

---

# When Does Closeness in Distribution Imply Representational Similarity? An Identifiability Perspective

---

Beatrix M. G. Nielsen<sup>\*,1</sup>

Emanuele Marconato<sup>2</sup>

Andrea Dittadi<sup>†,3,4,5</sup>

Luigi Gresele<sup>†,6</sup>

<sup>1</sup>Technical University of Denmark <sup>2</sup>University of Trento, Italy <sup>3</sup>Helmholtz AI, Munich  
<sup>4</sup>Technical University of Munich <sup>5</sup>Munich Center for Machine Learning (MCML) <sup>6</sup>University of Copenhagen

## Abstract

When and why representations learned by different deep neural networks are similar is an active research topic. We choose to address these questions from the perspective of identifiability theory, which suggests that a measure of representational similarity should be invariant to transformations that leave the model distribution unchanged. Focusing on a model family which includes several popular pre-training approaches, e.g., autoregressive language models, we explore when models which generate distributions that are close have similar representations. We prove that a small Kullback–Leibler divergence between the model distributions does not guarantee that the corresponding representations are similar. This has the important corollary that models with near-maximum data likelihood can still learn dissimilar representations—a phenomenon mirrored in our experiments with models trained on CIFAR-10. We then define a distributional distance for which closeness implies representational similarity, and in synthetic experiments, we find that wider networks learn distributions which are closer with respect to our distance and have more similar representations. Our results thus clarify the link between closeness in distribution and representational similarity.

## 1 Introduction

How to compare and relate the internal representations learned by different deep neural networks is a long-standing and active research topic [3, 30, 33, 38, 45, 49, 65]. Understanding when and why various kinds of *representational similarity* emerge has implications for model stitching [12, 16, 36, 41, 46], knowledge-distillation [53, 58, 73, 74] and interpretability for concept-based and neuro-symbolic models [5, 15, 42], see [61] for a review. A theory of **when similarity occurs** will chiefly depend on **how similarity is measured**. As argued by Bansal et al. [3], a similarity measure “should be invariant to operations that do not modify the ‘quality’ of the representation, but it is not always clear what these operations are”. Indeed, this depends on how one defines the ‘quality’ of a representation, and previous works have debated the pros and cons of different choices [3, 14, 30, 31, 45, 49, 65, 67].

We choose to address these questions from the perspective of identifiability theory. In identifiability of probabilistic models, representations are called *equivalent* if they result in equal likelihoods, and

---

\*Correspondence to: bmg@dtu.dk or beat@itu.dk.

†Joint last authors.

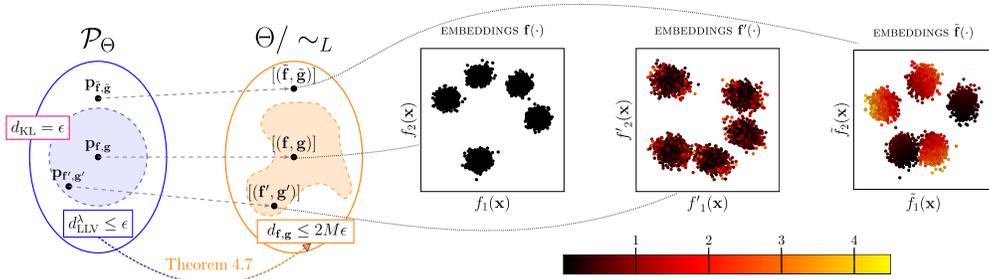


Figure 1: **When closeness in distribution does and does not imply representational similarity.** On the left, we show two distributions  $p_{\mathbf{f},\mathbf{g}}, p_{\mathbf{f}',\mathbf{g}'} \in \mathcal{P}_\Theta$  which are closer than  $\epsilon$  w.r.t. the distance  $d_{\text{LLV}}^\lambda$  (Definition 4.4), as illustrated by the shaded blue ball. We use  $[(\mathbf{f}, \mathbf{g})] \in \Theta / \sim_L$  to denote the *identifiability class* (Footnote 3) of a  $\sim_L$ -identifiable model with *embedding*  $\mathbf{f}$  and *unembedding*  $\mathbf{g}$  (Section 2). Theorem 4.7 implies that the identifiability classes  $[(\mathbf{f}, \mathbf{g})]$  and  $[(\mathbf{f}', \mathbf{g}')]$ , within the shaded orange area, will have *similar* representations—i.e., their dissimilarity under  $d_{\mathbf{f},\mathbf{g}}$  (Definition 4.6) is bounded above by  $2M\epsilon$ . We also consider a third distribution  $p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \in \mathcal{P}_\Theta$  which, while  $\epsilon$ -close in  $d_{\text{KL}}$  (in magenta), falls outside the blue region and has representations that are very *dissimilar* from those of  $[(\mathbf{f}, \mathbf{g})]$  and  $[(\mathbf{f}', \mathbf{g}')]$ , as described by our Theorem 3.1. On the right, we plot the three model embeddings: Taking  $\mathbf{f}$  as reference, we find the best linear fit to  $\mathbf{f}'$  and  $\tilde{\mathbf{f}}$ , and then color each of the points according to the residual error (brighter colors denote larger errors). The embeddings  $\mathbf{f}'$  are nearly a linear transformation of  $\mathbf{f}$ , while  $\tilde{\mathbf{f}}$  shows substantial deviation—visibly farther from being linearly related.

all models with equivalent representations together form an *identifiability class* [27].<sup>3</sup> Therefore, to preserve the ‘quality’ of representations, our notion of representational similarity needs to be invariant to precisely those transformations which leave the model likelihood unchanged. We focus on a model family including many popular pre-training approaches, e.g., autoregressive language models [8, 48], contrastive predictive coding [47], as well as standard supervised classifiers. Identifiability results for this model class show that, under a *diversity* condition, equal-likelihood models yield representations which are equal up to certain invertible linear transformations [28, 34, 51, 52].

As a first step toward a theory of representational similarity, we ask: **For what distributional and representational distances is it true that models whose output distributions are close have similar internal representations?** Addressing this requires going beyond classical identifiability, which assumes exact likelihood equality. We do so by proving in Theorem 3.1 that a small Kullback–Leibler (KL) divergence does not guarantee that the corresponding representations are similar—i.e., close to the identifiability class of our model family. As an important corollary, two models which are arbitrarily close to maximizing the data likelihood can still have dissimilar representations (Corollary 3.2). We then introduce a new distributional distance (Definition 4.4) and prove (Theorem 4.7) that closeness in this distance bounds representational dissimilarity (Definition 4.5).

We conduct classification experiments on CIFAR-10 and find substantial representational dissimilarity between some trained models with similarly good performance (Section 5.1). This dissimilarity appears to stem from a mechanism analogous to that identified in our KL divergence theory. We also run synthetic experiments showing that wider neural networks tend to learn distributions closer under our distance and have more similar representations (Section 5.2). This suggests that the inductive biases of larger models may promote representational similarity, something already observed in previous works.<sup>4</sup> Altogether, our findings indicate that the robustness of identifiability results [28, 51, 52] under the KL divergence may require additional assumptions, and that whether distributional closeness implies representational similarity depends critically on the choice of distributional and representational distance.

The **structure and main contributions** of our paper are as follows:

- We prove that small KL divergence between two models from our considered class (Equation (1)) does not guarantee similar representations (Section 3, Theorem 3.1); as a corollary, models with near-maximum data likelihood can have entirely dissimilar representations (Corollary 3.2).

<sup>3</sup> For a model family  $\Theta$  and an equivalence relation  $\sim_L$ ,  $\Theta / \sim_L$  denotes the quotient space whose elements are the *identifiability classes*  $[(\mathbf{f}, \mathbf{g})] := \{(\mathbf{f}', \mathbf{g}') \in \Theta : (\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}')\}$  [19, 27].

<sup>4</sup>E.g., Roeder et al. [52]: “as the representational capacity of the model and dataset size increases, learned representations indeed tend towards solutions that are equal up to only a linear transformation.”

- We define a distance between probability distributions (Section 4.1) and a dissimilarity measure between representations (Section 4.2), and prove that small distributional distance implies representational similarity up to a specific class of invertible linear transformations (Section 4.3).
- We empirically show that: (1) models trained on CIFAR-10 can exhibit dissimilar representations despite similarly good performance through a mechanism similar to that in the proof of Corollary 3.2 (Section 5.1); (2) on synthetic data, wider networks yield lower mean and variance of our distributional distance, and also have more similar representations (Section 5.2).

## 2 Model Class and Identifiability

We consider input variables  $\mathbf{x} \in \mathcal{X}$ , which can be real-valued (e.g., images) or discrete (e.g., text strings). We assume that the input data is sampled *i.i.d.* from a distribution  $p_{\mathcal{D}}(\mathbf{x})$ . Given an input  $\mathbf{x}$ , we consider the task of defining a conditional distribution over categorical outputs  $y \in \mathcal{Y}$ , which can be class labels in classification tasks or next-tokens following a context in language modeling.

**Model class.** Following Roeder et al. [52], we consider a model to be a pair of functions  $(\mathbf{f}, \mathbf{g}) \in \Theta$ , where the codomain of both  $\mathbf{f}$  and  $\mathbf{g}$  is a vector space  $\mathbb{R}^M$ , which we will also refer to as the *representation space*. Here,  $\Theta$  entails the space of all such pairs that can be generated by arbitrarily large neural networks.<sup>5</sup> We will refer to  $\mathbf{f}(\mathbf{x})$  as the model *embedding* and to  $\mathbf{g}(y)$  as the *unembedding*. The conditional distribution<sup>6</sup> defined by the model is given by

$$p_{\mathbf{f}, \mathbf{g}}(y|\mathbf{x}) = \frac{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y'))}. \quad (1)$$

We will also indicate the model distribution with  $p_{\mathbf{f}, \mathbf{g}} \in \mathcal{P}_\Theta$ , where  $\mathcal{P}_\Theta$  is the space of conditional distributions that can be constructed as in Equation (1) using models in  $\Theta$ . This model family captures a variety of models for different learning contexts [52] among which: (i) auto-regressive language models like GPT-2 [48] and GPT-3 [8]; (ii) supervised classifiers like energy-based models [28] and models trained with DIET [26]; (iii) pretrained language models like BERT [13]; (iv) self-supervised pretraining for image classification with Contrastive Predictive Coding (CPC) [47].

**Identifiability.** It is useful to fix a pivot input point  $\mathbf{x}_0 \in \mathcal{X}$  and a pivot label  $y_0 \in \mathcal{Y}$  and consider the displaced embeddings  $\mathbf{f}_0(\mathbf{x}) := \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)$  and unembeddings  $\mathbf{g}_0(y) := \mathbf{g}(y) - \mathbf{g}(y_0)$ . The conditional probability distribution generated by the model can then be rewritten as

$$p_{\mathbf{f}, \mathbf{g}}(y|\mathbf{x}) \propto \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y)) \exp(-\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y_0)) \quad (2)$$

$$= \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y)). \quad (3)$$

Next, we review *identifiability* of the model class in Equation (1). Identifiability theory characterizes the set of transformations of representations that leave a model’s output distribution unchanged; two models are therefore deemed equivalent when one can be obtained from the other by such a transformation. To proceed, we introduce an assumption about our considered models. This condition, also known as *diversity* [28, 34, 52], corresponds to only considering models that *span the whole representation space*, in the sense that there is no proper linear subspace of  $\mathbb{R}^M$  containing the embeddings or the unembeddings [43]. Formally, this can be written as:

**Definition 2.1** (Diversity condition). *A model  $(\mathbf{f}, \mathbf{g}) \in \Theta$  satisfies the diversity condition if there exists  $\mathbf{x}_1, \dots, \mathbf{x}_M \in \mathcal{X}$  and  $y_1, \dots, y_M \in \mathcal{Y}$  such that both the displaced embedding vectors  $\{\mathbf{f}_0(\mathbf{x}_i)\}_{i=1}^M$  and the displaced unembedding vectors  $\{\mathbf{g}_0(y_i)\}_{i=1}^M$  are linearly independent.*

The diversity condition can hold when the total number of unembeddings in  $\mathcal{Y}$  is strictly higher than the representation dimension  $M$  [52]. For example, in GPT-2 [48], while the representation dimensionality varies in the order of thousands (depending on the model size), the number of tokens in  $\mathcal{Y}$  is of the order of ten thousands ( $\sim 50\text{k}$  tokens), implying that diversity may be easily satisfied. With this assumption, we can prove the *identifiability* of the model in Equation (1) up to invertible linear transformations (a detailed proof can be found in Appendix B):

<sup>5</sup>Both  $\mathbf{f}$  and  $\mathbf{g}$  are therefore treated as nonparametric functions in the following, similar to [28, 34, 43, 52].

<sup>6</sup>We use the term “distribution” generically. In Equation (1),  $p_{\mathbf{f}, \mathbf{g}}(y|\mathbf{x})$  is a softmax over a finite label set, i.e., a probability *mass* function.

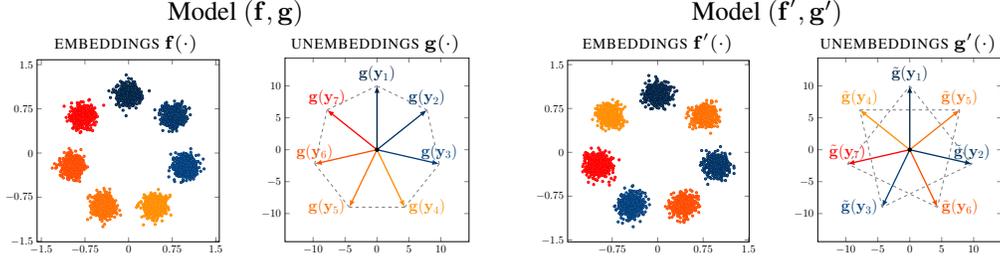


Figure 2: **Two models with small KL divergence but highly dissimilar representations.** We construct two models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  whose representations are related by a non-linear transformation: The embeddings and unembeddings of the  $(\mathbf{f}', \mathbf{g}')$  model are constructed by permuting the embedding clusters and the corresponding unembedding vectors of the  $(\mathbf{f}, \mathbf{g})$  model. As a result, the nearest unembedding vectors in  $\mathbf{g}(\cdot)$  (dashed lines) are mapped away from each other in  $\mathbf{g}'(\cdot)$ . In [Theorem 3.1](#), we show that as the norm of the unembedding vectors  $\rho$  grows for both models, their distributions  $p_{\mathbf{f}, \mathbf{g}}$  and  $p_{\mathbf{f}', \mathbf{g}'}$  become closer in KL divergence, whereas their representations remain dissimilar—i.e., far from being equal up to a linear transformation, see also [Table 1](#).

**Theorem 2.2** (Linear Identifiability [[28](#), [34](#), [52](#)]). *Let  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$ , and  $(\mathbf{f}, \mathbf{g})$  satisfy the diversity condition ([Definition 2.1](#)). Let  $\mathbf{L}$  (resp.  $\mathbf{L}'$ ) be the matrix with columns  $\mathbf{g}_0(y)$  (resp.  $\mathbf{g}'_0(y)$ ), and let  $\mathbf{A} := \mathbf{L}^{-\top} \mathbf{L}'^{\top} \in \mathbb{R}^{M \times M}$ . Then:*

$$p_{\mathbf{f}, \mathbf{g}}(y | \mathbf{x}) = p_{\mathbf{f}', \mathbf{g}'}(y | \mathbf{x}), \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \implies (\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}') \quad (4)$$

where the equivalence relation  $\sim_L$  is defined by

$$(\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}') \iff \begin{cases} \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{f}'(\mathbf{x}) \\ \mathbf{g}_0(y) = \mathbf{A}^{-\top} \mathbf{g}'_0(y) \end{cases} \quad (5)$$

The result of [Theorem 2.2](#) is central to the scope of our analysis because it shows that representations of two models generating the same conditional distribution are related by an invertible linear transformation. This also means that, under the diversity condition, to each probability distribution  $p_{\mathbf{f}, \mathbf{g}} \in \mathcal{P}_{\Theta}$  corresponds a unique (up to equivalence) choice of embeddings and unembeddings. See [Fig. 1](#) for an illustration.

**Representational similarity.** [Theorem 2.2](#) suggests that a natural notion of similarity between representations for our model class should account for the linear equivalence relation in [Equation \(5\)](#). Hereafter, we say two models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  have *equivalent representations* if their embeddings and unembeddings are related by an invertible linear transformation of the kind in [Theorem 2.2](#), i.e.,  $(\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}')$ , and we call representations *similar* if they are ‘close’ to having such a relation—which we will make precise in [Section 4.2](#). We note that several measures of representational similarity [[30](#), [45](#), [49](#)] are based on Canonical Correlation Analysis (CCA) [[22](#)], and are thus invariant to any invertible linear transformation. They thus attain their maximum when applied to representations that are equivalent in the sense above, but also for other invertible linear transformations. We discuss alternative choices of similarity measures in [Section 6](#).

### 3 Models Close in KL Divergence Can Have Dissimilar Representations

Given [Theorem 2.2](#), we next ask whether a similar conclusion holds approximately when the models have *nearly* the same distributions. More precisely, we ask: For a given choice of divergence between distributions, does a small difference in distributions imply representational similarity?

As a natural starting point, we consider the widely used Kullback–Leibler (KL) divergence. Specifically, we quantify the difference between two models by the KL divergence between their conditional distributions, averaged over the data distribution  $p_{\mathcal{D}}(\mathbf{x})$ . This is equivalent to the KL divergence between the corresponding joint distributions, assuming both share the marginal  $p_{\mathcal{D}}(\mathbf{x})$ . Formally:

$$d_{\text{KL}}(p, p') := \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [D_{\text{KL}}(p(y|\mathbf{x}) \| p'(y|\mathbf{x}))] = D_{\text{KL}}(p(\mathbf{x}, y) \| p'(\mathbf{x}, y)). \quad (6)$$

When  $D_{\text{KL}}(p(y|\mathbf{x}) \| p'(y|\mathbf{x})) = 0$  for all  $\mathbf{x}$ , we obtain models that are  $\sim_L$ -equivalent, as specified in [Theorem 2.2](#). However, only making the KL divergence small is insufficient to conclude much about the similarity of representations. The following result formalizes this claim (proof in [Appendix C.1](#)):

**Theorem 3.1** (Informal). *Let  $k = |\mathcal{Y}|$  and assume  $k \geq M + 1$ . Then  $\forall \epsilon > 0$ , there exist models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  for which  $d_{\text{KL}}(p_{\mathbf{f}, \mathbf{g}}, p_{\mathbf{f}', \mathbf{g}'}) \leq \epsilon$ , and whose embeddings are far from being linearly equivalent.<sup>7</sup> More precisely, one can construct a sequence of model pairs, depending smoothly on a real-valued parameter  $\rho$ , such that, as  $\rho \rightarrow \infty$ , the KL divergence between the models tends to zero while their embeddings remain fixed and not equal up to any linear transformation.*

*Proof sketch.* We here sketch the construction used for  $k \geq M + 2$  (see [Appendix C.1](#) for a construction which holds with  $k = M + 1$ ). Let  $(\mathbf{f}, \mathbf{g}) \in \Theta$  be a model which has unembeddings with fixed norm, i.e.,  $\|\mathbf{g}(y)\| = \rho$ , for all  $y \in \mathcal{Y}$  and  $\rho \in \mathbb{R}^+$ . Assume the unembeddings have non-zero angles between them and let the unembeddings satisfy diversity ([Definition 2.1](#)). For  $\mathbf{x} \in \mathcal{X}$ , let the embedding  $\mathbf{f}(\mathbf{x})$  have strictly largest cosine similarity with one  $\mathbf{g}(\hat{y})$ , for some  $\hat{y} \in \mathcal{Y}$ , and such that  $\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}}(p_{\mathbf{f}, \mathbf{g}}(y|\mathbf{x}))$ . As a consequence, note that the model  $(\mathbf{f}, \mathbf{g})$  also satisfies the diversity condition for the embeddings. Let  $(\mathbf{f}', \mathbf{g}') \in \Theta$  be another model which is constructed starting from  $(\mathbf{f}, \mathbf{g})$  and considering a permutation,  $\pi$ , of the label indices such that  $\mathbf{g}'(y_i) = \mathbf{g}(y_{\pi(i)})$  for  $i = 1, \dots, k$ , with a corresponding shift of the embedding clusters.<sup>8</sup> For every  $\mathbf{x} \in \mathcal{X}$ , we let  $\|\mathbf{f}'(\mathbf{x})\| = \|\mathbf{f}(\mathbf{x})\|$  and let the angle between  $\mathbf{f}'(\mathbf{x})$  and  $\mathbf{g}'(\hat{y})$  be equal to that between  $\mathbf{f}(\mathbf{x})$  and  $\mathbf{g}(\hat{y})$ . We illustrate this construction in [Fig. 2](#) for  $\mathbf{f}(\mathbf{x}), \mathbf{g}(y) \in \mathbb{R}^2$ . One can then find a permutation  $\pi$  for which  $\mathbf{f}(\mathbf{x})$  is not a linear transformation<sup>9</sup> of  $\mathbf{f}'(\mathbf{x})$  (thus the KL divergence between  $p_{\mathbf{f}, \mathbf{g}}$  and  $p_{\mathbf{f}', \mathbf{g}'}$  is non-zero). However, as  $\rho \rightarrow \infty$ , we have that  $d_{\text{KL}}(p_{\mathbf{f}, \mathbf{g}}, p_{\mathbf{f}', \mathbf{g}'}) \rightarrow 0$ , although the embeddings stay constant and not linearly equivalent.  $\square$

[Theorem 3.1](#) shows that, even when  $d_{\text{KL}}(p_{\mathbf{f}, \mathbf{g}}, p_{\mathbf{f}', \mathbf{g}'}) \leq \epsilon$ , there is no guarantee that the models are close to being  $\sim_L$ -equivalent. The proof constructs a  $\rho$ -parameterized sequence in which the embeddings remain fixed and not equal up to a linear transformation; consequently, the error of the best linear regression of one embedding set onto the other is strictly positive and stays constant as  $\rho \rightarrow \infty$ , while  $d_{\text{KL}} \rightarrow 0$ .

In [Theorem 3.1](#), we consider the KL divergence between two distributions from our model family. However, when practitioners study representational similarity in trained models, they fit each model to the data distribution—usually by maximum-likelihood estimation—and then compare their representations. We therefore also formulate the following corollary which is closer to the usual setup:

**Corollary 3.2** (Informal). *Let  $k = |\mathcal{Y}|$  and assume  $k \geq M + 1$ . Consider a dataset of  $(\mathbf{x}, y)$  pairs, where only one label is assigned to each unique input. Then, we can find two models that obtain an arbitrarily small cross-entropy loss on the data while having representations which are far from being linearly equivalent in the same sense as in [Theorem 3.1](#).*

The proof is in [Appendix C.2](#), and an illustration through synthetic experiments can be found in [Appendix F.4](#). This corollary shows that two models having close to optimal classification loss on training data, possibly also equal, may not possess similar representations at all. Importantly, we find that this situation can arise in practice with discriminative models trained on CIFAR-10 ([Section 5](#)), which can have dissimilar representations despite assigning high probability to the same labels.

## 4 When Closeness in Distribution Implies Representational Similarity

In this section, we introduce a distance between probability distributions ([Section 4.1](#)) and measure of representational similarity ([Section 4.2](#)) and prove that, for our model family ([Equation \(1\)](#)), small values of the former imply high similarity under the latter ([Section 4.3](#)).

<sup>7</sup>We show in [Section 5.3](#) how the construction from our proof can lead to high dissimilarity both according to  $m_{\text{CCA}}$  [[49](#), [52](#)] and under the measure introduced in [Definition 4.5](#).

<sup>8</sup>This is not a permutation of the representation dimensions—i.e., it is not  $\mathbf{f}'(\mathbf{x}) = \mathbf{P}\mathbf{f}(\mathbf{x})$  for some permutation matrix  $\mathbf{P}$ . The permutation of label indices, and corresponding shift of embedding clusters, is in general not a linear transformation of the representations (see [Fig. 2](#)); it also changes the model’s output distribution, since it alters the rank—by predicted probability—of every label except the highest-probability one for each input  $\mathbf{x}$ .

<sup>9</sup>In [Table 1](#), we show that one can construct models this way and such that the mean canonical correlation is close to zero, whereas it would be equal to one if their representations were linear transformations of each other.

## 4.1 A Distance Between Probability Distributions

We start by studying what relation holds between models in  $\Theta$  which need not have the same probability distributions. Hereafter, we will also consider the variance over input samples and outputs and denote it with  $\text{Var}_{\mathbf{x}}[\cdot] := \text{Var}_{\mathbf{x} \sim p_{\mathcal{D}}}[\cdot]$  and  $\text{Var}_y[\cdot] := \text{Var}_{y \sim \text{Unif}(\mathcal{Y})}[\cdot]$ , respectively. With this in mind, it is useful to introduce the following functions:

$$\psi_{\mathbf{x}}(y_i; p) := \sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]} \quad \text{and} \quad \psi_y(\mathbf{x}_j; p) := \sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_0)]} . \quad (7)$$

We also denote with  $\mathbf{S} \in \mathbb{R}^{M \times M}$  the diagonal matrix with entries  $S_{ii} := \psi_{\mathbf{x}}(y_i; p)^{-1}$ . We now have everything we need to state a Lemma relating any two model embeddings (the full statement and proof, including the relation between the unembeddings can be found in [Appendix E.1](#)):

**Lemma 4.1.** *Let  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  be models satisfying the diversity condition ([Definition 2.1](#)). Let  $\mathbf{L}, \mathbf{L}'$  be defined as in [Theorem 2.2](#). Let  $\tilde{\mathbf{A}} := \mathbf{L}^{-\top} \mathbf{S}^{-1} \mathbf{S}' \mathbf{L}'^{\top}$  and  $\mathbf{h}_{\mathbf{f}}(\mathbf{x}) := \mathbf{L}^{-\top} \mathbf{S}^{-1} \boldsymbol{\epsilon}_y(\mathbf{x})$ , where the  $i$ -th entry of  $\boldsymbol{\epsilon}_y(\mathbf{x})$  is given by*

$$\epsilon_{y_i}(\mathbf{x}) = \frac{\log p_{\mathbf{f}, \mathbf{g}}(y_i|\mathbf{x}) - \log p_{\mathbf{f}, \mathbf{g}}(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}, \mathbf{g}})} - \frac{\log p_{\mathbf{f}', \mathbf{g}'}(y_i|\mathbf{x}) - \log p_{\mathbf{f}', \mathbf{g}'}(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}', \mathbf{g}'})} . \quad (8)$$

Then, we have

$$\mathbf{f}(\mathbf{x}) = \tilde{\mathbf{A}} \mathbf{f}'(\mathbf{x}) + \mathbf{h}_{\mathbf{f}}(\mathbf{x}) . \quad (9)$$

**Remark 4.2.** *In the special case where the two models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  have the same distribution, the error term in [Equation \(8\)](#) vanishes and we recover  $(\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}')$  (see [Corollary E.1](#)).*

We note that the error term in [Lemma 4.1](#) depends entirely on differences of log probabilities. We also see that if  $\epsilon_{y_i}(\mathbf{x})$  is a constant (or equivalently,  $\text{Var}[\epsilon_{y_i}(\mathbf{x})] = 0$ ), the embeddings will be invertible linear transformations of each other. This suggests that if we want a distance between distributions which is related to the similarity of the embeddings, we can measure how far this error is from being a constant. We therefore define a distance which includes these weighted differences of log-likelihoods.

To proceed, we restrict our analysis to a subset of distributions that satisfy the following assumption:

**Assumption 4.3.** *We consider probability distributions  $p$  such that for sets  $\mathcal{X}_{\text{LLV}} \subset \mathcal{X}$  and  $\mathcal{Y}_{\text{LLV}} \subset \mathcal{Y}$  containing all labels except one, the following conditions are satisfied: (1) For all  $\mathbf{x} \in \mathcal{X}_{\text{LLV}}$  we have that  $\log p(y|\mathbf{x}) - \log p(y|\mathbf{x}_0)$  is not constant in  $y$  and  $p_{\mathcal{D}}(\mathbf{x}) > 0$ ; and (2) For all  $y \in \mathcal{Y}_{\text{LLV}}$  we have that  $\log p(y|\mathbf{x}) - \log p(y_0|\mathbf{x})$  is not constant in  $\mathbf{x}$ .*

Note that this assumption guarantees that, for any such  $p$ , the terms  $\psi_{\mathbf{x}}(y_i; p)$ , and  $\psi_y(\mathbf{x}_j; p)$  are non-zero for  $\mathbf{x}_j \in \mathcal{X}_{\text{LLV}}$  and  $y_i \in \mathcal{Y}_{\text{LLV}}$ . Under [Assumption 4.3](#), we exclude those probability distributions that assign the same distributions over the labels for all inputs, that is, we need at least one input to result in a different distribution. We also exclude those distributions which assign equal probability to two or more labels for all inputs.

**Definition 4.4** (Log-likelihood variance distance between distributions). *For any two probability distributions  $p, p'$  for which there exists a common choice of  $\mathcal{X}_{\text{LLV}} \subset \mathcal{X}$  and  $\mathcal{Y}_{\text{LLV}} \subset \mathcal{Y}$  such that they both satisfy [Assumption 4.3](#), for a fixed  $\lambda \in \mathbb{R}^+$ , and by considering the following terms:*

$$\begin{aligned} t_1 &:= \max_{y \in \mathcal{Y}_{\text{LLV}} \setminus \{y_0\}} \left\{ \sqrt{\text{Var}_{\mathbf{x}} \left[ \frac{\log p(y|\mathbf{x})}{\psi_{\mathbf{x}}(y; p)} - \frac{\log p'(y|\mathbf{x})}{\psi_{\mathbf{x}}(y; p')} \right]}, \sqrt{\text{Var}_{\mathbf{x}} \left[ \frac{\log p(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y; p)} - \frac{\log p'(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y; p')} \right]} \right\} \\ t_2 &:= \max_{\mathbf{x} \in \mathcal{X}_{\text{LLV}} \setminus \{\mathbf{x}_0\}} \left\{ \sqrt{\text{Var}_y \left[ \frac{\log p(y|\mathbf{x})}{\psi_y(\mathbf{x}; p)} - \frac{\log p'(y|\mathbf{x})}{\psi_y(\mathbf{x}; p')} \right]}, \sqrt{\text{Var}_y \left[ \frac{\log p(y|\mathbf{x}_0)}{\psi_y(\mathbf{x}; p)} - \frac{\log p'(y|\mathbf{x}_0)}{\psi_y(\mathbf{x}; p')} \right]} \right\} \\ t_3 &:= \max_{y \in \mathcal{Y}_{\text{LLV}} \setminus \{y_0\}} |\psi_{\mathbf{x}}(y; p) - \psi_{\mathbf{x}}(y; p')|, \quad t_4 := \max_{\mathbf{x} \in \mathcal{X}_{\text{LLV}} \setminus \{\mathbf{x}_0\}} |\psi_y(\mathbf{x}; p) - \psi_y(\mathbf{x}; p')|, \end{aligned}$$

the log-likelihood variance (LLV) distance between  $p$  and  $p'$  is given by

$$d_{\text{LLV}}^{\lambda}(p, p') := \max \{t_1, t_2, \lambda t_3, \lambda t_4\} . \quad (10)$$

In [Appendix E.2](#), we show that  $d_{\text{LLV}}^{\lambda}$  is a distance metric between sets of conditional probability distributions.<sup>10</sup> The non-zero weighting constant  $\lambda$  is introduced because, as we will show in

<sup>10</sup>Specifically,  $d_{\text{LLV}}^{\lambda}(p, p')$  is non-negative; zero iff  $p = p'$ , symmetric; and it satisfies the triangle inequality.

**Theorem 4.7**,  $t_1$  and  $t_2$  can be used alone to bound the similarity between model representations, but  $t_3$  and  $t_4$  need also to be considered to make  $d_{\text{LLV}}^\lambda$  a distance metric. In our experiments, we set  $\lambda$  to a small non-zero value, so that the  $t_1$  and  $t_2$  terms dominate (see [Appendix E.3](#) for details).

Note that the log-likelihood variance distance,  $d_{\text{LLV}}^\lambda$ , requires a choice of input,  $\mathcal{X}_{\text{LLV}}$ , and label,  $\mathcal{Y}_{\text{LLV}}$ , sets, and the exact value of the distance depends on this choice when the distributions are not equal. However,  $d_{\text{LLV}}^\lambda$  is a metric for any choice satisfying [Assumption 4.3](#). Therefore in the experiments, we draw a sample of possible sets and choose the ones which give the smallest value.

## 4.2 A Dissimilarity Measure Between Representations

Having defined a distance between distributions, we turn to a distance between representations which is related to invertible linear transformations. We define a dissimilarity measure based on the partial least squares (PLS) approach called PLS-SVD [[54](#), [60](#)]. For two random vectors  $\mathbf{z}, \mathbf{w}$ , let  $\Sigma_{\mathbf{z}\mathbf{w}}$  denote the covariance matrix whose  $(i, j)$ -th entry is

$$(\Sigma_{\mathbf{z}\mathbf{w}})_{ij} = \text{Cov}_{\mathbf{z}\mathbf{w}}[z_i, w_j] := \mathbb{E}_{(\mathbf{z}, \mathbf{w}) \sim p_{\mathbf{z}, \mathbf{w}}} [(z_i - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}}[z_i])(w_j - \mathbb{E}_{\mathbf{w} \sim p_{\mathbf{w}}}[w_j])], \quad (11)$$

where  $p_{\mathbf{z}, \mathbf{w}}$  is the joint distribution of  $\mathbf{z}, \mathbf{w}$ , and  $p_{\mathbf{z}}, p_{\mathbf{w}}$  are the respective marginals. PLS-SVD seeks pairs of unit-length vectors  $\mathbf{u}_\ell \in \mathbb{R}^{d_{\mathbf{z}}}$  and  $\mathbf{v}_\ell \in \mathbb{R}^{d_{\mathbf{w}}}$  ( $\ell = 1, \dots, r$ ) that maximize the covariance between the one-dimensional projections  $\mathbf{u}_\ell^\top \mathbf{z}$  and  $\mathbf{v}_\ell^\top \mathbf{w}$ , subject to mutual orthogonality of successive directions. This is equivalent to finding the left and right singular vectors of  $\Sigma_{\mathbf{z}\mathbf{w}}$  (more about PLS-SVD in [Appendix D](#)). Leveraging this procedure by PLS-SVD, we introduce a similarity and a dissimilarity measure:

**Definition 4.5** (PLS SVD distance between vectors). *Let  $\mathbf{z}, \mathbf{w}$  be two  $M$ -dimensional random vectors, and define  $\mathbf{z}', \mathbf{w}'$  by standardizing their components:  $z'_i = (z_i - \mathbb{E}[z_i]) / \text{std}(z_i)$ ,  $w'_i = (w_i - \mathbb{E}[w_i]) / \text{std}(w_i)$ . Let  $\{\mathbf{u}_i\}_{i=1}^M$  and  $\{\mathbf{v}_i\}_{i=1}^M$  be the left and right singular vectors of the cross-covariance matrix  $\Sigma_{\mathbf{z}'\mathbf{w}'}$ . We define*

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) := \frac{1}{M} \sum_{i=1}^M \text{Cov}_{\mathbf{z}'\mathbf{w}'}[\mathbf{u}_i^\top \mathbf{z}', \mathbf{v}_i^\top \mathbf{w}'], \quad d_{\text{SVD}}(\mathbf{z}, \mathbf{w}) := 1 - m_{\text{SVD}}(\mathbf{z}, \mathbf{w}). \quad (12)$$

Since  $m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) \leq m_{\text{SVD}}(\mathbf{z}, \mathbf{z}) = 1$ , it follows that  $d_{\text{SVD}}(\mathbf{z}, \mathbf{w}) \geq 0$  for all random vectors  $\mathbf{z}, \mathbf{w}$ . We also have that the dissimilarity measure is invariant to rotation followed by dimension-wise scaling (see [Appendix E.4](#)): if  $m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) = 1$  or, equivalently,  $d_{\text{SVD}}(\mathbf{z}, \mathbf{w}) = 0$ , then there exist an orthonormal matrix  $\mathbf{R}$  and diagonal matrices  $\mathbf{S}, \mathbf{S}'$  scaling  $\mathbf{z}, \mathbf{w}$  to unit variance such that  $\mathbf{z} = \mathbf{S}^{-1} \mathbf{R} \mathbf{S}' \mathbf{w}$ .

With this distance between vectors, we can define a dissimilarity between representations for models from our model family by seeing the embeddings and unembeddings as random vectors.

**Definition 4.6** (Representational dissimilarity measure  $d_{\mathbf{f}, \mathbf{g}}$ ). *Let  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$ . Let  $\mathbf{L}, \mathbf{L}'$  be defined as in [Theorem 2.2](#). Let  $\mathbf{N}$  (resp.  $\mathbf{N}'$ ) be the matrix with columns  $\mathbf{f}_0(\mathbf{x})$  (resp.  $\mathbf{f}'_0(\mathbf{x})$ ). The representational dissimilarity measure  $d_{\mathbf{f}, \mathbf{g}}$  is defined as follows:*

$$d_{\mathbf{f}, \mathbf{g}}((\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}')) := \max \left\{ d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})), d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) \right\}. \quad (13)$$

## 4.3 Bounding Representation Dissimilarity via Distributional Distance

We now prove that it is possible to relate a bound on the distance metric  $d_{\text{LLV}}^\lambda$  between model distributions ([Definition 4.4](#)) to a bound on the dissimilarity measure  $d_{\mathbf{f}, \mathbf{g}}$  between model representations ([Definition 4.6](#)). For the result below, We denote with  $\mathbf{z}_1 := \mathbf{L}^\top \mathbf{f}(\mathbf{x})$ ,  $\mathbf{z}_2 := \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})$ ,  $\mathbf{w}_1 := \mathbf{N}^\top \mathbf{g}(y)$  and  $\mathbf{w}_2 := \mathbf{N}'^\top \mathbf{g}'(y)$  (proof in [Appendix E.6](#)).

**Theorem 4.7.** *Let  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  be two models such that: (1) There exist  $\mathcal{X}_{\text{LLV}} \subset \mathcal{X}$  and  $\mathcal{Y}_{\text{LLV}} \subset \mathcal{Y}$ , consisting of a pivot point and all labels but one, such that all  $\mathbf{L}, \mathbf{L}'$  and  $\mathbf{N}, \mathbf{N}'$  matrices constructed from these sets are invertible;<sup>11</sup> (2) Both  $p_{\mathbf{f}, \mathbf{g}}$  and  $p_{\mathbf{f}', \mathbf{g}'}$  satisfy [Assumption 4.3](#) for  $\mathcal{X}_{\text{LLV}}$  and  $\mathcal{Y}_{\text{LLV}}$ ; (3) The covariance matrices  $\Sigma_{\mathbf{z}_1 \mathbf{z}_1}$ ,  $\Sigma_{\mathbf{w}_1 \mathbf{w}_1}$  and the cross-covariance matrices  $\Sigma_{\mathbf{z}_1 \mathbf{z}_2}$ ,  $\Sigma_{\mathbf{w}_1 \mathbf{w}_2}$  are non-singular. Then, for any weighting constant  $\lambda \in \mathbb{R}^+$ , we have*

$$d_{\text{LLV}}^\lambda(p_{\mathbf{f}, \mathbf{g}}, p_{\mathbf{f}', \mathbf{g}'}) \leq \epsilon \implies d_{\mathbf{f}, \mathbf{g}}((\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}')) \leq 2M\epsilon. \quad (14)$$

<sup>11</sup>This is slightly stronger than the diversity condition ([Definition 2.1](#)) in the sense that we need diversity to hold for all sets using labels from  $\mathcal{Y}_{\text{LLV}}$  and the chosen pivot point.

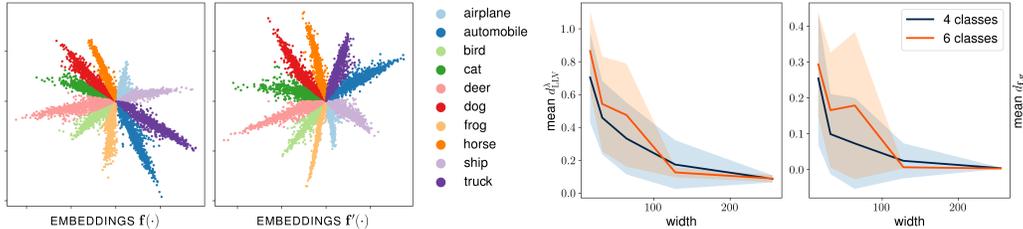


Figure 3: (Left) Embedding representations of two models trained on CIFAR-10. Representations for some of the labels are permuted, and  $m_{CCA}(\mathbf{f}(\mathbf{x}), \mathbf{f}'(\mathbf{x})) = 0.55$ . (Right) Mean  $d_{LLV}^\lambda$  and  $d_{\mathbf{f},\mathbf{g}}$  vs network width. Shaded area is standard deviation. Both mean and standard deviation decrease as the network width increases.

**Remark 4.8.** It can be shown (Lemma E.8) that  $d_{SVD}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x}))$  remains constant for models that are linearly equivalent to the members in the expression. That is, for any  $(\mathbf{f}^*, \mathbf{g}^*) \in \Theta$ , such that  $(\mathbf{f}^*, \mathbf{g}^*) \sim_L (\mathbf{f}, \mathbf{g})$ , we have that  $d_{SVD}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) = d_{SVD}(\mathbf{L}^{*\top} \mathbf{f}^*(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x}))$ . A similar argument also holds for  $d_{SVD}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y))$ .

Theorem 4.7 shows that for the measures  $d_{LLV}^\lambda$  and  $d_{\mathbf{f},\mathbf{g}}$ , closeness in distribution (i.e., small values of  $d_{LLV}^\lambda$ ) bounds representational dissimilarity (as measured by  $d_{\mathbf{f},\mathbf{g}}$ ).

Notice that in Theorem 4.7, we can freely choose any set of inputs  $\mathcal{X}_{LLV} \subset \mathcal{X}$  as long as they satisfy (1) and (2). Hence, if there is a choice making  $t_2$  in Definition 4.4 as small as possible, it will also make  $d_{SVD}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y))$  small. There is also some flexibility for  $\mathcal{Y}_{LLV} \subset \mathcal{Y}$ , where a careful choice of pivot point  $y_0$  and the left-out label can make the term  $t_1$  in Definition 4.4 small and, as a consequence,  $d_{SVD}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x}))$  also small.

## 5 Experiments

In this section, we present our key experimental findings: (1) For models trained for classification on CIFAR-10 [32], we observe cases of similarly well-performing models with highly dissimilar representations, arising from a mechanism analogous to the constructive proof of Theorem 3.1 (Section 5.1); (2) On synthetic data, training wider neural networks results in a reduction of both the mean and variance of  $d_{LLV}^\lambda$  between models and leads to higher representational similarity (Section 5.2). Additional experiments illustrating the phenomena in Theorem 3.1 and visualizing the bound in Theorem 4.7 are summarized in Section 5.3. For all experiments we use  $\lambda = 10^{-5}$ . See Appendix F for implementation details. The code can be found on GitHub.<sup>12</sup>

### 5.1 Dissimilar Representations in CIFAR-10 Models with Similar Performance

**Experimental setup.** We trained classification models on CIFAR-10 [32] with two-dimensional embedding and unembedding representations. The embedding network is a ResNet18 [21], and the unembedding network consists of three fully connected layers of width 128, followed by the final representation layer.

**Results.** Embeddings of images with the same labels form clusters, and certain label pairs—such as “truck” and “automobile”—consistently appear as neighbors across retrainings (See Appendix F.5). However, some clusters are permuted: Fig. 3 (Left) shows two retrainings where, in the first model, the “truck” cluster lies between “automobile” and “ship”, while in the second it lies between “automobile” and “horse”. This mirrors the construction in Theorem 3.1, where models close in KL divergence have dissimilar representations. The models in Fig. 3 (Left) have  $d_{KL}$  of 1.13 (0→1) and 1.12 (1→0), similar test losses (1.19 and 1.18), but a large  $d_{LLV}^\lambda$  ( $\sim 1.55$ ) and small  $m_{CCA}$  (0.55).

<sup>12</sup>[github.com/bemigini/close-dist-rep-sim](https://github.com/bemigini/close-dist-rep-sim), DOI: 10.5281/zenodo.17249361.

## 5.2 Wider Networks Learn Closer Distributions and More Similar Representations

**Experimental setup.** We train models on a 2D classification task with 4, 6, 10 or 18 labels. Data are samples from a 2D Gaussian ( $\mu = 0, \sigma = 3$ ), and labels are based on the angle to the first axis, consisting of a slice of the circle and the opposite slice (see Appendix F.2). We train twenty random seeds of models with two-dimensional representations. Each model consists of three fully connected layers with widths 16, 32, 64, 128 or 256. We retain only models achieving over 90% accuracy. For 4 classes, this yields 19 models at width 16 and 20 models for all other widths. For 6 classes, we retain 16, 17, 19, 19, and 20 models across the five widths, respectively.

**Results.** We find that  $d_{LLV}^\lambda$  and  $d_{f,g}$  between the models trained with wider networks have smaller both mean and standard deviation. In Fig. 3 (Right) we show the trend for models with 4 and 6 classes. Similar plots for 10 and 18 classes can be seen in Appendix F.7. While the bound in Theorem 4.7 is only non-vacuous when  $d_{f,g}((f, g), (f', g')) \leq 2M\epsilon < 1$  (i.e., when  $\epsilon < 1/2M$ ) our plots suggest that there is a broader relationship between  $d_{f,g}$  and  $d_{LLV}^\lambda$ , since the values of  $d_{LLV}^\lambda$  and  $d_{f,g}$  seem to be related even before the bound becomes non-vacuous.

## 5.3 Visualizing Representation Dissimilarity and the Bound from Theorem 4.7

Following the construction in Theorem 3.1, we simulate two models with fixed embeddings and unembeddings with increasing norm. We measure the KL divergence and  $d_{LLV}^\lambda$  between distributions; between representations, we measure the mean canonical correlation ( $m_{CCA}$ ) [45, 49, 52] and the maximum  $d_{SVD}$  of the embeddings over the input and label sets for  $d_{LLV}^\lambda$  (since this is what Theorem 4.7 bounds). The results are in Table 1: We find that as the unembeddings’ norm,  $\rho$ , increases,  $d_{KL}$  decreases, while  $d_{LLV}^\lambda$  and  $d_{SVD}$  remain high and  $m_{CCA}$  remains low and almost constant. This aligns with the result in Theorem 3.1, while it displays that  $d_{LLV}^\lambda$  is robust to these model constructions.

Also, for these model pairs and models trained on synthetic data, we find that the bound in Theorem 4.7 is always satisfied (refer to Appendix F.6 for details and to Fig. 16 for a visualization). However, for models trained on CIFAR-10, the distributions are not close enough for the bound to be non-vacuous.

Table 1: **Kullback–Leibler divergence rapidly vanishes while other measures stay constant.** Cells in the  $d_{KL}$  column are shaded red in proportion to their magnitude (darker  $\Rightarrow$  larger value). As the unembedding norm  $\rho$  grows,  $d_{KL}$  drops by almost four orders of magnitude, whereas  $d_{LLV}^\lambda$  and the maximal  $d_{SVD}$  between embeddings remain virtually unchanged.

$\rho$	$d_{KL}$	$d_{LLV}^\lambda$	$m_{CCA}$	$\max d_{SVD}$
3	0.8866	1.3176	0.0006	0.9996
6	0.2230	1.3169	0.0007	0.9998
9	0.0369	1.3171	0.0004	0.9995
12	0.0055	1.3178	0.0006	0.9998
15	0.0008	1.3175	0.0008	0.9999
18	0.0001	1.3172	0.0010	0.9998

## 6 Discussion

**Limitations.** To the best of our knowledge, this work provides the first theoretical result that upper-bounds representational dissimilarity by a distributional distance (Theorem 4.7)—a relationship that Theorem 3.1 shows is far from obvious. The bound is nonetheless not tight: it grows linearly even though the empirical trend of  $\max d_{SVD}$  is clearly sub-linear (Fig. 16), and it presently requires the label set to contain every class but one. We conjecture that a tighter, potentially non-linear bound could be obtained and that  $d_{LLV}^\lambda$  would remain a valid distance when computed on a suitably diverse subset of labels. The distributional distance itself (Definition 4.4) inherits the same label-diversity assumption (Definition 2.1), which prevents its applicability to the whole model family of Equation (1): diversity is more likely to hold in language models, but may fail in standard image classifiers—especially when the number of classes is smaller than the representation dimension [52]. It was recently proved by Marconato et al. [43] that, when relaxing the diversity assumption, the model family in Equation (1) can be identified up to an extended-linear equivalence relation. Extending our theory to that setting would enable comparing models with different representation dimension [43].

**Alternative similarity measures.** The invariances of our dissimilarity measure,  $d_{f,g}$ , comprise a subset of all invertible linear transforms and always consider embeddings and unembeddings in combination with each other (see Appendix E.7). By contrast, much prior work employs CCA-based measures that are invariant to any linear transformation [45, 49] and only measures (dis)similarity

between embeddings. Kornblith et al. [31] argue that similarity measures should be invariant only to orthogonal transformations, based on the task-relevance of relative activation scales and the invariance of gradient-descent dynamics [35]. What measure best captures representational similarity is still an open question [3], with different approaches reflecting distinct goals and trade-offs [30]. Our goal is not to claim our measure is *the* right one, but to present a pairing of dissimilarity measure and distributional distance that enables theoretical analysis and captures key empirical aspects of representational similarity. Extending such guarantees to other (dis)similarity measures and distributional divergences is an important direction for future work.

**Do similarly retrained models learn similar representations?** This has been extensively explored in empirical studies [31, 38, 46, 51, 52, 56, 65]. Retraining with a different seed can result in different likelihoods, and we show that even models with near-zero KL divergence can have dissimilar internal representations. While the existence of constructions such as the one in [Theorem 3.1](#) does not imply that a model will learn them during training, we find it interesting that a mechanism resembling our construction actually emerged in our CIFAR-10 experiments ([Section 5.1](#)), where we observe large representational dissimilarity despite close distributions in terms of KL divergence. Such dissimilarity can arise when most output labels have negligible probability—for example, an image might be confidently classified as a “cat” or “dog”, with all other classes (e.g., “spaceship”, “chair”) having near-zero probabilities; and in language modeling, often only few next-tokens are plausible. Overall, these observations suggest that similar performance alone is insufficient to ensure representational similarity [6], though higher network capacity may help, see [Section 5.2](#) and [52, 56].

**Identifiability and robustness.** In nonlinear ICA [10, 18, 20, 24, 28, 57, 72], disentangled and causal representation learning [2, 7, 11, 37, 39, 40, 44, 63, 64, 69–71] models are typically constructed so that their likelihood-maximising solutions are unique up to a fixed (and ‘small’) set of ambiguities, as guaranteed by identifiability [24, 68]. Robustness when the likelihood is only approximately maximized has been studied less, though Buchholz and Schölkopf [9] analyzed isometric-mixing ICA; Sliwa et al. [55] empirically probed *independent mechanism analysis* [19]; and Träuble et al. [62] showed that independence-based disentanglement lacks robustness when true factors are correlated. Our results further highlight the importance of understanding robustness in nonlinear representation learning.<sup>13</sup> In particular, our [Corollary 3.2](#) suggests that relying solely on identifiability results—such as those by Reizinger et al. [51]—may be insufficient to guarantee that similar representations are practically attainable when optimizing the cross-entropy loss: additional assumptions may be needed.

**Conclusion and Outlook.** We have studied the link between distributional closeness and representational similarity for the embedding and unembedding layers of our model family, focusing on different instances of the *same* model class. Extending our analysis to earlier layers would first require identifiability results for intermediate network representations, which remains an open problem in representation learning [25, 28]. In practice, learned representations often appear surprisingly robust across datasets, architectures, and training objectives [3, 4, 17, 23, 31, 46], and explaining this broader robustness is an exciting direction for future work. By clarifying when distributional closeness does—and does not—imply representational similarity for embeddings and unembeddings within our model family, we take an initial step toward a principled theory of representational similarity.

## Acknowledgments and Disclosure of Funding

We thank Simon Buchholz for helpful discussions and feedback. Beatrix M. G. Nielsen was supported by the Danish Pioneer Centre for AI, DNRF grant number P1. E.M. acknowledges support from TANGO, Grant Agreement No. 101120763, funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. A.D. acknowledges funding from the Helmholtz Foundation Model Initiative. L.G. was supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429).

<sup>13</sup> $\epsilon$ -non-identifiability [50] is closely related: our work illustrates it for learned representations and highlights that it depends on the choice of divergence—holding in  $d_{\text{KL}}$  ([Theorem 3.1](#)) but not in  $d_{\text{LLV}}^{\lambda}$  ([Theorem 4.7](#)).

## References

- [1] Ani Adhikari and Jim Pitman. *Probability for Data Science*. UC Berkeley, 2025. URL <https://textbook.prob140.org/>. [Cited on page 43.]
- [2] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, pages 372–407. PMLR, 2023. [Cited on page 10.]
- [3] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 225–236, 2021. [Cited on pages 1 and 10.]
- [4] Boaz Barak, Annabelle Carrell, Alessandro Favero, Weiyu Li, Ludovic Stephan, and Alexander Zlokapa. Computational complexity of deep learning: fundamental limitations and empirical phenomena. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104008, 2024. [Cited on page 10.]
- [5] Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. *arXiv preprint arXiv:2502.11245*, 2025. [Cited on page 1.]
- [6] Lukas Braun, Erin Grant, and Andrew M. Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *International Conference on Machine Learning (ICML)*, 2025. [Cited on page 10.]
- [7] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 38319–38331, 2022. [Cited on page 10.]
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 1877–1901, 2020. [Cited on pages 2 and 3.]
- [9] Simon Buchholz and Bernhard Schölkopf. Robustness of nonlinear representation learning. In *International Conference on Machine Learning (ICML)*, pages 4785–4821. PMLR, 2024. [Cited on page 10.]
- [10] Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:16946–16961, 2022. [Cited on page 10.]
- [11] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 45419–45462, 2023. [Cited on page 10.]
- [12] Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, Emanuele Rodolà, et al. From bricks to bridges: Product of invariances to enhance latent space communication. In *International Conference on Learning Representations (ICLR)*, 2024. [Cited on page 1.]
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*, pages 4171–4186, 2019. [Cited on page 3.]
- [14] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 1556–1568, 2021. [Cited on page 1.]

- [15] Hidde Fokkema, Tim van Erven, and Sara Magliacane. Sample-efficient learning of concepts with theoretical guarantees: from data to concepts without interventions. *arXiv preprint arXiv:2502.06536*, 2025. [Cited on page 1.]
- [16] Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Cited on page 1.]
- [17] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. *arXiv preprint arXiv:2010.08377*, 2020. [Cited on page 10.]
- [18] Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Uncertainty in Artificial Intelligence (UAI)*, pages 217–227. PMLR, 2020. [Cited on page 10.]
- [19] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems (NeurIPS)*, 34:28233–28248, 2021. [Cited on pages 2 and 10.]
- [20] Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Uncertainty in Artificial Intelligence (UAI)*, pages 939–948. PMLR, 2020. [Cited on page 10.]
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. [Cited on pages 8 and 49.]
- [22] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. [Cited on pages 4 and 35.]
- [23] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. [Cited on page 10.]
- [24] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 859–868. PMLR, 2019. [Cited on page 10.]
- [25] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024. [Cited on page 10.]
- [26] Mark Ibrahim, David Klindt, and Randall Balestriero. Occam’s razor for self supervised learning: What is sufficient to learn good representations? In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*, 2024. [Cited on page 3.]
- [27] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2207–2217. PMLR, 2020. [Cited on page 2.]
- [28] Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12768–12778, 2020. [Cited on pages 2, 3, 4, 10, and 25.]
- [29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. [Cited on pages 49 and 50.]
- [30] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. [Cited on pages 1, 4, and 10.]

- [31] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, pages 3519–3529. PMLR, 2019. [Cited on pages 1 and 10.]
- [32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. [Cited on pages 8, 49, and 54.]
- [33] Aarre Laakso and Garrison Cottrell. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76, 2000. [Cited on page 1.]
- [34] Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning (ICML)*, pages 18171–18206. PMLR, 2023. [Cited on pages 2, 3, 4, and 25.]
- [35] Yann LeCun, Ido Kanter, and Sara Solla. Second order properties of error surfaces: Learning time and generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 3, 1990. [Cited on page 10.]
- [36] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–999, 2015. [Cited on page 1.]
- [37] Adam Li, Yushu Pan, and Elias Bareinboim. Disentangled representation learning in non-markovian causal systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Cited on page 10.]
- [38] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Feature Extraction: Modern Questions and Challenges*, pages 196–212. PMLR, 2015. [Cited on pages 1 and 10.]
- [39] Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal Component Analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. [Cited on page 10.]
- [40] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning (ICML)*, pages 13557–13603. PMLR, 2022. [Cited on page 10.]
- [41] Valentino Maiorca, Luca Moschella, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via inverse relative projection. *arXiv preprint arXiv:2406.15057*, 2024. [Cited on page 1.]
- [42] Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25(12):1574, 2023. [Cited on page 1.]
- [43] Emanuele Marconato, Sébastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or none: Identifiable linear properties of next-token predictors in language modeling. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. [Cited on pages 3, 9, and 25.]
- [44] Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. [Cited on page 10.]
- [45] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. [Cited on pages 1, 4, and 9.]
- [46] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodola. Relative representations enable zero-shot latent space communication. *International Conference on Learning Representations (ICLR)*, 2023. [Cited on pages 1 and 10.]

- [47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [Cited on pages 2 and 3.]
- [48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. [Cited on pages 2 and 3.]
- [49] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. [Cited on pages 1, 4, 5, and 9.]
- [50] Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Position: Understanding LLMs requires more than statistical generalization. In *International Conference on Machine Learning (ICML)*, pages 42365–42390. PMLR, 2024. [Cited on page 10.]
- [51] 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. In *International Conference on Learning Representations (ICLR)*, 2025. [Cited on pages 2 and 10.]
- [52] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning (ICML)*, pages 9030–9039. PMLR, 2021. [Cited on pages 2, 3, 4, 5, 9, 10, and 25.]
- [53] Aninda Saha, Alina Bialkowski, and Sara Khalifa. Distilling representational similarity using centered kernel alignment (CKA). In *Proceedings of the the 33rd British Machine Vision Conference (BMVC 2022)*. British Machine Vision Association, 2022. [Cited on page 1.]
- [54] Paul D Sampson, Ann P Streissguth, Helen M Barr, and Fred L Bookstein. Neurobehavioral effects of prenatal alcohol: Part II. Partial least squares analysis. *Neurotoxicology and teratology*, 11(5):477–491, 1989. [Cited on pages 7 and 35.]
- [55] Joanna Sliwa, Shubhangi Ghosh, Vincent Stimper, Luigi Gresele, and Bernhard Schölkopf. Probing the robustness of independent mechanism analysis for representation learning. In *1st Workshop on Causal Representation Learning at the 38th Conference on Uncertainty in Artificial Intelligence (UAI CRL 2022)*. OpenReview. net, 2022. [Cited on page 10.]
- [56] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13689–13698. Institute of Electrical and Electronics Engineers, 2022. [Cited on page 10.]
- [57] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). *arXiv preprint arXiv:2001.04872*, 2020. [Cited on page 10.]
- [58] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in Neural Information Processing Systems (NeurIPS)*, 34:6906–6919, 2021. [Cited on page 1.]
- [59] Gilbert W Stewart. *Perturbation theory for the singular value decomposition*. Citeseer, 1998. [Cited on page 42.]
- [60] Ann Pytkowicz Streissguth, Fred L Bookstein, Paul D Sampson, and Helen M Barr. *The enduring effects of prenatal alcohol exposure on child development: Birth through seven years, a partial least squares solution*. The University of Michigan Press, 1993. [Cited on pages 7 and 35.]
- [61] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023. [Cited on page 1.]

- [62] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *International Conference on Machine Learning (ICML)*, pages 10401–10412. PMLR, 2021. [Cited on page 10.]
- [63] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2314–2322. PMLR, 2024. [Cited on page 10.]
- [64] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. [Cited on page 10.]
- [65] Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. [Cited on pages 1 and 10.]
- [66] Jacob A Wegelin et al. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. *University of Washington, Tech. Rep.*, 2000. [Cited on page 35.]
- [67] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021. [Cited on page 1.]
- [68] Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6912–6939. PMLR, 2023. [Cited on page 10.]
- [69] Danru Xu, Dingling Yao, Sebastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *International Conference on Machine Learning (ICML)*, pages 55389–55433. PMLR, 2024. [Cited on page 10.]
- [70] Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024.
- [71] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. [Cited on page 10.]
- [72] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:16411–16422, 2022. [Cited on page 10.]
- [73] Zikai Zhou, Yunhang Shen, Shitong Shao, Linrui Gong, and Shaohui Lin. Rethinking centered kernel alignment in knowledge distillation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5680–5688, 2024. [Cited on page 1.]
- [74] Martin Zong, Zengyu Qiu, Xinzhu Ma, Kunlin Yang, Chunya Liu, Jun Hou, Shuai Yi, and Wanli Ouyang. Better teacher better student: Dynamic prior knowledge for knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2023. [Cited on page 1.]

## 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: We list our three main claims in [Section 1](#) and they are mainly supported by [Section 3](#), [Section 4](#) and [Section 5](#).

Guidelines:

- The answer NA 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 NA 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: We discuss limitations in [Section 6](#).

Guidelines:

- The answer NA 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: Some proof ideas and sketches are in the main article. Full proofs are in the appendix due to size constraints. Proofs for [Section 2](#) are in [Appendix B](#), proofs for [Section 3](#) are in [Appendix C](#) and proofs for [Section 4](#) are in [Appendix E](#).

Guidelines:

- The answer NA 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: We give implementation details of our distance in [Appendix E.3](#) and details on how to recreate the experiments in [Appendix F](#). All the relevant code is available at [github](#)<sup>14</sup>.

Guidelines:

- The answer NA 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.

---

<sup>14</sup>[github.com/bemigini/close-dist-rep-sim](https://github.com/bemigini/close-dist-rep-sim)

## 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 will be released on publication of the article.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.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://nips.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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In [Section 5](#) we provide the necessary details to understand the results and in [Appendix F](#) and [Appendix H](#) we provide the details important for implementation. The code is also available at github, [github.com/bemigini/close-dist-rep-sim](https://github.com/bemigini/close-dist-rep-sim).

Guidelines:

- The answer NA 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: [No]

Justification: We do not have error bars in our plots.

Guidelines:

- The answer NA 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: We report computing resources in [Appendix G](#).

Guidelines:

- The answer NA 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 conforms with the principles outlined in the NeurIPS Code of Ethics.

Guidelines:

- The answer NA 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: [NA]

Justification: Our work focuses on improving our understanding a popular class of discriminative models including autoregressive language models. We hope our work can lead to safer deployment of such models in the future, however for now this is fundamental research that has no direct societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release anything with a high risk of misuse.

Guidelines:

- The answer NA 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: We cite original creators when relevant. Available licence information is in [Appendix H](#).

Guidelines:

- The answer NA 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](https://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: [Yes]

Justification: We describe how to generate the synthetic data used in [Appendix F.2](#). Our code also includes the necessary documentation.

Guidelines:

- The answer NA 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: [NA]

Justification: The work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA 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: [NA]

Justification: The work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs have not been used for any core components of this research.

Guidelines:

- The answer NA 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 (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

# Appendix

## Table of Contents

---

<b>A</b>	<b>Notation</b>	<b>24</b>
<b>B</b>	<b>Identifiability Result and Proof</b>	<b>25</b>
<b>C</b>	<b>Models close in KL divergence can have dissimilar representations - details</b>	<b>27</b>
C.1	KL goes to Zero but Representations are Dissimilar . . . . .	27
C.2	Loss goes to Zero but Representations are Dissimilar . . . . .	33
<b>D</b>	<b>Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA)</b>	<b>35</b>
D.1	PLS-SVD . . . . .	35
D.2	Canonical Correlation Analysis . . . . .	35
<b>E</b>	<b>When Closeness in Distribution Implies Representational Similarity - Details</b>	<b>36</b>
E.1	The Connection Between Representations and Log-Likelihoods . . . . .	36
E.2	Distance Between Distributions . . . . .	39
E.3	Implementation of the Log-Likelihood Variance Distance . . . . .	41
E.4	Using PLS SVD to Define a Dissimilarity Measure . . . . .	41
E.5	Lower Bound on Mean PLS-SVD . . . . .	42
E.6	Bounding Representational Similarity with Distribution Distance . . . . .	44
E.7	Invariances of our Representational Distance and CCA . . . . .	48
<b>F</b>	<b>Experimental Details</b>	<b>49</b>
F.1	Constructed Models . . . . .	49
F.2	Models Trained on Synthetic Data . . . . .	49
F.3	Models Trained on CIFAR-10 . . . . .	49
F.4	Loss Difference vs Embedding $m_{CCA}$ . . . . .	50
F.5	All Two-dimensional Representations of CIFAR-10 Models . . . . .	50
F.6	Illustration of Bound on Constructed and Trained Models . . . . .	53
F.7	Wider Models Have more Similar Distributions - Extra Plots . . . . .	53
<b>G</b>	<b>Computing Resources</b>	<b>54</b>
<b>H</b>	<b>Assets</b>	<b>54</b>

---

## A Notation

We here present a table (Table 2) with the notation used in the paper.

Table 2: **Notation.** We use bold letters for vectors and non-bold letters for scalars.

Symbol	Description	Reference
$\mathbf{x}$	vector and vector-valued variables	Section 2
$y$	scalar-valued variable	Section 2
$(\mathbf{f}, \mathbf{g})$	a model with embedding and unembedding functions $\mathbf{f}$ and $\mathbf{g}$ , see Equation (1)	Section 2
$p_{\mathbf{f}, \mathbf{g}}$	distributions of a model $(\mathbf{f}, \mathbf{g})$	Section 2
$\mathbf{f}_0(\mathbf{x})$	the displaced embedding functions $\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)$	Section 2
$\mathbf{g}_0(y)$	the displaced unembedding functions $\mathbf{g}(y) - \mathbf{g}(y_0)$	Section 2
$\mathbf{L}$	the matrix with columns $\mathbf{g}_0(y)$ , see Theorem 2.2	Section 2
$d_{\text{KL}}$	the distance between distributions arising from the KL divergence	Section 3
$\text{Var}_{\mathbf{x}}[\cdot]$	the variance over the inputs from some input distribution, $p_{\mathbf{x}}$	Section 4
$\text{Unif}(\mathcal{Y})$	the uniform distribution over the elements of $\mathcal{Y}$	Section 4
$d_{\text{LLV}}^{\lambda}$	the log-likelihood variance distance with weighting parameter $\lambda$	Section 4
$d_{\text{SVD}}$	the mean PLS-SVD distance between random vectors	Section 4
$d_{\mathbf{f}, \mathbf{g}}$	the distance between representations for our model class	Section 4
$\Sigma_{\mathbf{z}\mathbf{w}}$	the cross-covariance matrix of random vectors $\mathbf{z}, \mathbf{w}$	Section 4
$\mathbf{N}$	the matrix with columns $\mathbf{f}_0(\mathbf{x})$	Section 4
$m_{\text{CCA}}$	the mean canonical correlation coefficient	Appendix D.2

## B Identifiability Result and Proof

We present a unified and slightly adapted version of the identifiability results from [28, 34, 52], see also [43, Corollary 6].

**Theorem 2.2** (Linear Identifiability [28, 34, 52]). *Let  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$ , and  $(\mathbf{f}, \mathbf{g})$  satisfy the diversity condition (Definition 2.1). Let  $\mathbf{L}$  (resp.  $\mathbf{L}'$ ) be the matrix with columns  $\mathbf{g}_0(y)$  (resp.  $\mathbf{g}'_0(y)$ ), and let  $\mathbf{A} := \mathbf{L}^{-\top} \mathbf{L}'^{\top} \in \mathbb{R}^{M \times M}$ . Then:*

$$p_{\mathbf{f}, \mathbf{g}}(y | \mathbf{x}) = p_{\mathbf{f}', \mathbf{g}'}(y | \mathbf{x}), \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \implies (\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}') \quad (4)$$

where the equivalence relation  $\sim_L$  is defined by

$$(\mathbf{f}, \mathbf{g}) \sim_L (\mathbf{f}', \mathbf{g}') \iff \begin{cases} \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{f}'(\mathbf{x}) \\ \mathbf{g}_0(y) = \mathbf{A}^{-\top} \mathbf{g}'_0(y) \end{cases} . \quad (5)$$

*Proof.* We first prove that  $p_{\mathbf{f}, \mathbf{g}}(y | \mathbf{x}) = p_{\mathbf{f}', \mathbf{g}'}(y | \mathbf{x}), \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \implies \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{f}'(\mathbf{x})$  for  $\mathbf{A} := \mathbf{L}^{-\top} \mathbf{L}'^{\top}$  invertible.

By assumption, the models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}')$  have equal likelihoods. Moreover, by construction, the two models have representations in  $\mathbb{R}^M$ . Let  $Z(\mathbf{x}, \mathcal{Y}) = \sum_{y_j \in \mathcal{Y}} \exp(\mathbf{f}(\mathbf{x})^{\top} \mathbf{g}(y_j))$ , and similarly for  $Z'(\mathbf{x}, \mathcal{Y})$ . Then

$$p_{\mathbf{f}, \mathbf{g}}(y | \mathbf{x}) = p_{\mathbf{f}', \mathbf{g}'}(y | \mathbf{x}) \quad (15)$$

$$\mathbf{f}(\mathbf{x})^{\top} \mathbf{g}(y) - \log(Z(\mathbf{x}, \mathcal{Y})) = \mathbf{f}'(\mathbf{x})^{\top} \mathbf{g}'(y) - \log(Z'(\mathbf{x}, \mathcal{Y})) \quad (16)$$

for all  $\mathbf{x} \in \mathcal{X}$  and all  $y \in \mathcal{Y}$ . In particular, this equation holds for any fixed  $y$ . From the diversity condition (Definition 2.1), we can choose  $M + 1$   $y$ 's,  $y_0, \dots, y_M$ , such that the displaced unembeddings  $\{\mathbf{g}_0(y_i)\}_{i=1}^M$  are linearly independent. By subtracting the log-conditional distribution for the pivot  $y_0 \in \mathcal{Y}$ , we obtain for each  $y_i \in \mathcal{Y}$  the following:

$$\begin{aligned} & \mathbf{f}(\mathbf{x})^{\top} \mathbf{g}(y_i) - \mathbf{f}(\mathbf{x})^{\top} \mathbf{g}(y_0) + \log(Z(\mathbf{x}, \mathcal{Y})) - \log(Z(\mathbf{x}, \mathcal{Y})) \\ &= \mathbf{f}'(\mathbf{x})^{\top} \mathbf{g}'(y_i) - \mathbf{f}'(\mathbf{x})^{\top} \mathbf{g}'(y_0) + \log(Z'(\mathbf{x}, \mathcal{Y})) - \log(Z'(\mathbf{x}, \mathcal{Y})) \end{aligned} \quad (17)$$

$$\mathbf{f}(\mathbf{x})^{\top} \mathbf{g}(y_i) - \mathbf{f}(\mathbf{x})^{\top} \mathbf{g}(y_0) = \mathbf{f}'(\mathbf{x})^{\top} \mathbf{g}'(y_i) - \mathbf{f}'(\mathbf{x})^{\top} \mathbf{g}'(y_0) \quad (18)$$

$$\mathbf{f}(\mathbf{x})^{\top} \mathbf{g}_0(y_i) = \mathbf{f}'(\mathbf{x})^{\top} \mathbf{g}'_0(y_i), \quad (19)$$

where the last passage holds by definition of  $\mathbf{g}_0(y), \mathbf{g}'_0(y)$ , see Section 2. Let  $\mathbf{L}$  be the matrix which has  $\mathbf{g}_0(y_i)$  as columns and  $\mathbf{L}'$  be the matrix which has  $\mathbf{g}'_0(y_i)$  as columns. Notice that the diversity condition (Definition 2.1) implies that  $\mathbf{L}$  is an invertible matrix. We can then stack the equations to get

$$\mathbf{L}^{\top} \mathbf{f}(\mathbf{x}) = \mathbf{L}'^{\top} \mathbf{f}'(\mathbf{x}) \quad (20)$$

and since  $\mathbf{L}$  is invertible,

$$\mathbf{f}(\mathbf{x}) = \mathbf{L}^{-\top} \mathbf{L}'^{\top} \mathbf{f}'(\mathbf{x}). \quad (21)$$

If we set  $\mathbf{A} := (\mathbf{L}'/\mathbf{L}^{-1})^{\top}$ , we only need to show that  $\mathbf{A}$  is invertible. Using the diversity condition, we pick points  $\mathbf{x}_0, \dots, \mathbf{x}_M \in \mathcal{X}$  such that the displaced embedding vectors  $\{\mathbf{f}_0(\mathbf{x}_i)\}_{i=1}^M$  are linearly independent. Let  $\mathbf{N}$  be the matrix with  $\mathbf{f}_0(\mathbf{x}_i)$  as columns and  $\mathbf{N}'$  be the matrix with  $\mathbf{f}'_0(\mathbf{x}_i)$  as columns. Notice that, from the diversity condition  $\mathbf{N}$  is an invertible matrix. Then

$$\mathbf{N} = \mathbf{A}\mathbf{N}' . \quad (22)$$

Since any two matrices,  $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{M \times M}$ , have the property  $\text{rank}(\mathbf{BC}) \leq \min(\text{rank}(\mathbf{B}), \text{rank}(\mathbf{C}))$ , and  $\mathbf{N}$  has rank equal to  $M$ , we have that necessarily also  $\mathbf{A}$  and  $\mathbf{N}'$  have rank  $M$ . In particular,  $\mathbf{A}$  is an invertible matrix.

Next we prove that  $p_{\mathbf{f}, \mathbf{g}}(y | \mathbf{x}) = p_{\mathbf{f}', \mathbf{g}'}(y | \mathbf{x}), \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \implies \mathbf{g}_0(y) = \mathbf{A}^{-\top} \mathbf{g}'_0(y)$  for  $\mathbf{A} := \mathbf{L}^{-\top} \mathbf{L}'^{\top}$  invertible.

As before, we have that

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y) - \log(Z(\mathbf{x}, \mathcal{Y})) = \mathbf{f}'(\mathbf{x})^\top \mathbf{g}'(y) - \log(Z'(\mathbf{x}, \mathcal{Y})) \quad (23)$$

holds for all  $\mathbf{x} \in \mathcal{X}$  and all  $y \in \mathcal{Y}$ . In particular, this equation holds for any specific  $\mathbf{x}$ . From the diversity condition (Definition 2.1), we can choose  $M + 1$   $\mathbf{x}$ 's,  $\mathbf{x}_0, \dots, \mathbf{x}_M$ , such that the displaced embeddings  $\{\mathbf{f}_0(\mathbf{x}_i)\}_{i=1}^M$  are linearly independent. By subtracting the log-conditional distribution for the pivot  $\mathbf{x}_0 \in \mathcal{X}$ , we obtain for each  $\mathbf{x}_i \in \mathcal{X}$  the following:

$$\begin{aligned} & \mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y) - \mathbf{f}(\mathbf{x}_0)^\top \mathbf{g}(y) + \log(Z(\mathbf{x}_0, \mathcal{Y})) - \log(Z(\mathbf{x}_i, \mathcal{Y})) \\ &= \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'(y) - \mathbf{f}'(\mathbf{x}_0)^\top \mathbf{g}'(y) + \log(Z'(\mathbf{x}_0, \mathcal{Y})) - \log(Z'(\mathbf{x}_i, \mathcal{Y})) \end{aligned} \quad (24)$$

$$\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y) - \mathbf{f}(\mathbf{x}_0)^\top \mathbf{g}(y) = \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'(y) - \mathbf{f}'(\mathbf{x}_0)^\top \mathbf{g}'(y) + c_i \quad (25)$$

$$\mathbf{f}_0(\mathbf{x}_i)^\top \mathbf{g}(y) = \mathbf{f}'_0(\mathbf{x}_i)^\top \mathbf{g}'(y) + c_i, \quad (26)$$

where

$$c_i = \log\left(\frac{Z'(\mathbf{x}_0, \mathcal{Y})}{Z(\mathbf{x}_0, \mathcal{Y})}\right) + \log\left(\frac{Z(\mathbf{x}_i, \mathcal{Y})}{Z'(\mathbf{x}_i, \mathcal{Y})}\right). \quad (27)$$

Let  $\mathbf{N}$  be the matrix with  $\mathbf{f}_0(\mathbf{x}_i)$  as columns, let  $\mathbf{N}'$  be the matrix with  $\mathbf{f}'_0(\mathbf{x}_i)$  as columns and let  $\mathbf{c}$  be the vector with  $c_i$  as entries. Then, since  $\mathbf{N}$  is invertible

$$\mathbf{N}^\top \mathbf{g}(y) = \mathbf{N}'^\top \mathbf{g}'(y) + \mathbf{c} \quad (28)$$

$$\mathbf{g}(y) = \mathbf{N}^{-\top} \mathbf{N}'^\top \mathbf{g}'(y) + \mathbf{N}^{-\top} \mathbf{c}. \quad (29)$$

Since we found in Equation (22) that  $\mathbf{A}$  is invertible, we have that

$$\mathbf{N}' = \mathbf{A}^{-1} \mathbf{N}. \quad (30)$$

Therefore, we get

$$\mathbf{g}(y) = \mathbf{N}^{-\top} \mathbf{N}'^\top \mathbf{g}'(y) + \mathbf{N}^{-\top} \mathbf{c} \quad (31)$$

$$\mathbf{g}(y) = \mathbf{N}^{-\top} (\mathbf{A}^{-1} \mathbf{N})^\top \mathbf{g}'(y) + \mathbf{N}^{-\top} \mathbf{c} \quad (32)$$

$$\mathbf{g}(y) = \mathbf{N}^{-\top} \mathbf{N}^\top \mathbf{A}^{-\top} \mathbf{g}'(y) + \mathbf{N}^{-\top} \mathbf{c} \quad (33)$$

$$\mathbf{g}(y) = \mathbf{A}^{-\top} \mathbf{g}'(y) + \mathbf{b}, \quad (34)$$

where  $\mathbf{b} = \mathbf{N}^{-\top} \mathbf{c}$ . So, considering the displaced unembedding vectors, we get:

$$\mathbf{g}_0(y) = \mathbf{g}(y) - \mathbf{g}(y_0) \quad (35)$$

$$= \mathbf{A}^{-\top} \mathbf{g}'(y) + \mathbf{b} - \mathbf{A}^{-\top} \mathbf{g}'(y_0) - \mathbf{b} \quad (36)$$

$$= \mathbf{A}^{-\top} \mathbf{g}'_0(y) \quad (37)$$

This proves the claim.  $\square$

## C Models close in KL divergence can have dissimilar representations - details

### C.1 KL goes to Zero but Representations are Dissimilar

Here, we restate [Theorem 3.1](#) in full detail:

**Theorem** (Formal statement of [Theorem 3.1](#)). *Let  $(\mathbf{f}, \mathbf{g}_\rho) \in \Theta$  be a model with representations in  $\mathbb{R}^M$ . Assume  $M \geq 2$ . Let  $k := |\mathcal{Y}|$  be the total number of unembeddings and assume  $k \geq M + 1$ . Assume the model satisfies the following requirements:*

- i) *The unembeddings have fixed norm, that is  $\|\mathbf{g}_\rho(y_i)\| = \|\mathbf{g}_\rho(y_j)\| = \rho$ , for all  $y_i, y_j \in \mathcal{Y}$ , where  $\rho \in \mathbb{R}^+$ .*
- ii) *For every  $\mathbf{x}_i \in \mathcal{X}$ , let there be one element  $y_j \in \mathcal{Y}$  for which the cosine similarity  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > 0$  and  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell))$  for all  $y_\ell \neq y_j$ .*

*Let  $(\mathbf{f}', \mathbf{g}'_\rho)$  be a model which also satisfies i) and ii) and which is related to  $(\mathbf{f}, \mathbf{g}_\rho)$  in the following way:*

- iii) *For every  $\mathbf{x}_i \in \mathcal{X}$ ,  $\|\mathbf{f}'(\mathbf{x}_i)\| = \|\mathbf{f}(\mathbf{x}_i)\| = c(\mathbf{x}_i)$ .*
- iv) *For  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}}(p(y|\mathbf{x}_i))$  we have that the angle between  $\mathbf{f}'(\mathbf{x}_i)$  and  $\mathbf{g}'_\rho(y_j)$  is equal to the angle between  $\mathbf{f}(\mathbf{x}_i)$  and  $\mathbf{g}_\rho(y_j)$ , in particular,  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) = \cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_j))$ .*

*Then, making pairs of models satisfying assumptions i) to iv) and with increasing values of  $\rho$  gives us a sequence of pairs of models for which we have:*

$$d_{\text{KL}}(p_{\mathbf{f}, \mathbf{g}_\rho}, p_{\mathbf{f}', \mathbf{g}'_\rho}) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty. \quad (38)$$

*Also, it is possible to construct a model  $(\mathbf{f}', \mathbf{g}'_\rho) \in \Theta$  which satisfies the requirements above, but where as  $\rho \rightarrow \infty$ , the embeddings stay fixed and  $\mathbf{f}'(\mathbf{x})$  is not an invertible linear transformation of  $\mathbf{f}(\mathbf{x})$ .*

*Proof.* Let the models  $(\mathbf{f}, \mathbf{g}_\rho), (\mathbf{f}', \mathbf{g}'_\rho) \in \Theta$  be as described above. Below, we use the shorthand  $p := p_{\mathbf{f}, \mathbf{g}_\rho}$  and  $p' := p_{\mathbf{f}', \mathbf{g}'_\rho}$ .

We prove the results in two parts. (1) We show that, for any two models satisfying the requirements i) to iv) above, we get that  $d_{\text{KL}}(p, p') \rightarrow 0$  as  $\rho \rightarrow \infty$ . (2) We then specifically show how construct a model  $(\mathbf{f}, \mathbf{g}_\rho)$  and a model  $(\mathbf{f}', \mathbf{g}'_\rho)$ , which satisfies the requirements i) to iv) but where as  $\rho \rightarrow \infty$ , the embeddings stay fixed and  $\mathbf{f}'(\mathbf{x})$  is not an invertible linear transformation of  $\mathbf{f}(\mathbf{x})$ .

(1) Since  $D_{\text{KL}}(p(y|\mathbf{x}_i)||p'(y|\mathbf{x}_i)) \rightarrow 0$  for all  $\mathbf{x}_i \in \mathcal{X}$  implies that  $d_{\text{KL}}(p, p') \rightarrow 0$ , we will show that  $D_{\text{KL}}(p(y|\mathbf{x}_i)||p'(y|\mathbf{x}_i)) \rightarrow 0$  for all  $\mathbf{x}_i \in \mathcal{X}$ . Fix  $\mathbf{x}_i \in \mathcal{X}$ . For notational brevity, we denote with  $c_i \in \mathbb{R}^+$  the value of the norm of the embeddings on  $\mathbf{x}_i$ , i.e.,  $\|\mathbf{f}(\mathbf{x}_i)\| = \|\mathbf{f}'(\mathbf{x}_i)\| = c(\mathbf{x}_i)$  (assumption iii)). We then consider the KL divergence

$$D_{\text{KL}}(p(y|\mathbf{x}_i)||p'(y|\mathbf{x}_i)) = \sum_{j=1}^k p(y_j|\mathbf{x}_i) \log \frac{p(y_j|\mathbf{x}_i)}{p'(y_j|\mathbf{x}_i)}. \quad (39)$$

Since  $D_{\text{KL}}(p(y|\mathbf{x}_i)||p'(y|\mathbf{x}_i))$  involves a sum over  $j$ ,  $D_{\text{KL}}(p(y|\mathbf{x}_i)||p'(y|\mathbf{x}_i))$  will go to zero if each term in the sum goes to zero. To show that each term of this sum goes to zero, we first consider the term for  $y_j = \operatorname{arg max}_{y \in \mathcal{Y}}(p(y|\mathbf{x}_i))$ . In this case, we have that

$$p(y_j|\mathbf{x}_i) = \frac{\exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j))}{\sum_{\ell=1}^k \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_\ell))} \quad (40)$$

$$= \left( 1 + \sum_{\ell=1, \ell \neq j}^k \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_\ell) - \mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j)) \right)^{-1}. \quad (41)$$

From iii) we have  $\|\mathbf{f}(\mathbf{x}_i)\| = c_i$  and from i) we have that  $\|\mathbf{g}_\rho(y_j)\| = \rho$ . We can therefore write

$$\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) = \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) \cdot c_i \cdot \rho, \quad (42)$$

and substituting this into Equation (41), we see that

$$p(y_j|\mathbf{x}_i) = \left( 1 + \sum_{\ell=1, \ell \neq j}^k \exp(c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)))) \right)^{-1}. \quad (43)$$

Now since by assumption ii)  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > 0$  and for all  $y_\ell \neq y_j$

$$\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)), \quad (44)$$

we have that

$$\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) < 0, \quad (45)$$

Taking the limit in  $\rho$ , we get

$$\exp(c \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)))) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (46)$$

Since every term of the sum in Equation (43) goes to zero for  $\rho \rightarrow \infty$ , the whole sum goes to zero, which means that from Equation (43) we get that

$$p(y_j|\mathbf{x}_i) \rightarrow \frac{1}{1+0} = 1 \quad \text{for } \rho \rightarrow \infty. \quad (47)$$

Similarly, we also have that

$$p'(y_j|\mathbf{x}_i) \rightarrow 1 \quad \text{for } \rho \rightarrow \infty. \quad (48)$$

This means that for  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$ , from combining Equation (47), Equation (48) and Equation (39) we get that

$$p(y_j|\mathbf{x}_i) \log \frac{p(y_j|\mathbf{x}_i)}{p'(y_j|\mathbf{x}_i)} \rightarrow 1 \cdot \log \frac{1}{1} = 0 \quad \text{for } \rho \rightarrow \infty. \quad (49)$$

Next, we consider the case where  $y_j \neq \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$ . In this case, we consider

$$p(y_j|\mathbf{x}_i) \log \frac{p(y_j|\mathbf{x}_i)}{p'(y_j|\mathbf{x}_i)} = p(y_j|\mathbf{x}_i) (\log p(y_j|\mathbf{x}_i) - \log p'(y_j|\mathbf{x}_i)) \quad (50)$$

$$= p(y_j|\mathbf{x}_i) (\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) - \log Z(\mathbf{x}_i) - \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'_\rho(y_j) + \log Z'(\mathbf{x}_i)) \quad (51)$$

$$= p(y_j|\mathbf{x}_i) (\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) - \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'_\rho(y_j)) + p(y_j|\mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)}. \quad (52)$$

We will consider the two terms in Equation (52) separately. First, we consider

$$p(y_j|\mathbf{x}_i) (\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) - \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'_\rho(y_j)). \quad (53)$$

Similarly as for Equation (42), we have

$$\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) = \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) \cdot c_i \cdot \rho. \quad (54)$$

We use this to rewrite Equation (53) and see that

$$|\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) - \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'_\rho(y_j)| = |\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) \cdot c_i \cdot \rho - \cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_j)) \cdot c_i \cdot \rho| \quad (55)$$

$$= |c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) - \cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_j)))| \quad (56)$$

$$\leq 2c_i \cdot \rho. \quad (57)$$

This gives us that

$$p(y_j|\mathbf{x}_i) |\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) - \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'_\rho(y_j)| \leq 2c_i \rho \cdot p(y_j|\mathbf{x}_i) \quad (58)$$

Since  $|x| \rightarrow 0$  implies that  $x \rightarrow 0$  and we have shown that the absolute values of the term is upper bounded by  $2c_i \rho \cdot p(y_j | \mathbf{x}_i)$ , we will show that the term goes to zero by showing that  $2c_i \rho \cdot p(y_j | \mathbf{x}_i)$  goes to zero. We see that

$$2c_i \rho \cdot p(y_j | \mathbf{x}_i) = \frac{2c_i \cdot \rho \cdot \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j))}{\sum_{\ell=1}^k \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_\ell))} \quad (59)$$

$$= \frac{2c_i \cdot \rho}{1 + \sum_{\ell=1, \ell \neq j}^k \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_\ell) - \mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j))} \quad (60)$$

$$= \frac{2c_i \cdot \rho}{1 + \sum_{\ell=1, \ell \neq j}^k \exp(c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))))} \quad (61)$$

$$= 2c_i \left/ \left( 1/\rho + \sum_{\ell=1, \ell \neq j}^k \frac{\exp(c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))))}{\rho} \right) \right. \quad (62)$$

Consider now  $y_r = \operatorname{argmax}_{y \in \mathcal{Y}} p(y | \mathbf{x}_i)$ , recall we have  $y_j \neq y_r$ . Considering this  $y_r$ , we have by assumption ii) that  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_r)) > 0$  and

$$\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_r)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > 0, \quad (63)$$

Since for  $\alpha \in \mathbb{R}^+$ , we have  $\exp(\alpha x)/x \rightarrow \infty$  for  $x \rightarrow \infty$ , for this  $y_r$ , we have that

$$\frac{\exp(c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_r)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))))}{\rho} \rightarrow \infty \quad \text{for } \rho \rightarrow \infty. \quad (64)$$

We can now define the denominator of Equation (62) as

$$\begin{aligned} d(\rho) := & \frac{1}{\rho} + \frac{\exp(c_i \rho (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_r)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))))}{\rho} \\ & + \sum_{\ell=1, \ell \neq j, r}^k \frac{\exp(c_i \rho (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))))}{\rho} \end{aligned} \quad (65)$$

and see that the first term of Equation (65) goes to zero, the second term goes to infinity Equation (64) and the third term is always positive since  $\exp(x)$  is always positive. Therefore we have that

$$d(\rho) \rightarrow \infty \quad \text{for } \rho \rightarrow \infty \quad (66)$$

which means that

$$2c_i \rho \cdot p(y_j | \mathbf{x}_i) = \frac{2c_i}{d(\rho)} \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (67)$$

Thus, we get for the first term of Equation (52) that

$$p(y_j | \mathbf{x}_i) (\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}_\rho(y_j) - \mathbf{f}'(\mathbf{x}_i)^\top \mathbf{g}'_\rho(y_j)) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (68)$$

which was the desired result.

We now consider the second term in Equation (52)

$$p(y_j | \mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)}. \quad (69)$$

We see that

$$\frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} = \frac{\sum_{\ell=1}^k \exp(\cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_\ell)) \cdot c_i \cdot \rho)}{\sum_{m=1}^k \exp(\cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_m)) \cdot c_i \cdot \rho)} \quad (70)$$

$$\leq \frac{\sum_{\ell=1}^k \exp(1 \cdot c_i \cdot \rho)}{\sum_{m=1}^k \exp(-1 \cdot c_i \cdot \rho)} \quad (71)$$

$$= \frac{\exp(c_i \cdot \rho)}{\exp(-c_i \cdot \rho)} \quad (72)$$

$$= \exp(2c_i \cdot \rho). \quad (73)$$

Which means we have that

$$p(y_j|\mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} \leq p(y_j|\mathbf{x}_i) \log \exp(2c_i \cdot \rho) = 2c_i \rho \cdot p(y_j|\mathbf{x}_i). \quad (74)$$

We also have that

$$\frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} = \frac{\sum_{\ell=1}^k \exp(\cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_\ell)) \cdot c_i \cdot \rho)}{\sum_{m=1}^k \exp(\cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_m)) \cdot c_i \cdot \rho)} \quad (75)$$

$$\geq \frac{\sum_{\ell=1}^k \exp(-1 \cdot c_i \cdot \rho)}{\sum_{m=1}^k \exp(1 \cdot c_i \cdot \rho)} \quad (76)$$

$$= \frac{\exp(-c_i \cdot \rho)}{\exp(c_i \cdot \rho)} \quad (77)$$

$$= \exp(-2c_i \cdot \rho). \quad (78)$$

Which means we have that

$$p(y_j|\mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} \geq p(y_j|\mathbf{x}_i) \log \exp(-2c_i \cdot \rho) \quad (79)$$

$$= -2c_i \rho \cdot p(y_j|\mathbf{x}_i). \quad (80)$$

which gives us

$$-p(y_j|\mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} \leq 2c_i \rho \cdot p(y_j|\mathbf{x}_i). \quad (81)$$

Now Equation (74) and Equation (81) together give us that

$$\left| p(y_j|\mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} \right| \leq 2c_i \rho \cdot p(y_j|\mathbf{x}_i) \quad (82)$$

Thus, we have upper bounded the absolute value of the term with  $2c_i \rho \cdot p(y_j|\mathbf{x}_i)$  like in Equation (58). Using again that  $2c_i \rho \cdot p(y_j|\mathbf{x}_i) \rightarrow 0$  for  $\rho \rightarrow \infty$  (Equation (67)), we get that

$$p(y_j|\mathbf{x}_i) \log \frac{Z'(\mathbf{x}_i)}{Z(\mathbf{x}_i)} \rightarrow 0 \quad \text{for } \rho \rightarrow \infty. \quad (83)$$

Since both terms in the sum (Equation (52)) go to zero for  $\rho \rightarrow \infty$ , the entire sum goes to zero and for  $y_j \neq \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$ , we have that

$$p(y_j|\mathbf{x}_i) \log \frac{p(y_j|\mathbf{x}_i)}{p'(y_j|\mathbf{x}_i)} \rightarrow 0 \quad \text{for } \rho \rightarrow \infty. \quad (84)$$

Since we now have this result for both  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$  and  $y_j \neq \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$ , the sum over all  $j$  in  $D_{\text{KL}}(p(y|\mathbf{x}_i) \| p'(y|\mathbf{x}_i))$  (Equation (39)) also goes to zero. Thus,

$$D_{\text{KL}}(p(y|\mathbf{x}_i) \| p'(y|\mathbf{x}_i)) \rightarrow 0 \quad \forall \mathbf{x}_i \in \mathcal{X}, \quad (85)$$

which means that

$$d_{\text{KL}}(p, p') = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [D_{\text{KL}}(p(y|\mathbf{x}) \| p'(y|\mathbf{x}))] \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (86)$$

proving the first claim of the theorem.

(2) We now present two constructions which satisfy the requirements i) to iv) above, but where  $\mathbf{f}'(\mathbf{x})$  is not an invertible linear transformation of  $\mathbf{f}(\mathbf{x})$ . In the following, we use  $\mathbf{e}_i$  to denote the  $i$ -th unit vector in  $\mathbb{R}^M$ , whose  $i$ -th component is equal to 1, and all other components are equal to zero.

**Construction for  $k \geq M + 2$ .** For  $i = 1, \dots, M$ , let  $\mathbf{g}_\rho(y_i) = \rho \cdot \mathbf{e}_i$ . Let  $\mathbf{g}_\rho(y_{M+1}) = -\rho \cdot \mathbf{e}_1$  and  $\mathbf{g}_\rho(y_{M+2}) = -\rho \cdot \mathbf{e}_2$ . For any remaining labels with index  $j > M + 2$ , let them be such that  $\mathbf{g}_\rho(y_j) \neq \mathbf{g}_\rho(y_\ell)$  for  $j \neq \ell$  and more than  $\pi/2$  radian angle away from  $\mathbf{e}_2, -\mathbf{e}_1$  and  $-\mathbf{e}_2$ . Let  $\{\mathbf{x}_n\}_{n=1}^\infty \subset \mathcal{X}$  and  $\{\mathbf{x}_m\}_{m=1}^\infty \subset \mathcal{X}$  be two sets of inputs. We define

$$\mathbf{f}(\mathbf{x}_n) = \begin{pmatrix} \cos(\pi - \frac{\pi}{4}(1 - \frac{1}{n})) \\ \sin(\pi - \frac{\pi}{4}(1 - \frac{1}{n})) \\ 0 \\ \vdots \end{pmatrix} \quad (87)$$

Now  $\{\mathbf{f}(\mathbf{x}_n)\}_{n=1}^{\infty}$  is a sequence with a well-defined finite limit, which we for ease of reference shall call  $\mathbf{a}$

$$\mathbf{a} = \lim_{n \rightarrow \infty} \mathbf{f}(\mathbf{x}_n) = \begin{pmatrix} \cos(\frac{3\pi}{4}) \\ \sin(\frac{3\pi}{4}) \\ 0 \\ \vdots \end{pmatrix} \quad (88)$$

We now define  $\mathbf{f}$  for the other set.

$$\mathbf{f}(\mathbf{x}_m) = \begin{pmatrix} \cos(\frac{3\pi}{4} - \frac{\pi}{4m}) \\ \sin(\frac{3\pi}{4} - \frac{\pi}{4m}) \\ 0 \\ \vdots \end{pmatrix} \quad (89)$$

$\{\mathbf{f}(\mathbf{x}_m)\}_{m=1}^{\infty}$  is now a sequence with the same limit, that is,

$$\lim_{m \rightarrow \infty} \mathbf{f}(\mathbf{x}_m) = \begin{pmatrix} \cos(\frac{3\pi}{4}) \\ \sin(\frac{3\pi}{4}) \\ 0 \\ \vdots \end{pmatrix} = \mathbf{a} \quad (90)$$

For every remaining  $\mathbf{x} \in \mathcal{X}$ , construct  $\mathbf{f}(\mathbf{x})$  such that there exists a label  $y_j \in \mathcal{Y}$  such that we have the cosine similarities  $\cos(\mathbf{f}(\mathbf{x}), \mathbf{g}_\rho(y_j)) > 0$  and  $\cos(\mathbf{f}(\mathbf{x}), \mathbf{g}_\rho(y_j)) > \cos(\mathbf{f}(\mathbf{x}), \mathbf{g}_\rho(y_\ell))$  for  $y_\ell \neq y_j$ . With this  $(\mathbf{f}, \mathbf{g})$  satisfies i) and ii).

We now construct the second model  $(\mathbf{f}', \mathbf{g}')$ . Let  $\mathbf{g}'_\rho(y_i) = \mathbf{g}_\rho(y_i)$  for  $i = 1, \dots, M$  and  $i > M + 2$ . Let  $\mathbf{g}'_\rho(y_{M+1}) = -\rho \cdot \mathbf{e}_2$  and  $\mathbf{g}'_\rho(y_{M+2}) = -\rho \cdot \mathbf{e}_1$ . Note that we swapped the unembeddings for  $\mathbf{g}'$  of  $y_{M+1}$  and  $y_{M+2}$  compared to  $\mathbf{g}$ . For all  $\mathbf{x} \in \mathcal{X}$  and the corresponding  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x})$  such that  $y_j \neq y_{M+1}, y_{M+2}$ , let  $\mathbf{f}'(\mathbf{x}) = \mathbf{f}(\mathbf{x})$ . However, for all  $\mathbf{x} \in \mathcal{X}$  and the corresponding  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x})$  such that  $y_j = M + 1$ , let  $\|\mathbf{f}'(\mathbf{x})\| = \|\mathbf{f}(\mathbf{x})\|$  and let  $\mathbf{f}'(\mathbf{x})$  be rotated counterclockwise with respect to the first two axes (i.e., the first two components of the representations) such that  $\cos(\mathbf{f}'(\mathbf{x}), \mathbf{g}'_\rho(y_j)) = \cos(\mathbf{f}(\mathbf{x}), \mathbf{g}_\rho(y_j))$ . Also, for all  $\mathbf{x}_i \in \mathcal{X}$  and the corresponding  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$  such that  $y_j = M + 2$ , let  $\|\mathbf{f}'(\mathbf{x}_i)\| = \|\mathbf{f}(\mathbf{x}_i)\|$  and let  $\mathbf{f}'(\mathbf{x}_i)$  be rotated clockwise with respect to the first two axes such that  $\cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_j)) = \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))$ . This construction satisfies the requirements i) to iv).

We will show that  $\mathbf{f}'(\mathbf{x})$  is not an invertible linear transformation of  $\mathbf{f}(\mathbf{x})$  by showing that any function  $\tau : \mathbb{R}^M \rightarrow \mathbb{R}^M$  with  $\tau(\mathbf{f}(\mathbf{x})) = \mathbf{f}'(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$  cannot be continuous.

Let  $\tau : \mathbb{R}^M \rightarrow \mathbb{R}^M$  be such that  $\tau(\mathbf{f}(\mathbf{x})) = \mathbf{f}'(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$ . We recall that a function  $\mathbf{h} : \mathbb{R}^M \rightarrow \mathbb{R}^M$  is by definition continuous at  $\mathbf{c} \in \mathbb{R}^M$  if

$$\mathbf{h}(\mathbf{x}) \rightarrow \mathbf{h}(\mathbf{c}) \quad \text{for } \mathbf{x} \rightarrow \mathbf{c} \quad (91)$$

Therefore, we consider the limits of the sequences  $\{\tau(\mathbf{f}(\mathbf{x}_n))\}_{n=1}^{\infty}$  and  $\{\tau(\mathbf{f}(\mathbf{x}_m))\}_{m=1}^{\infty}$ . We first note that by our construction, we have for the rotation matrix

$$\mathbf{R} = \begin{bmatrix} \cos(\frac{\pi}{2}) & -\sin(\frac{\pi}{2}) & 0 & \cdots \\ \sin(\frac{\pi}{2}) & \cos(\frac{\pi}{2}) & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (92)$$

that

$$\mathbf{f}'(\mathbf{x}_n) = \mathbf{R}\mathbf{f}(\mathbf{x}_n) \quad (93)$$

Which means that since  $\tau(\mathbf{f}(\mathbf{x}_n)) = \mathbf{f}'(\mathbf{x}_n)$ , we have that

$$\tau(\mathbf{f}(\mathbf{x}_n)) = \mathbf{R}\mathbf{f}(\mathbf{x}_n) = \begin{pmatrix} \cos(\frac{3\pi}{2} - \frac{\pi}{4}(1 - \frac{1}{n})) \\ \sin(\frac{3\pi}{2} - \frac{\pi}{4}(1 - \frac{1}{n})) \\ 0 \\ \vdots \end{pmatrix} \quad (94)$$

and we see that when  $n \rightarrow \infty$ ,

$$\tau(\mathbf{f}(\mathbf{x}_n)) \rightarrow \mathbf{b} := \begin{pmatrix} \cos(\frac{3\pi}{2}) \\ \sin(\frac{3\pi}{2}) \\ 0 \\ \vdots \end{pmatrix} \quad \text{for } \mathbf{f}(\mathbf{x}_n) \rightarrow \mathbf{a} \quad (95)$$

However, we also have that  $\tau(\mathbf{f}(\mathbf{x}_m)) = \mathbf{f}'(\mathbf{x}_m) = \mathbf{f}(\mathbf{x}_m)$ . This means that for  $m \rightarrow \infty$ , we have

$$\tau(\mathbf{f}(\mathbf{x}_m)) \rightarrow \mathbf{a} \quad \text{for } \mathbf{f}(\mathbf{x}_m) \rightarrow \mathbf{a} \quad (96)$$

Now since for  $\mathbf{b}$  from Equation (95), we have  $\mathbf{b} \neq \mathbf{a}$ ,  $\tau$  cannot be continuous. In particular, it cannot be an invertible linear transformation, and thus  $\mathbf{f}'(\mathbf{x})$  is not an invertible linear transformation of  $\mathbf{f}(\mathbf{x})$ .

Note that if either  $\mathbf{f}$  or  $\mathbf{f}'$  is not injective, then both  $\mathbf{f}$  and  $\mathbf{f}'$  can be smooth. However, if both  $\mathbf{f}$  and  $\mathbf{f}'$  are injective, then either  $\mathbf{f}$  or  $\mathbf{f}'$  has to be non-continuous.

Notice also that we can permute additional labels to bring the embeddings further from a linear transformation. See for example Fig. 2 which has been constructed by permuting multiple labels.

**Construction for  $k = M + 1$ .** For  $i = 1, \dots, M$ , let  $\mathbf{g}_\rho(y_i) = \rho \cdot \mathbf{e}_i$  and  $\mathbf{g}_\rho(y_{M+1}) = -\rho \cdot \mathbf{e}_1$ . Let  $\{\mathbf{x}_n\}_{n=1}^\infty \subset \mathcal{X}$  and  $\{\mathbf{x}_m\}_{m=1}^\infty \subset \mathcal{X}$  be two sets of inputs. We define

$$\mathbf{f}(\mathbf{x}_n) = \begin{pmatrix} \cos(\pi - \frac{\pi}{4}(1 - \frac{1}{n})) \\ \sin(\pi - \frac{\pi}{4}(1 - \frac{1}{n})) \\ 0 \\ \vdots \end{pmatrix} \quad (97)$$

Now  $\{\mathbf{f}(\mathbf{x}_n)\}_{n=1}^\infty$  is a sequence with a well-defined finite limit, which we for ease of reference shall call  $\mathbf{a}$

$$\mathbf{a} = \lim_{n \rightarrow \infty} \mathbf{f}(\mathbf{x}_n) = \begin{pmatrix} \cos(\frac{3\pi}{4}) \\ \sin(\frac{3\pi}{4}) \\ 0 \\ \vdots \end{pmatrix} \quad (98)$$

We now define  $\mathbf{f}$  for the other set.

$$\mathbf{f}(\mathbf{x}_m) = \begin{pmatrix} \cos(\frac{3\pi}{4} - \frac{\pi}{4m}) \\ \sin(\frac{3\pi}{4} - \frac{\pi}{4m}) \\ 0 \\ \vdots \end{pmatrix} \quad (99)$$

$\{\mathbf{f}(\mathbf{x}_m)\}_{m=1}^\infty$  is now a sequence with the same limit, that is,

$$\lim_{m \rightarrow \infty} \mathbf{f}(\mathbf{x}_m) = \begin{pmatrix} \cos(\frac{3\pi}{4}) \\ \sin(\frac{3\pi}{4}) \\ 0 \\ \vdots \end{pmatrix} = \mathbf{a} \quad (100)$$

For every remaining  $\mathbf{x}_i \in \mathcal{X}$ , construct  $\mathbf{f}(\mathbf{x}_i)$  such that there exists a label  $y_j \in \mathcal{Y}$  such that we have the cosine similarities  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > 0$  and  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j)) > \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_\ell))$  for  $y_\ell \neq y_j$ . With this  $(\mathbf{f}, \mathbf{g})$  satisfies i) and ii).

We now construct the second model  $(\mathbf{f}', \mathbf{g}')$ . Let  $\mathbf{g}'_\rho(y_i) = \mathbf{g}_\rho(y_i)$  for  $i = 1, \dots, M$  and let  $\mathbf{g}'_\rho(y_{M+1}) = -\rho \cdot \mathbf{e}_2$ . For all  $\mathbf{x}_i \in \mathcal{X}$  and the corresponding  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$  such that  $y_j \neq y_{M+1}$ , let  $\mathbf{f}'(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)$ . For all  $\mathbf{x}_i \in \mathcal{X}$  and the corresponding  $y_j = \operatorname{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{x}_i)$  such that  $y_j = y_{M+1}$ , construct  $\mathbf{f}'(\mathbf{x}_i)$  from  $\mathbf{f}(\mathbf{x}_i)$  by rotating it counterclockwise with respect to the first two axes such that  $\cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'_\rho(y_j)) = \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}_\rho(y_j))$ . This construction satisfies the requirements i) to iv) and by the same argument as used for the case with  $k \geq M + 2$ ,  $\mathbf{f}'(\mathbf{x})$  is not an invertible linear transformation of  $\mathbf{f}(\mathbf{x})$ . □

## C.2 Loss goes to Zero but Representations are Dissimilar

We denote the negative log-likelihood for a model  $p$  with respect to a data distribution  $p_{\mathcal{D}}$  as:

$$\text{NLL}_{p_{\mathcal{D}}}(p) := \mathbb{E}_{(\mathbf{x}, y) \sim p_{\mathcal{D}}}[-\log p(y | \mathbf{x})]. \quad (101)$$

In the following, it is useful to establish the link between the negative log-likelihood and the  $D_{\text{KL}}$  divergence with the empirical distribution. To this end, we introduce the empirical distribution over inputs and outputs  $p_{\mathcal{D}}(\mathbf{x}, y)$ , whose marginal on the input is given by  $p_{\mathcal{D}}(\mathbf{x})$ . This can be rewritten as:

$$p_{\mathcal{D}}(\mathbf{x}, y) = p_{\mathcal{D}}(y | \mathbf{x})p_{\mathcal{D}}(\mathbf{x}). \quad (102)$$

Notice that the conditional distribution underlies how labels are associated with the input. We can connect the negative log-likelihood to the KL divergence with the empirical distribution, a well-known fact in the literature, as shown in the next Lemma:

**Lemma C.1.** *Let  $p_{\mathcal{D}}(\mathbf{x}, y) = p_{\mathcal{D}}(y|\mathbf{x})p_{\mathcal{D}}(\mathbf{x})$  be the ground-truth data distribution. Let  $p(y|\mathbf{x})$  be a model such that, for all  $\mathbf{x}$  in the support of  $p_{\mathcal{D}}(\mathbf{x})$ , we have  $p_{\mathcal{D}}(y | \mathbf{x}) \ll p(y | \mathbf{x})$ .<sup>15</sup> Define the negative log-likelihood as  $\text{NLL}_{p_{\mathcal{D}}}(p) := \mathbb{E}_{(\mathbf{x}, y) \sim p_{\mathcal{D}}}[-\log p(y | \mathbf{x})]$ . Then:*

$$\text{NLL}_{p_{\mathcal{D}}}(p) = d_{\text{KL}}(p_{\mathcal{D}}, p) + \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}}[H(p_{\mathcal{D}}(y | \mathbf{x}))], \quad (103)$$

where  $H(q) := -\mathbb{E}_{x \sim q}[\log q(x)]$  denotes the entropy of a distribution  $q$ .

*Proof.* From the definition of KL divergence and entropy:

$$D_{\text{KL}}(p_{\mathcal{D}}(y | \mathbf{x}) \| p(y | \mathbf{x})) = \mathbb{E}_{y \sim p_{\mathcal{D}}(y|\mathbf{x})} [\log p_{\mathcal{D}}(y | \mathbf{x}) - \log p(y | \mathbf{x})] \quad (104)$$

$$= -H(p_{\mathcal{D}}(y | \mathbf{x})) + \mathbb{E}_{y \sim p_{\mathcal{D}}(y|\mathbf{x})} [-\log p(y | \mathbf{x})]. \quad (105)$$

Moving the entropy to the other side and taking expectations w.r.t.  $p_{\mathcal{D}}(\mathbf{x})$ , we get:

$$\mathbb{E}_{(\mathbf{x}, y) \sim p_{\mathcal{D}}} [-\log p(y | \mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [D_{\text{KL}}(p_{\mathcal{D}}(y | \mathbf{x}) \| p(y | \mathbf{x}))] + \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [H(p_{\mathcal{D}}(y | \mathbf{x}))] \quad (106)$$

The statement follows by the definition of  $d_{\text{KL}}$  (Equation (6)) and NLL.  $\square$

Notice that, when the conditional distribution  $p_{\mathcal{D}}(y | \mathbf{x})$  is degenerate, i.e., to each  $\mathbf{x} \in \mathcal{X}$  a single  $y \in \mathcal{Y}$  is associated, the expectation of the conditional Shannon entropy vanishes.

Next, we restate [Corollary 3.2](#) and prove in full details:

**Corollary** (Formal version of [Corollary 3.2](#)). *Assume we have a data distribution where only one label is associated with each unique input. Let  $(\mathbf{f}, \mathbf{g}) \in \Theta$  be a model with representations in  $\mathbb{R}^M$ . Assume  $M \geq 2$ . Let  $k := |\mathcal{Y}|$  be the total number of unembeddings and assume  $k \geq M + 1$ . Assume the model satisfies the following requirements:*

- i) *The unembeddings have fixed norm, that is  $\|\mathbf{g}(y_i)\| = \|\mathbf{g}(y_j)\| = \rho$ , for all  $y_i, y_j \in \mathcal{Y}$ , where  $\rho \in \mathbb{R}^+$ .*
- ii) *For every  $\mathbf{x}_i \in \mathcal{X}$ , let there be one element  $y_j \in \mathcal{Y}$  for which the cosine similarity  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)) > 0$  and  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)) > \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_\ell))$  for all  $y_\ell \neq y_j$ . Also, assume  $y_j$  is the only label associated with  $\mathbf{x}_i$ .*

Let  $(\mathbf{f}', \mathbf{g}')$  be a model which also satisfies i) and ii) and which is related to  $(\mathbf{f}, \mathbf{g})$  in the following way:

- iii) *For every  $\mathbf{x}_i \in \mathcal{X}$ ,  $\|\mathbf{f}'(\mathbf{x}_i)\| = \|\mathbf{f}(\mathbf{x}_i)\|$ .*

- iv) *For  $\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}}(p(y|\mathbf{x}_i))$  we have that the angle between  $\mathbf{f}'(\mathbf{x}_i)$  and  $\mathbf{g}'(\hat{y})$  is equal to the angle between  $\mathbf{f}(\mathbf{x}_i)$  and  $\mathbf{g}(\hat{y})$ , in particular,  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(\hat{y})) = \cos(\mathbf{f}'(\mathbf{x}_i), \mathbf{g}'(\hat{y}))$ .*

Then, making a sequence of pairs of models indexed by  $\rho$ , we have

$$\text{NLL}(p_{\mathbf{f}, \mathbf{g}}) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (107)$$

and

$$\text{NLL}(p_{\mathbf{f}', \mathbf{g}'}) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty. \quad (108)$$

<sup>15</sup>Here  $\ll$  denotes absolute continuity.

*Proof.* Below, we use the shorthand  $p := p_{\mathbf{f}, \mathbf{g}}$  and  $p' := p_{\mathbf{f}', \mathbf{g}'}$ . We make use of the same calculations as for the proof of [Theorem 3.1](#). We consider  $p(y_j|\mathbf{x}_i)$  for  $y_j$  the label assigned to  $\mathbf{x}_i$  according to the data. We denote with  $c_i \in \mathbb{R}^+$  the value of the norm of the embeddings on  $\mathbf{x}_i$ , i.e.,  $\|\mathbf{f}(\mathbf{x}_i)\| = \|\mathbf{f}'(\mathbf{x}_i)\| = c_i$ . In this case we have that

$$p(y_j|\mathbf{x}_i) = \frac{\exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y_j))}{\sum_{\ell=1}^k \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y_\ell))} \quad (109)$$

$$= \frac{1}{1 + \sum_{\ell=1, \ell \neq j}^k \exp(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y_\ell) - \mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y_j))} \quad (110)$$

By construction,  $\|\mathbf{f}(\mathbf{x}_i)\| = c_i$ . We use that

$$\mathbf{f}(\mathbf{x}_i)^\top \mathbf{g}(y_j) = \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)) \cdot c_i \cdot \rho \quad (111)$$

and see that

$$p(y_j|\mathbf{x}_i) = \left( 1 + \sum_{\ell=1, \ell \neq j}^k \exp(c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)))) \right)^{-1} \quad (112)$$

Now since by construction,  $\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)) > 0$ , and for all  $\ell \neq j$

$$\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)) > \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_\ell)) \quad (113)$$

we have for all  $\ell$  that

$$\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)) < 0 \quad (114)$$

and

$$\exp(c_i \cdot \rho \cdot (\cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_\ell)) - \cos(\mathbf{f}(\mathbf{x}_i), \mathbf{g}(y_j)))) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (115)$$

which means that combined with [Equation \(112\)](#) we get

$$p(y_j|\mathbf{x}_i) \rightarrow \frac{1}{1+0} = 1 \quad \text{for } \rho \rightarrow \infty \quad (116)$$

Similarly, we have that

$$p'(y_j|\mathbf{x}_i) \rightarrow 1 \quad \text{for } \rho \rightarrow \infty. \quad (117)$$

Therefore, we have for the NLL:

$$\text{NLL}(p) = - \sum_{i=1}^n \log p(y_j|\mathbf{x}_i) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty \quad (118)$$

and

$$\text{NLL}(p') = - \sum_{i=1}^n \log p'(y_j|\mathbf{x}_i) \rightarrow 0 \quad \text{for } \rho \rightarrow \infty. \quad (119)$$

□

## D Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA)

Partial least squares (PLS) algorithms are techniques that, for two random variables  $\mathbf{z}$  and  $\mathbf{w}$ , derive the cross-covariance matrix  $\Sigma_{\mathbf{z}\mathbf{w}}$  using latent variables. We give here a short description of the two variants that are relevant to our work. A more detailed description can be found in [66].

In our case, we will work with random variables of the same dimension. So in the following, we consider  $\mathbf{z}, \mathbf{w} \in \mathbb{R}^M$ . We will mostly work with the centered and normalized versions of the random variables. So if  $z_i$  is the  $i$ 'th dimension of the original random variable, we will consider

$$z'_i = \frac{z_i - \mathbb{E}[z_i]}{\sqrt{\text{Var}[z_i]}}. \quad (120)$$

### D.1 PLS-SVD

PLS-SVD [54, 60] is a variant of PLS where the latent vectors are simply the left and right singular vectors in the singular value decomposition of the cross-covariance matrix,  $\Sigma_{\mathbf{z}\mathbf{w}}$ .

In [Algorithm 1](#), we report the algorithm in the case where we have  $n$  samples of our  $M$ -dimensional random variables. In this case, we work with  $(n \times M)$  matrices  $\mathbf{Z}, \mathbf{W}$ , where each row is a sample. The aim of the algorithm is to get projection vectors  $\mathbf{u}^{(r)}, \mathbf{v}^{(r)}$  such that  $\text{Cov}_{\mathbf{Z}\mathbf{W}}[\mathbf{Z}\mathbf{u}^{(r)}, \mathbf{W}\mathbf{v}^{(r)}]$  is as large as possible under the constraint that  $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ . The original algorithm includes a choice of signs for the covariances. In this version of the algorithm, we have chosen all signs to be positive.  $R > 0$  denotes the maximal rank of the algorithm.

---

**Algorithm 1** Iterative SVD Projection Extraction from Cross-Covariance Matrix

---

- 1: Set  $r \leftarrow 1$
  - 2: Center and scale  $\mathbf{Z}$  and  $\mathbf{W}$
  - 3: Compute cross-covariance matrix:  $\mathbf{C} \leftarrow \mathbf{Z}^\top \mathbf{W}$
  - 4: Set  $\mathbf{C}^{(1)} \leftarrow \mathbf{C}$
  - 5: **while**  $\mathbf{C}^{(r)} \neq 0$  and  $r \leq R$  **do**
  - 6:     Perform SVD:  $\mathbf{C}^{(r)} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$
  - 7:     Extract leading singular vectors:  $\mathbf{u}_1^{(r)} \leftarrow$  first column of  $\mathbf{U}$ ,  $\mathbf{v}_1^{(r)} \leftarrow$  first column of  $\mathbf{V}$
  - 8:     Save  $\mathbf{u}_1^{(r)}$  and  $\mathbf{v}_1^{(r)}$  as the  $r$ -th projection vectors
  - 9:     Let  $\sigma_r \leftarrow$  leading singular value of  $\mathbf{C}^{(r)}$
  - 10:    Update matrix:  $\mathbf{C}^{(r+1)} \leftarrow \mathbf{C}^{(r)} - \sigma_r \mathbf{u}_1^{(r)} \mathbf{v}_1^{(r)\top}$
  - 11:     $r \leftarrow r + 1$
- 

Considering this algorithm, we see that the projection vectors we get,  $\mathbf{u}^{(r)}, \mathbf{v}^{(r)}$ , are exactly the first  $m$  singular vectors of  $\mathbf{Z}^\top \mathbf{W}$  corresponding to the largest  $m$  singular values, where  $m$  is the value of  $r$  at the exit point. Note that this relates to principal component analysis (PCA) in the sense that when doing PCA, one uses the singular value decomposition of the covariance matrix to obtain components that capture the variance in a dataset (a random variable). When doing PLS-SVD, one uses singular value decomposition of the cross-covariance matrix to obtain components that capture the covariance between two random variables.

### D.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) [22] seeks pairs of unit-length vectors  $\mathbf{u}_\ell \in \mathbb{R}^{d_z}$  and  $\mathbf{v}_\ell \in \mathbb{R}^{d_w}$  ( $\ell = 1, \dots, r$ ) that maximize the correlation between the one-dimensional projections  $\mathbf{u}_\ell^\top \mathbf{z}$  and  $\mathbf{v}_\ell^\top \mathbf{w}$ , subject to mutual orthogonality of successive directions. In other words, CCA finds the singular value decomposition of the matrix  $\mathbf{Q} = \Sigma_{\mathbf{z}\mathbf{z}}^{-\frac{1}{2}} \Sigma_{\mathbf{z}\mathbf{w}} \Sigma_{\mathbf{w}\mathbf{w}}^{-\frac{1}{2}}$ , where  $\Sigma_{\mathbf{z}\mathbf{w}}$  is a cross-covariance matrix and  $\Sigma_{\mathbf{z}\mathbf{z}}, \Sigma_{\mathbf{w}\mathbf{w}}$  are covariance matrices. Equivalently,  $\mathbf{Q}$  is the cross-covariance of the whitened variables  $\Sigma_{\mathbf{z}\mathbf{z}}^{-\frac{1}{2}} \mathbf{z}$  and  $\Sigma_{\mathbf{w}\mathbf{w}}^{-\frac{1}{2}} \mathbf{w}$ , which have identity covariance. Whitening removes the influence of marginal variances, so the eigenvalues found by SVD are the canonical correlations. The results of CCA and PLS-SVD will differ if there are high covariances between the  $z_i$ 's or the  $w_i$ 's.

We can use CCA to define a similarity measure between random variables  $\mathbf{z}$  and  $\mathbf{w}$ , known as the mean canonical correlation  $m_{\text{CCA}}(\mathbf{z}, \mathbf{w})$ , which is the mean of the maxima correlations for the

centered and rescaled random variables. Because of these centering and rescaling operations,  $m_{\text{CCA}}$  will be invariant to any invertible linear transformation.

## E When Closeness in Distribution Implies Representational Similarity - Details

We here include all the technical details of [Section 4](#).

### E.1 The Connection Between Representations and Log-Likelihoods

Let  $(\mathbf{f}, \mathbf{g}) \in \Theta$  be a model satisfying the diversity condition ([Definition 2.1](#)) and consider  $\mathbf{x}_1, \dots, \mathbf{x}_M \in \mathcal{X}$  and  $y_1, \dots, y_M \in \mathcal{Y}$  such that both the vectors  $\{\mathbf{f}_0(\mathbf{x})\}_{i=1}^M$  and  $\{\mathbf{g}_0(y_i)\}_{i=1}^M$  are linearly independent. We recall some notation for use in the Lemma. We denote with  $\mathbf{L}$  and  $\mathbf{N}$  the following matrices:

$$\mathbf{L} := (\mathbf{g}_0(y_1), \dots, \mathbf{g}_0(y_M)), \quad \mathbf{N} := (\mathbf{f}_0(\mathbf{x}_1), \dots, \mathbf{f}_0(\mathbf{x}_M)). \quad (121)$$

We recall the definition of the following functions:

$$\psi_{\mathbf{x}}(y_i; p) := \sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]} \quad \text{and} \quad \psi_y(\mathbf{x}_j; p) := \sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_0)]}. \quad (122)$$

We also denote with  $\mathbf{S}, \mathbf{S}', \mathbf{D}, \mathbf{D}' \in \mathbb{R}^{M \times M}$  the diagonal matrices with entries  $S_{ii} := \frac{1}{\psi_{\mathbf{x}}(y_i; p)}$ ,  $D_{ii} := \frac{1}{\psi_y(\mathbf{x}_j; p)}$ ,  $S'_{ii} := \frac{1}{\psi_{\mathbf{x}}(y_i; p')}$ ,  $D'_{ii} := \frac{1}{\psi_y(\mathbf{x}_j; p')}$ . We now have everything we need to state a Lemma relating any two model representations.

Here, we provide the full statement for [Lemma 4.1](#) and prove it:

**Lemma** (Complete version of [Lemma 4.1](#)). *For any  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  satisfying the diversity condition ([Definition 2.1](#)). Let  $\tilde{\mathbf{A}} := \mathbf{L}^{-\top} \mathbf{S}^{-1} \mathbf{S}' \mathbf{L}'^{\top}$  and  $\mathbf{h}_{\mathbf{f}}(\mathbf{x}) := \mathbf{L}^{-\top} \mathbf{S}^{-1} \epsilon_y(\mathbf{x})$ , where the  $i$ -th entry of  $\epsilon_y(\mathbf{x})$  is given by*

$$\epsilon_{y_i}(\mathbf{x}) = \frac{\log p_{\mathbf{f}, \mathbf{g}}(y_i|\mathbf{x}) - \log p_{\mathbf{f}, \mathbf{g}}(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}, \mathbf{g}})} - \frac{\log p_{\mathbf{f}', \mathbf{g}'}(y_i|\mathbf{x}) - \log p_{\mathbf{f}', \mathbf{g}'}(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}', \mathbf{g}'})}.$$

Let  $\mathbf{B} := \mathbf{N}^{-\top} \mathbf{D}^{-1} \mathbf{D}' \mathbf{N}'^{\top}$  and  $\mathbf{h}_{\mathbf{g}}(y) := \mathbf{N}^{-\top} \mathbf{D}^{-1} \epsilon_{\mathbf{x}}(y)$ , where the  $j$ -th entry of  $\epsilon_{\mathbf{x}}(y)$  is given by

$$\epsilon_{\mathbf{x}_j}(y) = \frac{\log p_{\mathbf{f}, \mathbf{g}}(y|\mathbf{x}_j) - \log p_{\mathbf{f}, \mathbf{g}}(y|\mathbf{x}_0)}{\psi_y(\mathbf{x}_j; p_{\mathbf{f}, \mathbf{g}})} - \frac{\log p_{\mathbf{f}', \mathbf{g}'}(y|\mathbf{x}_j) - \log p_{\mathbf{f}', \mathbf{g}'}(y|\mathbf{x}_0)}{\psi_y(\mathbf{x}_j; p_{\mathbf{f}', \mathbf{g}'})}.$$

Then we have:

$$\mathbf{f}(\mathbf{x}) = \tilde{\mathbf{A}} \mathbf{f}'(\mathbf{x}) + \mathbf{h}_{\mathbf{f}}(\mathbf{x}), \quad \mathbf{g}(y) = \mathbf{B} \mathbf{g}'(y) + \mathbf{h}_{\mathbf{g}}(y). \quad (123)$$

*Proof.* Using the definition of  $\epsilon(\mathbf{x})$ , we can relate the embeddings of the two models as follows:

$$\mathbf{S} \mathbf{L} \mathbf{f}(\mathbf{x}) = \mathbf{S}' \mathbf{L}' \mathbf{f}'(\mathbf{x}) + \mathbf{S} \mathbf{L} \mathbf{f}(\mathbf{x}) - \mathbf{S}' \mathbf{L}' \mathbf{f}'(\mathbf{x}) \quad (124)$$

$$= \mathbf{S}' \mathbf{L}' \mathbf{f}'(\mathbf{x}) + \epsilon_y(\mathbf{x}). \quad (125)$$

Therefore, multiplying by the inverse of  $\mathbf{S}$  and of  $\mathbf{L}$ , we get

$$\mathbf{f}(\mathbf{x}) = \mathbf{L}^{-\top} \mathbf{S}^{-1} \mathbf{S}' \mathbf{L}'^{\top} \mathbf{f}'(\mathbf{x}) + \mathbf{L}^{-\top} \mathbf{S}^{-1} \epsilon_y(\mathbf{x}) \quad (126)$$

$$= \tilde{\mathbf{A}} \mathbf{f}'(\mathbf{x}) + \mathbf{h}_{\mathbf{f}}(\mathbf{x}). \quad (127)$$

With similar steps, we obtain the relation between the unembeddings of the two models, proving [Equation \(123\)](#).  $\square$

Notice that the error term  $\epsilon_y(\mathbf{x})$  is a function of  $\mathbf{x}$  and comprises the  $y_i$ 's from the diversity condition ([Definition 2.1](#)). Also,  $\mathbf{B}$  results from the product of the scaling matrices and those constructed from the diversity condition applied to  $\mathbf{f}$ .  $\mathbf{h}_{\mathbf{g}}(y)$  depends on  $\epsilon_{\mathbf{x}}(y)$ , which is a function of  $y$  using the  $\mathbf{x}_i$ 's from the diversity condition.

The following corollary connects [Lemma 4.1](#) to the identifiability results of [Theorem 2.2](#):

**Corollary E.1.** *Under the same assumptions of [Lemma 4.1](#), if we assume that the distributions of the models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  are equal, then we get (as in [Theorem 2.2](#)):*

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{f}'(\mathbf{x}) \quad (128)$$

$$\mathbf{g}_0(y) = \mathbf{A}^{-\top} \mathbf{g}'_0(y), \quad (129)$$

where  $\mathbf{A} = \mathbf{L}^{-\top} \mathbf{L}'^\top$ .

*Proof.* Below, we use the shorthand  $p := p_{\mathbf{f}, \mathbf{g}}$  and  $p' := p_{\mathbf{f}', \mathbf{g}'}$ .

When the two models entail the same distribution, we get that

$$\psi_{\mathbf{x}}(y_i; p) = \sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]} = \psi_{\mathbf{x}}(y_i; p'). \quad (130)$$

which means that

$$\begin{aligned} \epsilon_{y_i}(\mathbf{x}) &= \frac{\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y_i)}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}_i^\top \mathbf{f}(\mathbf{x})]}} - \frac{\mathbf{f}'(\mathbf{x})^\top \mathbf{g}'_0(y_i)}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}'_i{}^\top \mathbf{f}'(\mathbf{x})]}} \\ &= \frac{\log p_{\mathbf{f}, \mathbf{g}}(y_i|\mathbf{x}) - \log p_{\mathbf{f}, \mathbf{g}}(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}, \mathbf{g}})} - \frac{\log p_{\mathbf{f}', \mathbf{g}'}(y_i|\mathbf{x}) - \log p_{\mathbf{f}', \mathbf{g}'}(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}', \mathbf{g}'})} \\ &= \frac{1}{\psi_{\mathbf{x}}(y_i; p_{\mathbf{f}, \mathbf{g}})} (\log p_{\mathbf{f}, \mathbf{g}}(y_i|\mathbf{x}) - \log p_{\mathbf{f}', \mathbf{g}'}(y_i|\mathbf{x}) + \log p_{\mathbf{f}', \mathbf{g}'}(y_0|\mathbf{x}) - \log p_{\mathbf{f}, \mathbf{g}}(y_0|\mathbf{x})) \\ &= 0. \end{aligned}$$

Also, since

$$\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}_i^\top \mathbf{f}(\mathbf{x})]} = \sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]} \quad (131)$$

$$= \sqrt{\text{Var}_{\mathbf{x}}[\log p'(y_i|\mathbf{x}) - \log p'(y_0|\mathbf{x})]} \quad (132)$$

$$= \sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}'_i{}^\top \mathbf{f}'(\mathbf{x})]} \quad (133)$$

and  $\mathbf{S}$  is a diagonal matrix, we have that

$$\mathbf{S}^{-1} \mathbf{S}' = \mathbf{S}^{-1} \mathbf{S} = \mathbf{I}. \quad (134)$$

Finally, we get

$$\mathbf{A} = \mathbf{L}^{-\top} \mathbf{S}^{-1} \mathbf{S}' \mathbf{L}'^\top = \mathbf{L}^{-\top} \mathbf{L}'^\top. \quad (135)$$

and since the error term vanishes, we obtain:

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{f}'(\mathbf{x}), \quad (136)$$

showing the identifiability of the embeddings. The result for the unembeddings proceeds in a similar way, by noticing that also  $\epsilon_{\mathbf{x}}$  is zero, therefore getting the same result of [Theorem 2.2](#).  $\square$

The result of [Lemma 4.1](#) also gives us the following corollary, showing a connection between the distributions and the representations:

**Corollary E.2.** *Under the same assumptions as in [Lemma 4.1](#), we have:*

$$\text{Var}_{\mathbf{x}}[\epsilon_{y_i}(\mathbf{x})] = 2(1 - \text{Corr}[\mathbf{L}_i \mathbf{f}(\mathbf{x}), \mathbf{L}'_i \mathbf{f}'(\mathbf{x})]). \quad (137)$$

*Proof.* We can write

$$\mathbf{S}\mathbf{L}^\top \mathbf{f}(\mathbf{x}) = \mathbf{S}'\mathbf{L}'^\top \mathbf{f}'(\mathbf{x}) + \mathbf{S}\mathbf{L}^\top \mathbf{f}(\mathbf{x}) - \mathbf{S}'\mathbf{L}'^\top \mathbf{f}'(\mathbf{x}). \quad (138)$$

So setting

$$\epsilon_{y_i}(\mathbf{x}) = \mathbf{S}\mathbf{L}^\top \mathbf{f}(\mathbf{x}) - \mathbf{S}'\mathbf{L}'^\top \mathbf{f}'(\mathbf{x}) \quad (139)$$

gives  $\epsilon_{y_i}(\mathbf{x})$  as in [Equation \(8\)](#). Since  $\mathbf{L}^\top$  and  $\mathbf{S}$  are invertible, the inverses exist, and we can left multiply with these inverses. We now consider the variance of  $\epsilon_{y_i}(\mathbf{x})$ . Recall that

$$\text{Var}_{zw}[z + w] = \text{Var}_z[z] + \text{Var}_w[w] + 2\text{Cov}_{zw}[z, w] \quad (140)$$

This results in

$$\text{Var}_{\mathbf{x}}[\epsilon_{y_i}(\mathbf{x})] = \text{Var}_{\mathbf{x}}\left[\frac{\mathbf{L}_i \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}_i \mathbf{f}(\mathbf{x})]}} - \frac{\mathbf{L}'_i \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}'_i \mathbf{f}'(\mathbf{x})]}}\right] \quad (141)$$

$$= \text{Var}_{\mathbf{x}}\left[\frac{\mathbf{L}_i \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}_i \mathbf{f}(\mathbf{x})]}}\right] + \text{Var}_{\mathbf{x}}\left[\frac{\mathbf{L}'_i \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}'_i \mathbf{f}'(\mathbf{x})]}}\right] \quad (142)$$

$$- 2 \text{Cov}_{\mathbf{x}\mathbf{x}}\left[\frac{\mathbf{L}_i \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}_i \mathbf{f}(\mathbf{x})]}} , \frac{\mathbf{L}'_i \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}'_i \mathbf{f}'(\mathbf{x})]}}\right] \quad (143)$$

$$= 2 - 2 \text{Cov}_{\mathbf{x}\mathbf{x}}\left[\frac{\mathbf{L}_i \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}_i \mathbf{f}(\mathbf{x})]}} , \frac{\mathbf{L}'_i \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\mathbf{L}'_i \mathbf{f}'(\mathbf{x})]}}\right] \quad (144)$$

$$= 2(1 - \text{Corr}[\mathbf{L}_i \mathbf{f}(\mathbf{x}), \mathbf{L}'_i \mathbf{f}'(\mathbf{x})]), \quad (145)$$

where we use [Equation \(140\)](#) for the second equality and that  $\text{Var}_z[z/\text{std}(z)] = 1$  in the third equality. We thus have the result.  $\square$

There are three things that are important from this result.

Firstly, as seen in [Corollary E.1](#), when the distributions are equal, the error term becomes zero and we are back in the case of [Theorem 2.2](#).

Secondly, the shape of the error term suggests what a distance should measure if we want it to give a guarantee for how far representations are from being invertible linear transformations of each other. We notice that [Equation \(8\)](#) can be rewritten as

$$\epsilon_{y_i}(\mathbf{x}) = \frac{\log p(y_i|\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]}} - \frac{\log p'(y_i|\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\log p'(y_i|\mathbf{x}) - \log p'(y_0|\mathbf{x})]}} \quad (146)$$

$$+ \frac{\log p'(y_0|\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\log p'(y_i|\mathbf{x}) - \log p'(y_0|\mathbf{x})]}} - \frac{\log p(y_0|\mathbf{x})}{\sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]}} \quad (147)$$

that is, entirely in terms of the logarithm of distributions. We also see that if  $\epsilon_{y_i}(\mathbf{x})$  is a constant (or equivalently,  $\text{Var}[\epsilon_{y_i}(\mathbf{x})] = 0$ ), the embeddings will be invertible linear transformations of each other.

Thirdly, [Lemma 4.1](#) fits with the observation made in [Section 3](#). The KL divergence depends on the difference of the logarithms of distributions, but it also multiplies that difference by the likelihood of the label. Which means that we can have a relatively large difference in logarithms of distributions, even though the KL divergence would result in a small value.

A similar result for the unembeddings can be found below.

**Corollary E.3.** *Under the same assumptions as in [Lemma 4.1](#), we have:*

$$\text{Var}_y[\epsilon_{\mathbf{x}_i}(y)] = 2(1 - \text{Corr}[\mathbf{N}_i \mathbf{g}(y), \mathbf{N}'_i \mathbf{g}'(y)]). \quad (148)$$

*Proof.* We can write

$$\mathbf{S}\mathbf{N}^\top \mathbf{g}(y) = \mathbf{S}'\mathbf{N}'^\top \mathbf{g}'(y) + \mathbf{S}\mathbf{N}^\top \mathbf{g}(y) - \mathbf{S}'\mathbf{N}'^\top \mathbf{g}'(y). \quad (149)$$

So setting

$$\epsilon_{\mathbf{x}}(y) = \mathbf{S}\mathbf{N}^\top \mathbf{g}(y) - \mathbf{S}'\mathbf{N}'^\top \mathbf{g}'(y) \quad (150)$$

gives  $\epsilon_{\mathbf{x}_i}(y)$  as the error term for the unembeddings.

We now consider the variance of  $\epsilon_{\mathbf{x}_i}(y)$ .

$$\text{Var}_y[\epsilon_{\mathbf{x}_i}(y)] = \text{Var}_y \left[ \frac{\mathbf{N}_i \mathbf{g}(y)}{\sqrt{\text{Var}_y[\mathbf{N}_i \mathbf{g}(y)]}} - \frac{\mathbf{N}'_i \mathbf{g}'(y)}{\sqrt{\text{Var}_y[\mathbf{N}'_i \mathbf{g}'(y)]}} \right] \quad (151)$$

$$= \text{Var}_y \left[ \frac{\mathbf{N}_i \mathbf{g}(y)}{\sqrt{\text{Var}_y[\mathbf{N}_i \mathbf{g}(y)]}} \right] + \text{Var}_y \left[ \frac{\mathbf{N}'_i \mathbf{g}'(y)}{\sqrt{\text{Var}_y[\mathbf{N}'_i \mathbf{g}'(y)]}} \right] \quad (152)$$

$$- 2 \text{Cov}_{yy} \left[ \frac{\mathbf{N}_i \mathbf{g}(y)}{\sqrt{\text{Var}_y[\mathbf{N}_i \mathbf{g}(y)]}}, \frac{\mathbf{N}'_i \mathbf{g}'(y)}{\sqrt{\text{Var}_y[\mathbf{N}'_i \mathbf{g}'(y)]}} \right] \quad (153)$$

$$= 2 - 2 \text{Cov}_{yy} \left[ \frac{\mathbf{N}_i \mathbf{g}(y)}{\sqrt{\text{Var}_y[\mathbf{N}_i \mathbf{g}(y)]}}, \frac{\mathbf{N}'_i \mathbf{g}'(y)}{\sqrt{\text{Var}_y[\mathbf{N}'_i \mathbf{g}'(y)]}} \right] \quad (154)$$

$$= 2(1 - \text{Corr}[\mathbf{N}_i \mathbf{g}(y), \mathbf{N}'_i \mathbf{g}'(y)]), \quad (155)$$

which gives us the result.  $\square$

## E.2 Distance Between Distributions

Recall that we use the following notation:

$$\psi_{\mathbf{x}}(y_i; p) := \sqrt{\text{Var}_{\mathbf{x}}[\log p(y_i|\mathbf{x}) - \log p(y_0|\mathbf{x})]} \quad \text{and} \quad \psi_y(\mathbf{x}_j; p) := \sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_0)]}.$$

We restate [Definition 4.4](#) and then prove its a distance metric.

**Definition 4.4** (Log-likelihood variance distance between distributions). *For any two probability distributions  $p, p'$  for which there exists a common choice of  $\mathcal{X}_{\text{LLV}} \subset \mathcal{X}$  and  $\mathcal{Y}_{\text{LLV}} \subset \mathcal{Y}$  such that they both satisfy [Assumption 4.3](#), for a fixed  $\lambda \in \mathbb{R}^+$ , and by considering the following terms:*

$$\begin{aligned} t_1 &:= \max_{y \in \mathcal{Y}_{\text{LLV}} \setminus \{y_0\}} \left\{ \sqrt{\text{Var}_{\mathbf{x}} \left[ \frac{\log p(y|\mathbf{x})}{\psi_{\mathbf{x}}(y; p)} - \frac{\log p'(y|\mathbf{x})}{\psi_{\mathbf{x}}(y; p')} \right]}, \sqrt{\text{Var}_{\mathbf{x}} \left[ \frac{\log p(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y; p)} - \frac{\log p'(y_0|\mathbf{x})}{\psi_{\mathbf{x}}(y; p')} \right]} \right\} \\ t_2 &:= \max_{\mathbf{x} \in \mathcal{X}_{\text{LLV}} \setminus \{\mathbf{x}_0\}} \left\{ \sqrt{\text{Var}_y \left[ \frac{\log p(y|\mathbf{x})}{\psi_y(\mathbf{x}; p)} - \frac{\log p'(y|\mathbf{x})}{\psi_y(\mathbf{x}; p')} \right]}, \sqrt{\text{Var}_y \left[ \frac{\log p(y|\mathbf{x}_0)}{\psi_y(\mathbf{x}; p)} - \frac{\log p'(y|\mathbf{x}_0)}{\psi_y(\mathbf{x}; p')} \right]} \right\} \\ t_3 &:= \max_{y \in \mathcal{Y}_{\text{LLV}} \setminus \{y_0\}} |\psi_{\mathbf{x}}(y; p) - \psi_{\mathbf{x}}(y; p')|, \quad t_4 := \max_{\mathbf{x} \in \mathcal{X}_{\text{LLV}} \setminus \{\mathbf{x}_0\}} |\psi_y(\mathbf{x}; p) - \psi_y(\mathbf{x}; p')|, \end{aligned}$$

the log-likelihood variance (LLV) distance between  $p$  and  $p'$  is given by

$$d_{\text{LLV}}^\lambda(p, p') := \max \{t_1, t_2, \lambda t_3, \lambda t_4\}. \quad (10)$$

To show that  $d_{\text{LLV}}^\lambda$  is a distance metric between sets of conditional probability distributions, we have to prove that: i) it is non-negative, ii) it is zero if and only if the models are equal (that is, the distributions are equal for all labels and all inputs) iii) it is symmetric and iv) it satisfies the triangle inequality.

*Proof.* i) Since variance is non-negative, the first two terms are non-negative. The second two terms are absolute values, so also non-negative. Thus, the maximum of these four terms is also non-negative.

ii) From the expression of  $d_{\text{LLV}}^\lambda$ , if the distributions are equal for all  $\mathbf{x} \in \mathcal{X}$  and all  $y \in \mathcal{Y}$ , then the distance becomes zero.

Now, assume the distance is zero. Then, in particular,  $t_4 = 0$ . So we have

$$\sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_0)]} = \sqrt{\text{Var}_y[\log p'(y|\mathbf{x}_j) - \log p'(y|\mathbf{x}_0)]} \quad (156)$$

for all  $\mathbf{x}_j \in \mathcal{X}_{\text{LLV}} \setminus \{\mathbf{x}_0\}$ . So since  $t_2 = 0$ , we have

$$\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p'(y|\mathbf{x}_j)] = 0, \quad (157)$$

for all  $\mathbf{x}_j \in \mathcal{X}_{\text{LLV}}$ . This means that for each  $\mathbf{x}_j$ , we have that the log difference of the distributions is a constant for all  $y_i \in \mathcal{Y}_{\text{LLV}}$ , and so

$$\log p(y_i|\mathbf{x}_j) - \log p'(y_i|\mathbf{x}_j) = c_j, \quad (158)$$

for all  $y_i$ . This means that

$$\log p(y_i|\mathbf{x}_j) = \log p'(y_i|\mathbf{x}_j) + c_j \quad (159)$$

$$p(y_i|\mathbf{x}_j) = p'(y_i|\mathbf{x}_j) \cdot \exp(c_j). \quad (160)$$

Now since  $p$  and  $p'$  are probability distributions over the  $y_i$ 's, we have

$$1 = \sum_{i=1}^c p(y_i|\mathbf{x}_j) = \exp(c_j) \sum_{i=1}^c p'(y_i|\mathbf{x}_j) = \exp(c_j) \quad (161)$$

This means that  $p(y|\mathbf{x}_j) = p'(y|\mathbf{x}_j)$  for all  $y \in \mathcal{Y}$  and for the  $\mathbf{x}_j \in \mathcal{X}_{LLV}$ . Since we have  $t_1 = t_3 = 0$ , we get that

$$\text{Var}_{\mathbf{x}} [\log p(y_i|\mathbf{x}) - \log p'(y_i|\mathbf{x})] = 0, \quad (162)$$

for all  $y_i \in \mathcal{Y}_{LLV}$  and all inputs  $\mathbf{x} \in \mathcal{X}$ . Thus, we have

$$\log p(y_i|\mathbf{x}) - \log p'(y_i|\mathbf{x}) = c_i \quad (163)$$

for every choice of  $y_i$  except one left out of the label set and all  $\mathbf{x}$ . This results in

$$p(y_i|\mathbf{x}) = p'(y_i|\mathbf{x}) \cdot \exp(c_i) \quad (164)$$

for all  $\mathbf{x} \in \mathcal{X}$ . In particular, this is true for  $\mathbf{x}_j \in \mathcal{X}_{LLV}$ . So we have

$$p(y_i|\mathbf{x}_j) = p'(y_i|\mathbf{x}_j) \cdot \exp(c_i) = p'(y_i|\mathbf{x}_j) \quad (165)$$

and thus,  $\exp(c_i) = 1$  for all  $y_i \in \mathcal{Y}_{LLV}$ . Since the probability of the last label,  $y_k$  is fixed by the other labels, i.e.

$$p(y_k|\mathbf{x}) = 1 - \sum_{i \neq k} p(y_i|\mathbf{x}) \quad (166)$$

this gives us that

$$p(y_i|\mathbf{x}) = p'(y_i|\mathbf{x}) \quad (167)$$

for all  $y_i \in \mathcal{Y}$  and all  $\mathbf{x} \in \mathcal{X}$ .

iii)  $d_{LLV}^\lambda$  is symmetric because it depends on variances and on absolute values, which are symmetric.

iv) We prove that  $d_{LLV}^\lambda$  satisfies the triangle inequality by showing that all terms in it satisfy the triangle inequality. Assume we have three random variables  $X, Y, Z$ , then

$$\sqrt{\text{Var}[X - Z]} = \sqrt{\text{Var}[X - Y + Y - Z]} \quad (168)$$

$$= \sqrt{\text{Var}[X - Y] + \text{Var}[Y - Z] + 2 \text{Cov}[X - Y, Y - Z]} \quad (169)$$

$$\leq \sqrt{\text{Var}[X - Y] + \text{Var}[Y - Z] + 2\sqrt{\text{Var}[X - Y]}\sqrt{\text{Var}[Y - Z]}} \quad (170)$$

$$= \sqrt{\left(\sqrt{\text{Var}[X - Y]} + \sqrt{\text{Var}[Y - Z]}\right)^2} \quad (171)$$

$$= \sqrt{\text{Var}[X - Y]} + \sqrt{\text{Var}[Y - Z]}. \quad (172)$$

So the square root of the variance of differences of random variables satisfies the triangle inequality. Taking the maximum also preserves the triangle inequality. Therefore,  $t_1$  and  $t_2$  satisfy the triangle inequality. Concerning the last two terms  $t_3$  and  $t_4$ , we see that for models  $p, p', p^*$  we have

$$\begin{aligned} & \left| \sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_k)]} - \sqrt{\text{Var}_y[\log p'(y|\mathbf{x}_j) - \log p'(y|\mathbf{x}_k)]} \right| \\ &= \left| \sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_k)]} - \sqrt{\text{Var}_y[\log p^*(y|\mathbf{x}_j) - \log p^*(y|\mathbf{x}_k)]} \right| \\ & \quad + \left| \sqrt{\text{Var}_y[\log p^*(y|\mathbf{x}_j) - \log p^*(y|\mathbf{x}_k)]} - \sqrt{\text{Var}_y[\log p'(y|\mathbf{x}_j) - \log p'(y|\mathbf{x}_k)]} \right| \\ &\leq \left| \sqrt{\text{Var}_y[\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_k)]} - \sqrt{\text{Var}_y[\log p^*(y|\mathbf{x}_j) - \log p^*(y|\mathbf{x}_k)]} \right| \\ & \quad + \left| \sqrt{\text{Var}_y[\log p^*(y|\mathbf{x}_j) - \log p^*(y|\mathbf{x}_k)]} - \sqrt{\text{Var}_y[\log p'(y|\mathbf{x}_j) - \log p'(y|\mathbf{x}_k)]} \right|. \end{aligned}$$

So the second two terms also satisfy the triangle inequality. Thus, the sum of these four terms satisfies the triangle inequality.  $\square$

### E.3 Implementation of the Log-Likelihood Variance Distance

Here, we collect the implementation details of  $d_{LLV}^\lambda$ . We set the weighting constant  $\lambda$  to  $10^{-5}$ . To choose the best pivot  $y_0 \in \mathcal{Y}$  and the left-out label  $y_\emptyset \in \mathcal{Y}$ , we evaluate  $t_1$  for all possible choices of  $(y_0, y_\emptyset)$ , then select those giving the smallest value of  $d_{LLV}^\lambda$ . For the input set  $\mathcal{X}_{LLV}$ , we use  $M + 1$  inputs. In our experiments, the representational dimension is 2, so we collect 3 inputs. We randomly draw 200 samples of sets from  $\mathcal{X}$  containing 3 inputs, and choose the set giving the smallest  $t_2$ .

In our experiments, we only make pairwise comparisons of models. Notice that, in the case where one would like to compare three or more models, the same pivot  $y_0$ , left-out label  $y_\emptyset$ , and input set  $\mathcal{X}_{LLV}$  must be chosen for all models.

### E.4 Using PLS SVD to Define a Dissimilarity Measure

We restate [Definition 4.5](#) and then prove some important properties.

**Definition 4.5** (PLS SVD distance between vectors). *Let  $\mathbf{z}, \mathbf{w}$  be two  $M$ -dimensional random vectors, and define  $\mathbf{z}', \mathbf{w}'$  by standardizing their components:  $z'_i = (z_i - \mathbb{E}[z_i]) / \text{std}(z_i)$ ,  $w'_i = (w_i - \mathbb{E}[w_i]) / \text{std}(w_i)$ . Let  $\{\mathbf{u}_i\}_{i=1}^M$  and  $\{\mathbf{v}_i\}_{i=1}^M$  be the left and right singular vectors of the cross-covariance matrix  $\Sigma_{\mathbf{z}'\mathbf{w}'}$ . We define*

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) := \frac{1}{M} \sum_{i=1}^M \text{Cov}_{\mathbf{z}'\mathbf{w}'}[\mathbf{u}_i^\top \mathbf{z}', \mathbf{v}_i^\top \mathbf{w}'], \quad d_{\text{SVD}}(\mathbf{z}, \mathbf{w}) := 1 - m_{\text{SVD}}(\mathbf{z}, \mathbf{w}). \quad (12)$$

We show that this measure is zero from a vector to itself and is invariant to multiplications of the scaled variables with orthonormal matrices. Before proceeding, we recall the following property of the trace of two matrices:

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}). \quad (173)$$

In the next Proposition, we prove that  $d_{\text{SVD}}(\mathbf{z}, \mathbf{z}) = 0$ .

**Proposition E.4.** *Let  $\mathbf{z}$  be an  $M$ -dimensional random variable and assume that the covariance matrix  $\Sigma_{\mathbf{z}'\mathbf{z}'}$  of the centered and scaled variable  $\mathbf{z}'$  is non-singular. Then*

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{z}) = \frac{1}{M} \sum_{i=1}^M \text{Cov}_{\mathbf{z}'\mathbf{z}'}[\mathbf{u}_i^\top \mathbf{z}', \mathbf{u}_i^\top \mathbf{z}'] = 1, \quad (174)$$

where  $\mathbf{u}_i$  is the  $i$ 'th singular vector of  $\Sigma_{\mathbf{z}'\mathbf{z}'}$ .

*Proof.* Let

$$\Sigma_{\mathbf{z}'\mathbf{z}'} = \mathbf{UDU}^\top \quad (175)$$

be the singular value decomposition of the covariance matrix of  $\mathbf{z}'$  with itself. Then, using [Equation \(173\)](#), we have that

$$\text{tr}(\Sigma_{\mathbf{z}'\mathbf{z}'}) = \text{tr}(\mathbf{UDU}^\top) = \text{tr}(\mathbf{U}^\top \mathbf{UD}) = \text{tr}(\mathbf{D}) = \sum_{i=1}^M \sigma_i, \quad (176)$$

where  $\sigma_i$  are the singular values of  $\Sigma_{\mathbf{z}'\mathbf{z}'}$ . Since  $\text{Var}[z'_i] = 1$  for all  $i \in \{1, \dots, M\}$ , we also have that

$$\text{tr}(\Sigma_{\mathbf{z}'\mathbf{z}'}) = \sum_{i=1}^M \text{Var}[z'_i] = M. \quad (177)$$

Combining the two results, we get

$$\sum_{i=1}^M \sigma_i = M. \quad (178)$$

Therefore, the mean of the covariances becomes

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{z}) = \frac{1}{M} \sum_{i=1}^M \text{Cov}_{\mathbf{z}'\mathbf{z}'}[\mathbf{u}_i^\top \mathbf{z}', \mathbf{u}_i^\top \mathbf{z}'] \quad (179)$$

$$= \frac{1}{M} \text{tr}(\mathbf{U}^\top \text{Cov}_{\mathbf{z}'\mathbf{z}'}[\mathbf{z}', \mathbf{z}'] \mathbf{U}) \quad (180)$$

$$= \frac{1}{M} \sum_{i=1}^M \sigma_i = \frac{M}{M} = 1 \quad (181)$$

which proves the claim.  $\square$

Next, we prove the invariance to translation and orthogonal transformations of the members in  $m_{\text{SVD}}$ :

**Proposition E.5.** *Let  $\mathbf{z}$  and  $\mathbf{w}$  be  $M$ -dimensional random variables. Then, the measure*

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) = \frac{1}{M} \sum_{i=1}^M \text{Cov}_{\mathbf{z}'\mathbf{w}'}[\mathbf{u}_i^\top \mathbf{z}', \mathbf{v}_i^\top \mathbf{w}'] \quad (182)$$

*is invariant to translations and multiplications with an orthonormal matrix after the scaling.*

*Proof. Invariance to translations.* Since  $m_{\text{SVD}}(\mathbf{z}, \mathbf{w})$  is based on covariances, and covariances are invariant to translations,  $m_{\text{SVD}}(\mathbf{z}, \mathbf{w})$  is invariant to translations.

**Invariance to multiplication with orthonormal matrix.** Assume  $\mathbf{T}$  and  $\mathbf{R}$  are orthonormal matrices. Let  $\mathbf{h} = \mathbf{T}\mathbf{z}'$  and  $\mathbf{k} = \mathbf{R}\mathbf{w}'$ . Let the singular value decomposition of  $\Sigma_{\mathbf{z}'\mathbf{w}'}$  be:

$$\Sigma_{\mathbf{z}'\mathbf{w}'} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \quad (183)$$

then

$$\Sigma_{\mathbf{h}\mathbf{k}} = \mathbf{T}\mathbf{U}\mathbf{D}\mathbf{V}^\top \mathbf{R}^\top \quad (184)$$

$$= (\mathbf{T}\mathbf{U})\mathbf{D}(\mathbf{R}\mathbf{V})^\top. \quad (185)$$

Since  $\mathbf{T}$  and  $\mathbf{R}$  are orthonormal,  $\mathbf{T}\mathbf{U}$  and  $\mathbf{R}\mathbf{V}$  are also orthonormal, so  $\text{Cov}_{\mathbf{h}\mathbf{k}}[\mathbf{h}, \mathbf{k}]$  has the same singular values as  $\Sigma_{\mathbf{z}'\mathbf{w}'}$ . Thus, the measure is invariant to multiplications with orthonormal matrices (rotations) after the scaling.  $\square$

## E.5 Lower Bound on Mean PLS-SVD

To show how the probability distributions are connected to the representations, we need an intermediate result, where we bound  $m_{\text{SVD}}$ . To prove this bound, we will make use of Weyl's Inequality and we therefore restate it from [59].

**Theorem E.6** (Weyl's Inequality). *Let  $\mathbf{A}$  be a  $n \times m$  matrix. Let  $\mathbf{B} = \mathbf{A} + \mathbf{E}$ . Let  $\sigma_i$  be the  $i$ 'th singular value of  $\mathbf{A}$  and let  $\tilde{\sigma}_i$  be the  $i$ 'th singular value of  $\mathbf{B}$  (ordered from largest to smallest). Then*

$$|\sigma_i - \tilde{\sigma}_i| \leq \|\mathbf{E}\|_s, \quad (186)$$

where  $\|\cdot\|_s$  is the spectral norm defined as

$$\|\mathbf{E}\|_s = \max_{\|\mathbf{v}\|=1} \|\mathbf{E}\mathbf{v}\|_2 \quad (187)$$

and  $\|\cdot\|_2$  is the Euclidean vector norm. Or equivalently:

$$\|\mathbf{E}\|_s = \sigma_{\max}(\mathbf{E}), \quad (188)$$

where  $\sigma_{\max}(\mathbf{E})$  is the largest singular value of  $\mathbf{E}$ .

We now state and prove a bound on  $m_{\text{SVD}}$ .

**Lemma E.7.** Let  $\mathbf{z}, \mathbf{w}$  be two  $M$ -dimensional random vectors, and define  $\mathbf{z}', \mathbf{w}'$  by standardizing their components:  $z'_i = (z_i - \mathbb{E}[z_i]) / \text{std}(z_i)$ ,  $w'_i = (w_i - \mathbb{E}[w_i]) / \text{std}(w_i)$ . Assume the  $M \times M$  matrices  $\Sigma_{\mathbf{z}'\mathbf{z}'}$  and  $\Sigma_{\mathbf{z}'\mathbf{w}'}$   $\in \mathbb{R}^{M \times M}$  are full-rank. Let  $\{\mathbf{u}_i\}_{i=1}^M$  and  $\{\mathbf{v}_i\}_{i=1}^M$  be the left and right singular vectors of  $\Sigma_{\mathbf{z}'\mathbf{w}'}$ . Then, the following bound holds:

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) = \frac{1}{M} \sum_{i=1}^M \text{Cov}_{\mathbf{z}'\mathbf{w}'}[\mathbf{u}_i^\top \mathbf{z}', \mathbf{v}_i^\top \mathbf{w}'] \geq 1 - \sqrt{M \sum_{l=1}^M \text{Var}_{\mathbf{z}', \mathbf{w}'}[z'_l - w'_l]} \quad (189)$$

where  $\text{Var}_{\mathbf{z}, \mathbf{w}}[z'_l - w'_l]$  denotes the variance over the joint distribution over  $\mathbf{z}'$  and  $\mathbf{w}'$ .

*Proof.* For ease of notation, assume  $\mathbf{z}$  and  $\mathbf{w}$  are  $M$ -dimensional random variables which are already centered and scaled. So  $\text{Var}[z_i] = 1$  for all components  $i = 1, \dots, M$ . Below we will make use of the covariance inequality, which says that for scalar variables  $z, w$ ,

$$-\sqrt{\text{Var}_z[z]} \sqrt{\text{Var}_w[w]} \leq \text{Cov}_{zw}[z, w] \leq \sqrt{\text{Var}_z[z]} \sqrt{\text{Var}_w[w]}. \quad (190)$$

Let  $\mathbf{A} := \Sigma_{\mathbf{z}\mathbf{w}}$  be the cross-covariance matrix of  $\mathbf{z}$  and  $\mathbf{w}$ , and let  $\mathbf{B} := \Sigma_{\mathbf{z}\mathbf{z}}$  be the covariance matrix of  $\mathbf{z}$ . Then

$$\mathbf{B} = \mathbf{A} + (\mathbf{B} - \mathbf{A}), \quad (191)$$

and we define  $\mathbf{E} := \mathbf{B} - \mathbf{A}$ . We can write the  $i, j$ 'th entry of  $\mathbf{E}$  as

$$(\mathbf{E})_{i,j} = \text{Cov}_{\mathbf{z}\mathbf{z}}[z_i, z_j] - \text{Cov}_{\mathbf{z}\mathbf{w}}[z_i, w_j] = \text{Cov}_{\mathbf{z}\mathbf{w}}[z_i, z_j - w_j], \quad (192)$$

because of the bilinearity of the covariance, see section 13.2.7 of Adhikari and Pitman [1], "The Main Property: Bilinearity".

The singular value decompositions of  $\mathbf{A}$  and  $\mathbf{B}$  always exist and can be written as follows:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \quad \mathbf{B} = \mathbf{W}\Sigma\mathbf{W}^\top, \quad (193)$$

where  $\mathbf{U}, \mathbf{V}$  and  $\mathbf{W}$  are orthonormal matrices and  $\mathbf{D}$  and  $\Sigma$  are diagonal matrices containing the singular values. Note that since  $\mathbf{B}$  is symmetric and positive definite, the left and right singular vectors coincide. Let  $\tilde{\sigma}_i$  be the  $i$ 'th singular value of  $\mathbf{A}$  (sorted from largest to smallest) and let  $\sigma_i$  be the  $i$ 'th singular value of  $\mathbf{B}$ . Now Weyl's inequality E.6 gives us that

$$|\sigma_i - \tilde{\sigma}_i| \leq \|\mathbf{E}\|_s. \quad (194)$$

Since the spectral norm of  $\mathbf{E}$  is the largest singular value of  $\mathbf{E}$

$$\|\mathbf{E}\|_s = \sigma_{\max}(\mathbf{E}) \quad (195)$$

and the frobenius norm is the square root of the sum of the squared singular values

$$\|\mathbf{E}\|_F = \sqrt{\sum_{i=1}^M \sigma_i^2(\mathbf{E})} = \sqrt{\text{trace}(\mathbf{E}^\top \mathbf{E})}, \quad (196)$$

we have that

$$\|\mathbf{E}\|_s \leq \|\mathbf{E}\|_F = \sqrt{\text{trace}(\mathbf{E}^\top \mathbf{E})}. \quad (197)$$

Next, we bound the trace of  $\mathbf{E}^\top \mathbf{E}$ :

$$\text{trace}(\mathbf{E}^\top \mathbf{E}) = \sum_{l=1}^M \sum_{k=1}^M \text{Cov}_{\mathbf{z}\mathbf{w}}[z_k, z_l - w_l]^2 \quad (198)$$

$$\leq \sum_{l=1}^M \sum_{k=1}^M (\sqrt{\text{Var}_z[z_k]} \sqrt{\text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]})^2 \quad (199)$$

$$= \sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l], \quad (200)$$

where in the first step we use the covariance inequality (Equation (190)) and in the second we use that  $\text{Var}[z_k] = 1$  by construction. Combining the inequalities from Equations (194), (197) and (200) we obtain

$$|\sigma_i - \tilde{\sigma}_i| \leq \sqrt{\sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]}, \quad (201)$$

which means that the difference between  $\sigma_i$  and  $\tilde{\sigma}_i$  is at most  $\sqrt{\sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]}$ . This means that if  $\tilde{\sigma}_i < \sigma_i$ , we still have :

$$\tilde{\sigma}_i \geq \sigma_i - \sqrt{\sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]}. \quad (202)$$

and if  $\tilde{\sigma}_i \geq \sigma_i$ , we also have

$$\tilde{\sigma}_i \geq \sigma_i - \sqrt{\sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]}. \quad (203)$$

since  $\sqrt{\sum_{l=1}^M M \text{Var}[z_l - w_l]} \geq 0$ . Let  $\mathbf{v}_i$  be the  $i$ 'th column of  $\mathbf{V}$  and recall that from the SVD of we have

$$\tilde{\sigma}_i = \mathbf{u}_i^\top \mathbf{A} \mathbf{v}_i = \text{Cov}_{\mathbf{z}\mathbf{w}} \left[ \sum_{k=1}^M \mathbf{u}_{ik} z_k, \sum_{l=1}^M \mathbf{v}_{il} w_l \right] \quad (204)$$

$$= \text{Cov}_{\mathbf{z}\mathbf{w}} [\mathbf{u}_i^\top \mathbf{z}, \mathbf{v}_i^\top \mathbf{w}], \quad (205)$$

because of the bilinearity of the covariance. Thus, Equation (202) gives us the bound

$$\text{Cov}_{\mathbf{z}\mathbf{w}} [\mathbf{u}_i^\top \mathbf{z}, \mathbf{v}_i^\top \mathbf{w}] \geq \sigma_i - \sqrt{\sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]} \quad (206)$$

and therefore

$$m_{\text{SVD}}(\mathbf{z}, \mathbf{w}) = \frac{1}{m} \sum_{i=1}^m [\text{Cov}_{\mathbf{z}\mathbf{w}} [\mathbf{u}_i^\top \mathbf{z}, \mathbf{v}_i^\top \mathbf{w}]] \geq 1 - \sqrt{\sum_{l=1}^M M \text{Var}_{\mathbf{z}, \mathbf{w}}[z_l - w_l]}, \quad (207)$$

where we used that  $\frac{1}{m} \sum_{i=1}^m \sigma_i = 1$  from Proposition E.4.  $\square$

## E.6 Bounding Representational Similarity with Distribution Distance

We can now use Definitions 4.4 and 4.5 and Lemma E.7 to show how a bound on the distance between probability distributions can give us a bound on the distance between representations. For the result below, let  $\mathbf{L}, \mathbf{L}'$  be defined as in Theorem 2.2. Let  $\mathbf{N}$  (resp.  $\mathbf{N}'$ ) be the matrix with columns  $\mathbf{f}_0(\mathbf{x})$  (resp.  $\mathbf{f}'_0(\mathbf{x})$ ). We denote with  $\mathbf{z}_1 := \mathbf{L}^\top \mathbf{f}(\mathbf{x})$ ,  $\mathbf{z}_2 := \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})$ ,  $\mathbf{w}_1 := \mathbf{N}^\top \mathbf{g}(y)$  and  $\mathbf{w}_2 := \mathbf{N}'^\top \mathbf{g}'(y)$ .

**Theorem 4.7.** *Let  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  be two models such that: (1) There exist  $\mathcal{X}_{\text{LLV}} \subset \mathcal{X}$  and  $\mathcal{Y}_{\text{LLV}} \subset \mathcal{Y}$ , consisting of a pivot point and all labels but one, such that all  $\mathbf{L}, \mathbf{L}'$  and  $\mathbf{N}, \mathbf{N}'$  matrices constructed from these sets are invertible;<sup>16</sup> (2) Both  $p_{\mathbf{f}, \mathbf{g}}$  and  $p_{\mathbf{f}', \mathbf{g}'}$  satisfy Assumption 4.3 for  $\mathcal{X}_{\text{LLV}}$  and  $\mathcal{Y}_{\text{LLV}}$ ; (3) The covariance matrices  $\Sigma_{\mathbf{z}_1 \mathbf{z}_1}$ ,  $\Sigma_{\mathbf{w}_1 \mathbf{w}_1}$  and the cross-covariance matrices  $\Sigma_{\mathbf{z}_1 \mathbf{z}_2}$ ,  $\Sigma_{\mathbf{w}_1 \mathbf{w}_2}$  are non-singular. Then, for any weighting constant  $\lambda \in \mathbb{R}^+$ , we have*

$$d_{\text{LLV}}^\lambda(p_{\mathbf{f}, \mathbf{g}}, p_{\mathbf{f}', \mathbf{g}'}) \leq \epsilon \implies d_{\mathbf{f}, \mathbf{g}}(\mathbf{f}, \mathbf{g}, \mathbf{f}', \mathbf{g}') \leq 2M\epsilon. \quad (14)$$

<sup>16</sup>This is slightly stronger than the diversity condition (Definition 2.1) in the sense that we need diversity to hold for all sets using labels from  $\mathcal{Y}_{\text{LLV}}$  and the chosen pivot point.

*Proof.* We start from

$$d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) = 1 - m_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) \quad (208)$$

and consider the components  $\mathbf{L}_l, \mathbf{L}'_l$ , which are  $l$ 'th row of  $\mathbf{L}^\top, \mathbf{L}'^\top$ , respectively. In the following, notice that

$$\psi_{\mathbf{x}}(y_l; p) := \sqrt{\text{Var}_{\mathbf{x}}[\log p(y_l|\mathbf{x}) - \log p(y_0|\mathbf{x})]} = \sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]} \quad (209)$$

Using [Lemma E.7](#), we have that

$$m_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) \geq 1 - \sqrt{\sum_{l=1}^M M \text{Var} \left[ \frac{\mathbf{L}_l \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\mathbf{L}'_l \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right]}. \quad (210)$$

Therefore, we have

$$d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) \leq \sqrt{\sum_{l=1}^M M \text{Var} \left[ \frac{\mathbf{L}_l \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\mathbf{L}'_l \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right]}. \quad (211)$$

Considering the variance term, we can rewrite it as follows:

$$\begin{aligned} \text{Var} \left[ \frac{\mathbf{L}_l \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\mathbf{L}'_l \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right] &= \text{Var} \left[ \frac{\log p(y_l|x) - \log p(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\log p'(y_l|x) - \log p'(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right] \\ &= \text{Var} \left[ \frac{\log p(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\log p'(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right] \\ &\quad + \text{Var} \left[ \frac{\log p'(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} - \frac{\log p(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} \right] \\ &\quad + 2 \text{Cov} \left[ \frac{\log p(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\log p'(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right], \frac{\log p'(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} - \frac{\log p(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} \left] \\ &\leq \text{Var} \left[ \frac{\log p(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\log p'(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right] \\ &\quad + \text{Var} \left[ \frac{\log p'(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} - \frac{\log p(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} \right] \\ &\quad + 2 \sqrt{\text{Var} \left[ \frac{\log p(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\log p'(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right]} \sqrt{\text{Var} \left[ \frac{\log p'(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} - \frac{\log p(y_0|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} \right]}. \end{aligned}$$

By assumption,  $d_{\text{LLV}}^\lambda(p, p') \leq \epsilon$ , which means that for all  $y_l \in \mathcal{Y}_{\text{LLV}}$ :

$$\sqrt{\text{Var} \left[ \frac{\log p(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\log p'(y_l|x)}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right]} \leq \epsilon. \quad (212)$$

Hence, we obtain

$$d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) \leq \sqrt{\sum_{l=1}^M M \text{Var} \left[ \frac{\mathbf{L}_l \mathbf{f}(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}_l \mathbf{f}(\mathbf{x})]}} - \frac{\mathbf{L}'_l \mathbf{f}'(\mathbf{x})}{\sqrt{\text{Var}[\mathbf{L}'_l \mathbf{f}'(\mathbf{x})]}} \right]} \quad (213)$$

$$\leq \sqrt{\sum_{l=1}^M M(\epsilon^2 + \epsilon^2 + 2\epsilon^2)} \quad (214)$$

$$= \sqrt{M^2 4\epsilon^2} = 2M\epsilon, \quad (215)$$

giving the first part of the result. Next, we consider

$$d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) = 1 - m_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)). \quad (216)$$

Let  $\mathbf{N}_j, \mathbf{N}'_j$  be the  $j$ 'th row of  $\mathbf{N}^\top, \mathbf{N}'^\top$ . Using [Lemma E.7](#), we obtain:

$$m_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) \geq 1 - \sqrt{\sum_{l=1}^M M \text{Var} \left[ \frac{\mathbf{N}_j \mathbf{g}(y)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\mathbf{N}'_j \mathbf{g}'(y)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right]}. \quad (217)$$

This implies the following:

$$d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) \leq \sqrt{\sum_{l=1}^M M \text{Var} \left[ \frac{\mathbf{N}_j \mathbf{g}(y)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\mathbf{N}'_j \mathbf{g}'(y)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right]}. \quad (218)$$

Considering the variance term, we can rework it as follows:

$$\begin{aligned} \text{Var} \left[ \frac{\mathbf{N}_j \mathbf{g}(y)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\mathbf{N}'_j \mathbf{g}'(y)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right] &= \text{Var} \left[ \frac{\log p(y|\mathbf{x}_j) - \log p(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_j) - \log p'(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right] \\ &= \text{Var} \left[ \frac{\log p(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right] \\ &\quad + \text{Var} \left[ \frac{\log p'(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} - \frac{\log p(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} \right] \\ &\quad + 2 \text{Cov} \left[ \frac{\log p(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}}, \frac{\log p'(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} - \frac{\log p(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} \right] \\ &\leq \text{Var} \left[ \frac{\log p(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right] \\ &\quad + \text{Var} \left[ \frac{\log p'(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} - \frac{\log p(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} \right] \\ &\quad + 2 \sqrt{\text{Var} \left[ \frac{\log p(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right]} \sqrt{\text{Var} \left[ \frac{\log p'(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} - \frac{\log p(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} \right]}. \end{aligned}$$

Since by assumption  $d_{\text{LLV}}^\lambda(p, p') \leq \epsilon$ , this implies that  $t_2 \leq \epsilon$  and that for  $\mathbf{x}_j \in \mathcal{X}_{\text{LLV}}$ :

$$\sqrt{\text{Var} \left[ \frac{\log p(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_j)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right]} \leq \epsilon \quad (219)$$

and

$$\sqrt{\text{Var} \left[ \frac{\log p(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\log p'(y|\mathbf{x}_0)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right]} \leq \epsilon. \quad (220)$$

Therefore, we get:

$$d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) \leq \sqrt{\sum_{l=1}^M M \text{Var} \left[ \frac{\mathbf{N}_j \mathbf{g}(y)}{\sqrt{\text{Var}[\mathbf{N}_j \mathbf{g}(y)]}} - \frac{\mathbf{N}'_j \mathbf{g}'(y)}{\sqrt{\text{Var}[\mathbf{N}'_j \mathbf{g}'(y)]}} \right]} \quad (221)$$

$$\leq \sqrt{\sum_{l=1}^M M(\epsilon^2 + \epsilon^2 + 2\epsilon^2)} \quad (222)$$

$$= \sqrt{M^2 4\epsilon^2} = 2M\epsilon \quad (223)$$

showing the second part of the result. Taking the maximum between [Equation \(215\)](#) and [Equation \(223\)](#), we get the result of the claim.  $\square$

To see how this result connects to how far the embedding functions  $\mathbf{f}(\mathbf{x}), \mathbf{f}'(\mathbf{x})$  are from being invertible linear transformations of each other, notice that if  $\mathbf{L}^\top$  and  $\mathbf{L}'^\top$  are both invertible, then if there exists an invertible linear transformation,  $\mathbf{B}$ , such that  $\mathbf{L}^\top \mathbf{f}(\mathbf{x}) = \mathbf{B} \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})$ , then we also have an invertible linear transformation  $\mathbf{A} = \mathbf{L}^{-\top} \mathbf{B} \mathbf{L}'^\top$  such that  $\mathbf{f}(\mathbf{x}) = \mathbf{A} \mathbf{f}'(\mathbf{x})$ .

Since  $m_{\text{SVD}}(\mathbf{x}, y)$  is always non-negative, for this bound in [Theorem 4.7](#) to be non-vacuous, we need

$$d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) \leq 2M\epsilon < 1 \implies \epsilon < \frac{1}{2M}. \quad (224)$$

This means that for higher dimensional representations, we need the variance of the differences of log-likelihoods to be smaller, if we want a guarantee from this result.

Next, we prove that the dissimilarity between representations induced by [Theorem 4.7](#) is invariant to substituting the members with models that are  $\sim_L$ -equivalent.

**Lemma E.8.** *For any two models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$ , and for any other model  $(\mathbf{f}^*, \mathbf{g}^*) \in \Theta$  such that  $(\mathbf{f}^*, \mathbf{g}^*) \sim_L (\mathbf{f}, \mathbf{g})$ , we have that:*

$$d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) = d_{\text{SVD}}(\mathbf{L}^{*\top} \mathbf{f}^*(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})), \quad (225)$$

$$d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) = d_{\text{SVD}}(\mathbf{N}^{*\top} \mathbf{g}^*(y), \mathbf{N}'^\top \mathbf{g}'(y)). \quad (226)$$

*Proof.* The proof follows using the linear equivalence relation [Theorem 2.2](#). For any two models  $(\mathbf{f}, \mathbf{g}) \sim (\mathbf{f}^*, \mathbf{g}^*)$  we have that:

$$\mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{x}) = \mathbf{g}'_0(y)^\top \mathbf{f}'(\mathbf{x}), \quad (227)$$

for all  $\mathbf{x} \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . By considering  $M$  elements in  $\mathcal{Y}$ , we get

$$\mathbf{L}^\top \mathbf{f}(\mathbf{x}) = \mathbf{L}'^\top \mathbf{f}'(\mathbf{x}), \quad (228)$$

where  $\mathbf{L}, \mathbf{L}' \in \mathbb{R}^{M \times M}$  are the matrices constructed with columns the vectors  $\mathbf{g}_0(y)$  and  $\mathbf{g}'_0(y)$ , respectively. Therefore, we get [Equation \(225\)](#):

$$d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})) = d_{\text{SVD}}(\mathbf{L}^{*\top} \mathbf{f}^*(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})). \quad (229)$$

To obtain [Equation \(226\)](#), notice that with similar steps we can write

$$\mathbf{N}^\top \mathbf{g}(y) = \mathbf{N}'^\top \mathbf{g}'(y) + \mathbf{b}. \quad (230)$$

where  $\mathbf{b} \in \mathbb{R}$  is a displacement vector. We have

$$d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) = d_{\text{SVD}}(\mathbf{N}^{*\top} \mathbf{g}^*(y) + \mathbf{b}, \mathbf{N}'^\top \mathbf{g}'(y)) \quad (231)$$

and given that  $d_{\text{SVD}}$  is invariant to translations, we arrive at the final result

$$d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) = d_{\text{SVD}}(\mathbf{N}^{*\top} \mathbf{g}^*(y), \mathbf{N}'^\top \mathbf{g}'(y)). \quad (232)$$

This shows the claim.  $\square$

## E.7 Invariances of our Representational Distance and CCA

As noted in [Appendix D.2](#),  $m_{\text{CCA}}(\mathbf{f}, \mathbf{f}')$  is invariant to any invertible linear transformation of  $\mathbf{f}$  and  $\mathbf{f}'$ . In contrast, when considering our dissimilarity measure,  $d_{\mathbf{f}, \mathbf{g}}$ , and the transformations to which it is invariant, it is important to note that it relies on both  $d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x}))$  and  $d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y))$ . In both these expressions, embeddings and unembeddings are coupled, since the matrices  $\mathbf{L}, \mathbf{L}'$  (resp.  $\mathbf{N}, \mathbf{N}'$ ) depend on the unembeddings  $\mathbf{g}, \mathbf{g}'$  (resp. the embeddings  $\mathbf{f}, \mathbf{f}'$ ). By contrast,  $m_{\text{CCA}}(\mathbf{f}, \mathbf{f}')$  can be computed independently of the unembeddings  $\mathbf{g}, \mathbf{g}'$ .

Consequently, the two similarity measures (dissimilarity in the case of  $d_{\mathbf{f}, \mathbf{g}}$ ) differ in their invariance properties, as demonstrated below. Consider two models  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$  such that  $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{f}'(\mathbf{x})$  and  $\mathbf{g}_0(y) = \mathbf{B}\mathbf{g}'_0(y)$ , with  $\mathbf{A}$  and  $\mathbf{B}$  invertible matrices but such that  $\mathbf{B} \neq \mathbf{A}^{-\top}$ , i.e.,  $(\mathbf{f}, \mathbf{g})$  and  $(\mathbf{f}', \mathbf{g}')$  are not in the same identifiability class. Since  $\mathbf{A}$  and  $\mathbf{B}$  are invertible matrices, we have  $m_{\text{CCA}}(\mathbf{f}, \mathbf{f}') = m_{\text{CCA}}(\mathbf{g}, \mathbf{g}') = 1$ . In contrast, we get:

$$d_{\mathbf{f}, \mathbf{g}}((\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}')) = \max \left\{ d_{\text{SVD}}(\mathbf{L}^\top \mathbf{f}(\mathbf{x}), \mathbf{L}'^\top \mathbf{f}'(\mathbf{x})), d_{\text{SVD}}(\mathbf{N}^\top \mathbf{g}(y), \mathbf{N}'^\top \mathbf{g}'(y)) \right\} \geq 0, \quad (233)$$

where the equality holds if and only if there exist orthogonal matrices  $\mathbf{O}, \mathbf{O}'$  and diagonal matrices  $\mathbf{S}, \mathbf{S}', \mathbf{D}, \mathbf{D}' \in \mathbb{R}^{M \times M}$  with entries  $S_{ii} := (\psi_{\mathbf{x}}(y_i; p))^{-1}, D_{ii} := (\psi_{\mathbf{y}}(\mathbf{x}_j; p))^{-1}, S'_{ii} := (\psi_{\mathbf{x}}(y_i; p'))^{-1}, D'_{ii} := (\psi_{\mathbf{y}}(\mathbf{x}_j; p'))^{-1}$ , and displacement vectors  $\mathbf{a}, \mathbf{b}$  such that:

$$\mathbf{L}^\top \mathbf{f}(\mathbf{x}) = \mathbf{S}^{-1} \mathbf{O} \mathbf{S}' \mathbf{L}'^\top \mathbf{f}'(\mathbf{x}) + \mathbf{a}, \quad \mathbf{N}^\top \mathbf{g}(y) = \mathbf{D}^{-1} \mathbf{O}' \mathbf{D}' \mathbf{N}'^\top \mathbf{g}'(y) + \mathbf{b}. \quad (234)$$

Because the set of matrices described above is a subset of all linear invertible transformations, it is then possible to find cases where, in [Equation \(233\)](#), the equality does not hold for a careful choice of  $(\mathbf{f}, \mathbf{g}), (\mathbf{f}', \mathbf{g}') \in \Theta$ .

## F Experimental Details

This section contains details of the implementation of all our experiments. A repository containing the code to reproduce our experiments is available on GitHub.<sup>17</sup>

### F.1 Constructed Models

In the following, we will detail how to construct the models which we will use to illustrate [Theorem 3.1](#) and generate [Table 1](#); and to illustrate the bound in [Theorem 4.7](#) (see [Appendix F.6](#)).

We choose classification models with a representation space equal to  $\mathbb{R}^2$ . To construct the reference model  $(\mathbf{f}, \mathbf{g}) \in \Theta$ , we distribute its unembeddings uniformly on the unit circle, ensuring that the angle between any two adjacent unembeddings is equal. We then sample its embeddings such that each embedding corresponding to a given label lies closer—measured by angular distance—to its associated unembedding than to any other.

We construct another model  $(\mathbf{f}', \mathbf{g}') \in \Theta$  by permuting the unembeddings and their associated embeddings. For both models, we then vary the norm of the unembeddings and measure  $d_{\text{KL}}$  and  $d_{\text{LLV}}^\lambda$  between models and the maximum of  $d_{\text{SVD}}$  between embeddings.

To illustrate how  $d_{\text{LLV}}^\lambda$  and the bound derived in [Theorem 4.7](#) increase with increasing differences in representations, we compare a reference model with a perturbed version constructed by adding a small amount of Gaussian noise to the reference model’s embeddings. By increasing the amount of noise,  $d_{\text{LLV}}^\lambda$  grows, as well as the maximum of  $d_{\text{SVD}}$ .

### F.2 Models Trained on Synthetic Data

**Synthetic data generation.** We consider data with two-dimensional input and with  $c$  classes, where each class consists of a slice of the circle together with the opposite slice. See [Fig. 4](#) for an example with  $c = 6$ . We construct the data by drawing 20,000 samples from a two-dimensional Gaussian ( $\mu = \mathbf{0}, \sigma = 3$ ) and assigning labels based on angle to the first axis.

**Model training.** We trained classification models with  $c \in \{4, 6, 10, 18\}$  classes, using a representation space equal to  $\mathbb{R}^2$ . Each model consists of three fully-connected layers, with layer sizes chosen from  $\{16, 32, 64, 128, 256\}$ . We use LeakyReLU activation functions. We train four types of models: one where the norms of both the embeddings and unembeddings are constrained to be equal to 20, one where the norm of the embeddings are equal to 20 and with unconstrained unembeddings, one where the norm of the embeddings is unconstrained and the norm of the unembeddings is 20, and one with no constraints on the norms. To obtain the results in [Section 5.2](#), we only consider models with unconstrained norms. For each combination of the number of classes, the layer size, and whether the restriction is applied or not, we train with 20 random seeds. All models are trained with a batch size of 128 for 15,000 steps using the ADAM optimizer [\[29\]](#).

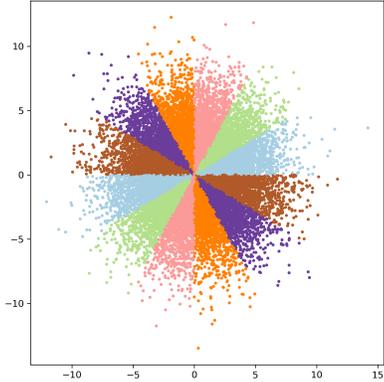


Figure 4: Illustration of training data for 6 classes. Each color represents a different class label.

### F.3 Models Trained on CIFAR-10

We trained classification models on CIFAR-10 [\[32\]](#), where the embedding network consisted of a ResNet18 [\[21\]](#)<sup>18</sup> but choosing the representation space to be 2, 3 or 5-dimensional. For each dimension we trained 10 seeds. The unembedding network consisted of three fully connected layers of width 128, followed by an output layer of size 2, 3 or 5, thus giving us representations in the desired number of dimensions. The models were trained for 20,000 steps with a batch size of 32

<sup>17</sup>[github.com/bemigini/close-dist-rep-sim](https://github.com/bemigini/close-dist-rep-sim)

<sup>18</sup>For this, we used code based on <https://github.com/kuangliu/pytorch-cifar/blob/master/models/resnet.py>

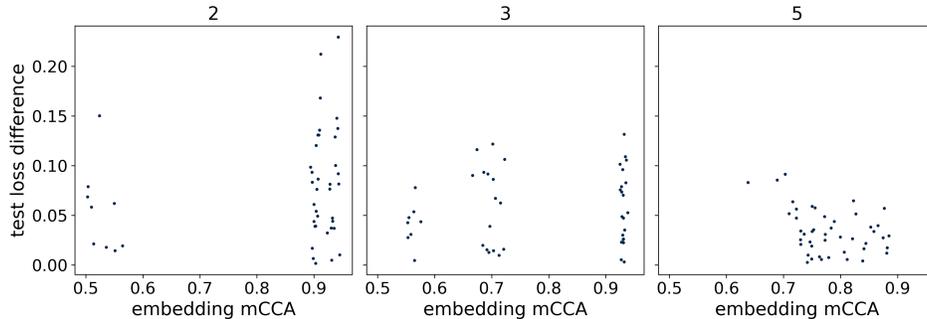


Figure 5: For models trained on CIFAR-10 with representational dimensions of 2, 3 and 5, difference in test loss vs  $m_{CCA}$  of the embeddings of the models. We see that there can both be a small difference in loss and a larger difference in representations or a larger difference in loss and a smaller difference in representations.

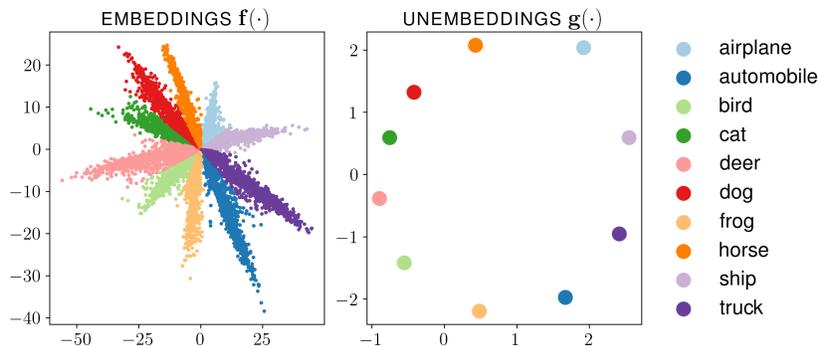


Figure 6: Illustration of representations of model trained on CIFAR-10, seed 0.

using the ADAM optimizer [29]. The ResNet model used ReLU activation functions, while the networks for the unembeddings used LeakyReLU activation functions.

#### F.4 Loss Difference vs Embedding $m_{CCA}$

To illustrate [Corollary 3.2](#) and that a small difference in test loss does not guarantee similar representations for our models trained on CIFAR-10, we present [Fig. 5](#). Here we compare difference in test loss with  $m_{CCA}$  of embeddings of models, where the dimension of the representations are 2, 3 and 5. We see that there can both be a small difference in loss and a larger difference in representations or a larger difference in loss and a smaller difference in representations.

#### F.5 All Two-dimensional Representations of CIFAR-10 Models

We here present all the embedding and unembedding representations of our models trained on CIFAR-10 with 2-dimensional representations: seed 0 in [Fig. 6](#), seed 1 in [Fig. 7](#), seed 2 in [Fig. 8](#), seed 3 in [Fig. 9](#), seed 4 in [Fig. 10](#), seed 5 in [Fig. 11](#), seed 6 in [Fig. 12](#), seed 7 in [Fig. 13](#), seed 8 in [Fig. 14](#), seed 9 in [Fig. 15](#).

We see that some labels are neighbours for all ten models. For example, “automobile” and “truck” are neighbours for all seeds and “cat” and “dog” are neighbours for all seeds. However, other labels sometimes have varying neighbours. For example “airplane” and “bird” are neighbours for seeds 2, 3, 5, 6, 7, but not for the remaining seeds. In seeds 1, 4, 8 and 9, the “frog” label is put between them, and in seed 0 the labels are permuted even further. This behaviour might arise because inputs for some labels are so similar that it would adversely affect the performance of the model to place them far apart, while the embeddings of inputs for other labels can be placed in several equally good ways.

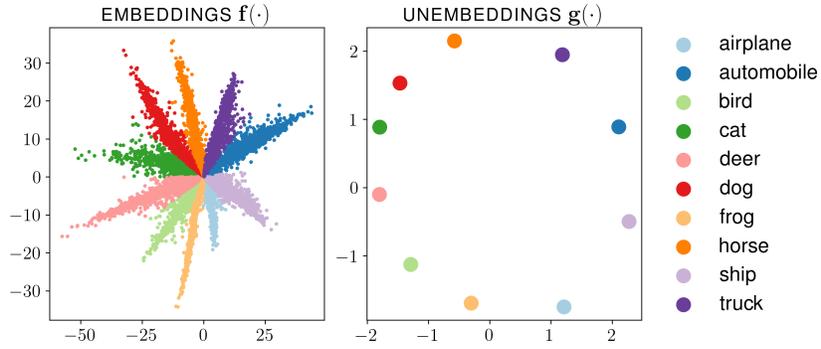


Figure 7: Illustration of representations of model trained on CIFAR-10, seed 1.

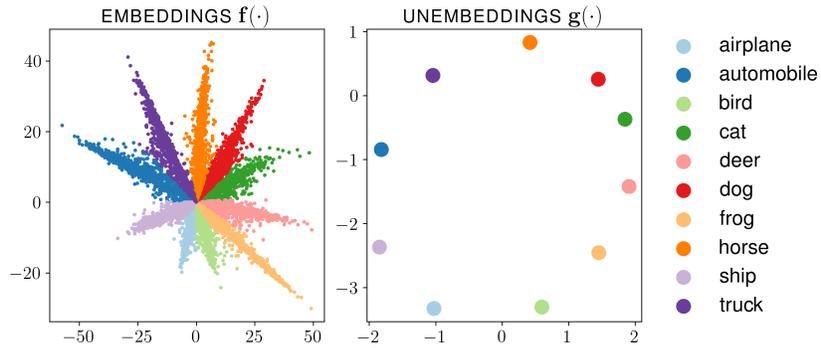


Figure 8: Illustration of representations of model trained on CIFAR-10, seed 2.

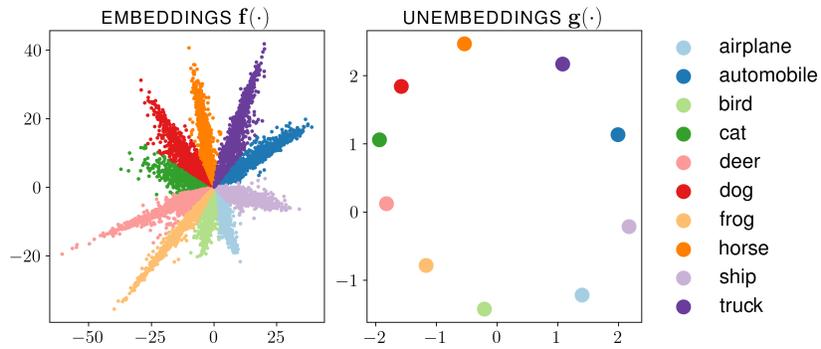


Figure 9: Illustration of representations of model trained on CIFAR-10, seed 3.

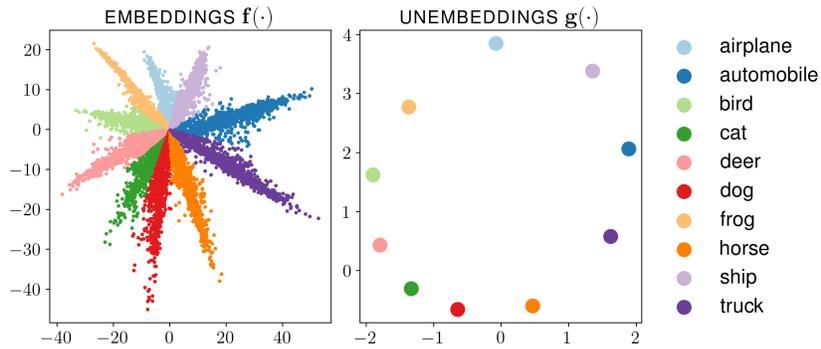


Figure 10: Illustration of representations of model trained on CIFAR-10, seed 4.

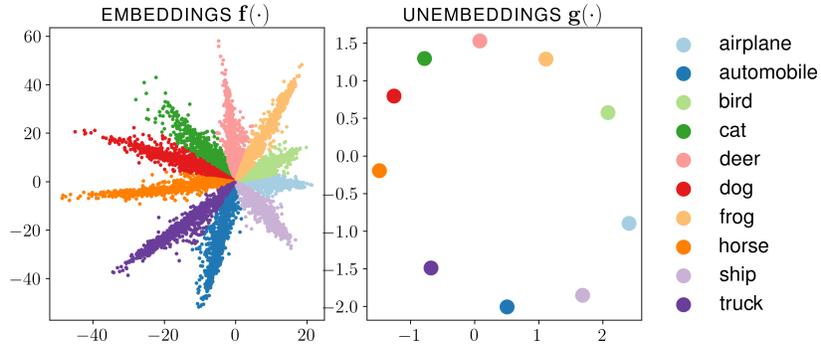


Figure 11: Illustration of representations of model trained on CIFAR-10, seed 5.

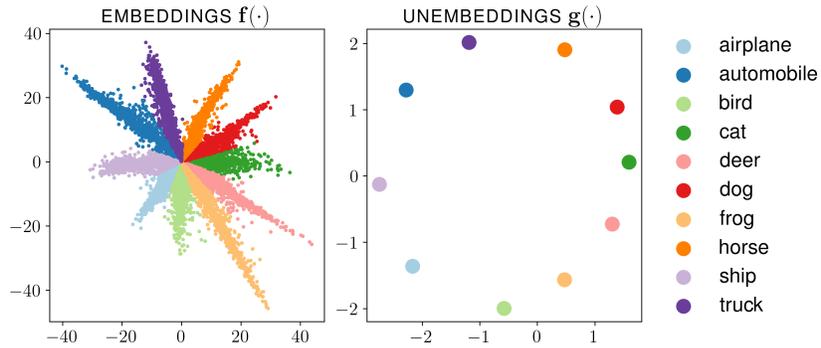


Figure 12: Illustration of representations of model trained on CIFAR-10, seed 6.

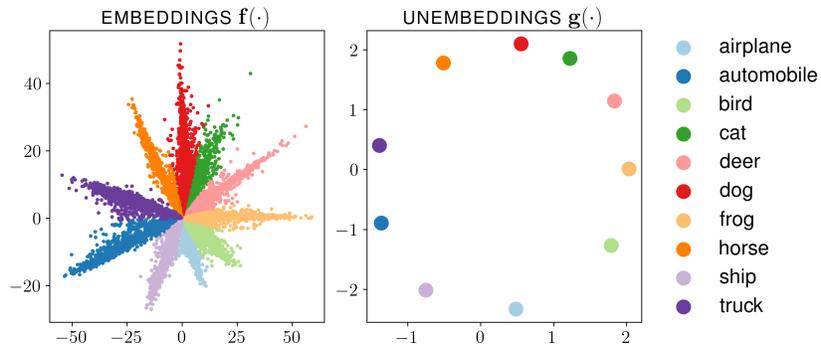


Figure 13: Illustration of representations of model trained on CIFAR-10, seed 7.

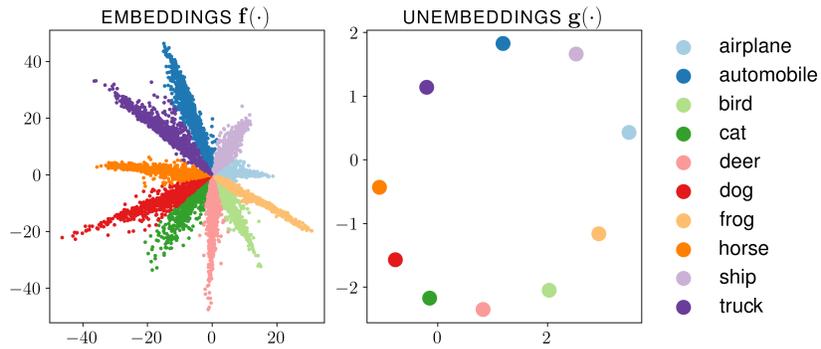


Figure 14: Illustration of representations of model trained on CIFAR-10, seed 8.

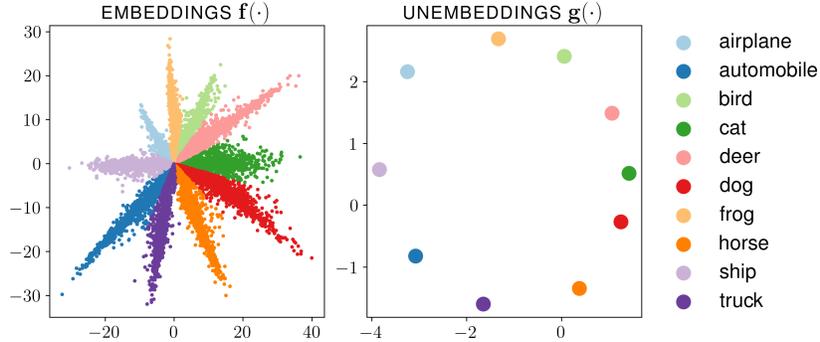


Figure 15: Illustration of representations of model trained on CIFAR-10, seed 9.

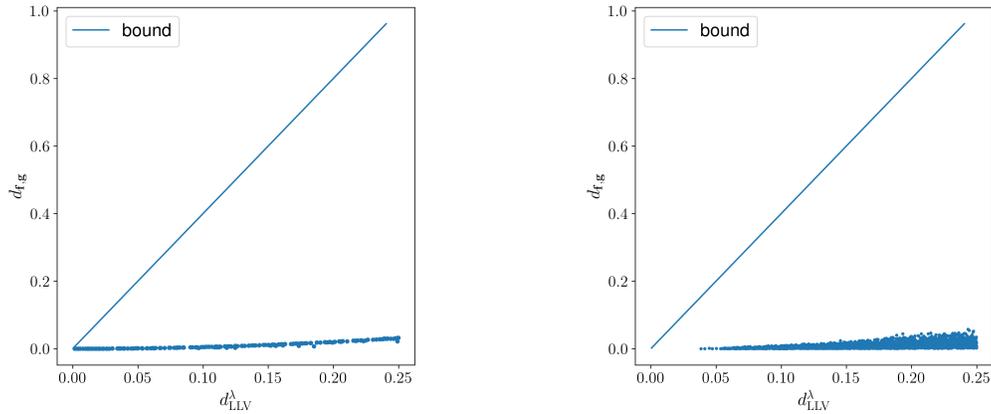


Figure 16: (Left) We evaluated  $d_{LLV}^\lambda$  and  $d_{f,g}$  for a reference model and a constructed one where we progressively increase the noise added to the embeddings. As more noise is injected, the two models display higher  $d_{LLV}^\lambda$  (up to 0.25), while the representation dissimilarity  $d_{f,g}$  does not exceed 0.1. (Right) We observe a similar trend for trained models, with a slight increase in  $d_{f,g}$  when  $d_{LLV}^\lambda$  approaches 0.25. We note that the bound given by [Theorem 4.7](#) remains valid for both cases.

## F.6 Illustration of Bound on Constructed and Trained Models

**Experimental setup.** We compare both constructed models and models trained on synthetic data. For the constructed models, we compare a reference model to a perturbed version constructed by adding Gaussian noise to the reference model’s embedding representations (as described in [Appendix F.1](#)). For models trained on synthetic data, the setup is the same as described in [Appendix F.2](#)

**Results.** In [Fig. 16](#) (left), we observe that the maximum distance between representations gradually increases with the distance between probability distributions while staying below the bound of the [Theorem 4.7](#). We find that not all the trained models have distance  $d_{LLV}^\lambda$  small enough for the bound to be non-vacuous. In this case, we report only those distances that are small enough for the bound to be non-vacuous between trained models in [Fig. 16](#) (right) and see that the bound remains valid.

## F.7 Wider Models Have more Similar Distributions - Extra Plots

We here report the empirical trend that wider networks induce more similar distributions and more similar representations for models with 10 classes ([Fig. 17](#)  $d_{LLV}$ (left) and  $\max d_{SVD}$ (right)) and 18 classes ([Fig. 18](#)  $d_{LLV}$ (left) and  $\max d_{SVD}$ (right)). Note that with more classes, we have less models among the most narrow networks which achieve more than 90% accuracy. We have therefore left out results in the plots where we had fewer than 5 models for the comparison. For 10 classes, this means that starting at a width of 32, we have 9, 16, 19, and 19 models for the comparisons in the plot. For 18 classes, starting at a width of 64, we have 10, 12, and 19 models for the comparisons in the plot.

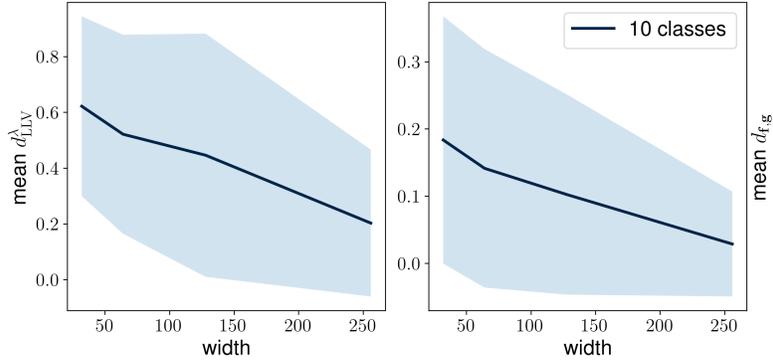


Figure 17: We display how mean  $d_{LLV}$  and max  $d_{LLV}$  varies with increasing the neural network width for models trained to classify among 10 classes. (Left) We observe decreasing mean  $d_{LLV}$  as the network width grows. The shaded area represents the standard deviations evaluated from different random seeds retraining. (Right) A similar trend is also observed for max  $d_{LLV}$  when increasing the network width.

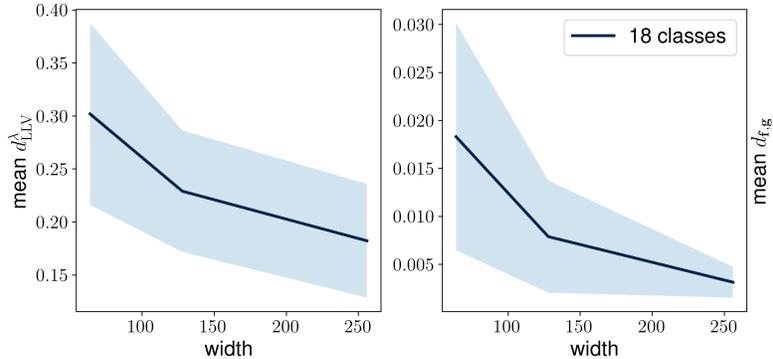


Figure 18: We display how mean  $d_{LLV}$  and max  $d_{LLV}$  varies with increasing the neural network width for models trained to classify among 18 classes. (Left) We observe decreasing mean  $d_{LLV}$  as the network width grows. The shaded area represents the standard deviations evaluated from different random seeds retraining. (Right) A similar trend is also observed for max  $d_{LLV}$  when increasing the network width.

## G Computing Resources

Each model was trained using a single NVIDIA RTX A5000. For each number of classes (4, 6, 10, 18), training 20 seeds on the synthetic data took about 34 hours, summing to a total of 136 hours to train the models on synthetic data. For the models on CIFAR-10, training 10 seeds took  $\sim 27$  hours.

The distances are calculated on a CPU-only machine: computing the distances for the models on synthetic data required less than 2 hours, whereas the evaluation on models in CIFAR-10 required around 4 hours. Evaluating the accuracy of models on synthetic data was also done on the CPU, taking less than 20 minutes in total.

## H Assets

**CIFAR-10.** We used the CIFAR-10 dataset as loaded with torchvision package<sup>19</sup>. The dataset from [32] and contains 50,000 images for train and validation, and 10,000 images for testing.

**Python Packages.** All experiments are conducted with Python 3.11 and used pytorch 2.5.1. Other packages are reported in the repository at [github.com/bemigini/close-dist-rep-sim](https://github.com/bemigini/close-dist-rep-sim).

<sup>19</sup><https://docs.pytorch.org/vision/main/generated/torchvision.datasets.CIFAR10.html>