance, while additional components contribute progressively less. Representations concentrate around a low-dimensional linear subspace embedded in a higher-dimensional representation space. This behavior is consistent with the notion of effective complexity: although models use high-dimensional representations, only a small number of directions appear to capture the structure relevant for prediction.

C. Student and teacher predictive subspaces are linearly aligned

Figure 2 reports canonical correlation analysis (CCA) between teacher and student embeddings. Across all architectures, the first two canonical correlations are consistently close to one, while subsequent correlations are substantially lower. This indicates that student models recover a predictive subspace that is linearly aligned with the teacher's predictive subspace. To further assess whether remaining dimensions contain shared structure, we perform residual CCA after projecting out the leading canonical directions, as shown in Figure 3. Residual correlations drop markedly, suggesting that dimensions outside the shared predictive subspace are weakly aligned and largely unconstrained. Together, these results support the prediction that models trained to induce the same conditional distribution agree on a low-dimensional predictive subspace, while retaining freedom in orthogonal directions.

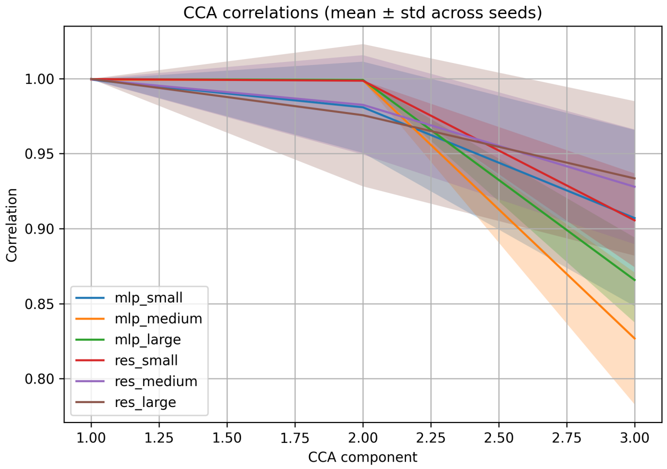

Fig. 2. Canonical correlation analysis on embeddings components

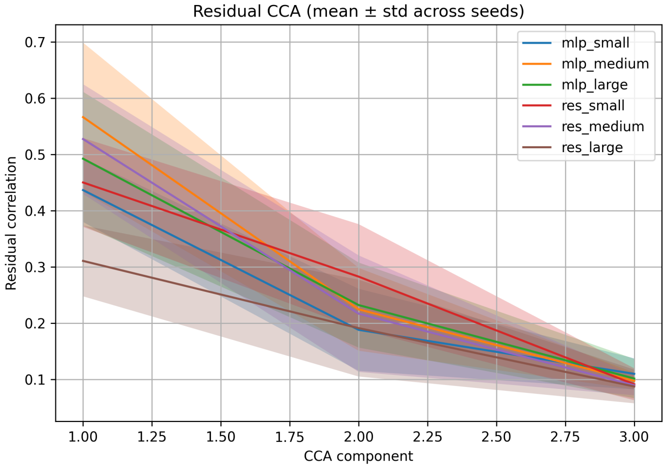

Fig. 3. Residual correlation of components

V. Conclusions

We presented an empirical study of linear identifiability and effective complexity in discriminative models using a controlled teacher–student setting. Across architectures and embedding sizes, student networks trained to match the same conditional distribution consistently learn representations that concentrate on a low-dimensional predictive subspace and are linearly aligned with the teacher on that subspace, in line with theoretical predictions.

Outside this predictive subspace, alignment is substantially weaker, confirming that these directions are largely unconstrained by the distribution. However, we observe that residual directions of the student embedding space still exhibit small but nonzero correlations across independently trained models. We hypothesize that this residual structure arises from training artifacts and shared inductive biases rather than from the task itself. Future work should investigate training strategies and model designs that tune or suppress these residual correlations as much as possible, in order to better isolate purely distribution-determined representations.

References

[1] G. Roeder, L. Metz, and D. P. Kingma, "On linear identifiability of learned representations," 2020. [Online]. Available: https://arxiv.org/abs/2007.00810

[2] E. Marconato, S. Lachapelle, S. Weichwald, and L. Gresele, "All or none: Identifiable linear properties of next-token predictors in language modeling," 2025. [Online]. Available: https://arxiv.org/abs/2410.23501