# Improved Generalization Bounds for Transfer Learning via Neural Collapse

**Tomer Galanti** [1] **András György** [2] **Marcus Hutter** [2]

## Abstract

Using representations learned by large, pretrained models, also called foundation models, in new tasks with fewer data has been successful in a wide range of machine learning problems. Recently, Galanti et al. (2022) introduced a theoretical framework for studying this transfer learning setting for classification. Their analysis is based on the recently observed phenomenon that the features learned by overparameterized deep classification networks show an interesting clustering property, called neural collapse (Papyan et al., 2020). A cornerstone of their analysis demonstrates that neural collapse generalizes from the source classes to new target classes. However, this analysis is limited as it relies on several unrealistic assumptions. In this work, we provide an improved theoretical analysis significantly relaxing these modeling assumptions.

## 1. Introduction

Transfer learning is a prominent approach for dealing with overfitting (see, e.g., Caruana, 1995; Bengio, 2012; Yosinski et al., 2014). A popular approach for transfer learning suggests to pretrain a large neural network on a large-scale source task (e.g., ResNet-50 on ImageNet ILSVRC, Russakovsky et al., 2015), and then to train a relatively small network on top of the penultimate layer of the pretrained model, using the available data in the target task. In fact, due to the impressive success of transfer learning, large pretrained models that can be effectively adapted to a wide variety of tasks (Brown et al., 2020; Ramesh et al., 2021) have recently been characterized as 'foundation models' (Bommasani et al., 2021), emphasizing their adaptive nature.

While foundation models are intended to be generic and widely adaptive to downstream tasks, when some specifics of the target tasks are known, often special-purpose algorithms are designed to utilize this information. This is the case, for example, in few-shot learning, where it is known in advance that the target problems come with a very small training set (Vinyals et al., 2016; Ravi & Larochelle, 2017; Finn et al., 2017; Lee et al., 2019). Even though these approaches significantly improved the state of the art for a long time, surprisingly, recent works have demonstrated that predictors trained on top of pretrained models can achieve better performance on few-shot learning benchmarks (Tian et al., 2020; Dhillon et al., 2020; Galanti et al., 2022).

Complementing the empirical results, Galanti et al. (2022) also studied this approach theoretically,and provided an explanation for its success based on the recently discovered phenomenon of neural collapse (Papyan et al., 2020). Informally, neural collapse (see Section 3) identifies training dynamics of deep networks for standard classification tasks, where the features (the output of the penultimate layer) associated with training samples belonging to the same class concentrate around their class feature mean. Galanti et al. (2022) showed that this property generalizes to new data points and new classes (e.g., the target classes), when the model is trained on several classes with many samples for each. In addition, they showed that in the presence of neural collapse, training a linear classifier on top of the learned feature map can generalize well, even if trained with few samples only. However, their analysis demonstrating that neural collapse generalizes to new classes relies on some hard-to-justify assumptions regarding the feature maps, making their results less relevant in practical situations.

In this paper we provide a stronger theoretical analysis for this problem without relying on this kind of assumptions.

## 2. Problem Setup

We consider a transfer learning setting, where a model is pretrained on some source task and is adapted to solve many downstream tasks. To model this problem, we assume that a final downstream task is a $k$-class classification problem (a *target* problem), and the auxiliary task where the feature representation is learned on is an $l$-class classification problem, called the *source* problem. Formally, a target task is defined by a distribution $P$ over samples $(x, y) \in \mathcal{X} \times \mathcal{Y}_k$, where $\mathcal{X} \subset \mathbb{R}^d$ is the instance space, and $\mathcal{Y}_k = [k] := \{1, \ldots, k\}$

[1]MIT, Cambridge, MA, USA [2]Deepmind, London, UK. Correspondence to: Tomer Galanti <galanti@mit.edu>, András György <agyorgy@deepmind.com>.

*First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

is a label space. For a pair $(x, y)$ with distribution $P$, we denote by $P_c$ the class-conditional distribution of $x$ given $y = c$ (i.e., $P_c(\cdot) := \mathbb{P}[x \in \cdot \mid y = c]$).

For a given target task, our goal is to learn a classifier $h : \mathcal{X} \to \mathcal{Y}_k$ that minimizes the target test error

$$L_P(h) \;:=\; \mathbb{E}_{(x,y)\sim P}\big[\mathbb{I}[h(x) \neq y]\big], \tag{1}$$

where $\mathbb{I}[E]$ denotes the indicator function of an event $E$ (equals 1 if the event holds and 0 otherwise). As usual, $P$ is unknown and the algorithm is provided with some training data $S = \{(x_i, y_i)\}_{i=1}^{nk}$. In this work we assume that $P$ is a balanced distribution (i.e., $P[y = c] = 1/k$) with class-conditionals $\{P_c\}_{c=1}^k$ and that the data $S$ is also a balanced set over the $k$ classes, that is, $S = \cup_{c=1}^k \{(x_{ci}, c)\}_{i=1}^n$, where the sets $S_c = \{x_{ci}\}_{i=1}^n \sim P_c^n$ are drawn independently.

When $n$ is small, training a classifier on $S$ may not be sufficient to achieve reasonable performance. To facilitate finding a good solution, we aim to find a classifier of the form $h = g \circ f$, where $f \in \mathcal{F} \subset \{f' : \mathbb{R}^d \to \mathbb{R}^p\}$ is a feature map and $g \in \mathcal{G} \subset \{g' : \mathbb{R}^p \to \mathbb{R}^k\}$ is a classifier used on the feature space $\mathbb{R}^p$. The idea is that the feature map $f$ can be learned on some other problem where more data is available, and then $g$ is trained to solve the hopefully simpler classification problem of finding $y$ based on $f(x)$, instead of $x$. That is, $g$ is actually a function of the modified training data $\{(f(x_i), y_i)\}_{i=1}^{nk}$; to emphasize this dependence, we use the notation $g_{f,S}$ for the trained classifier based on the features, and $h_{f,S} = g_{f,S} \circ f$ for the full classifier (to be specified in Section 4).

We assume that the source task helping to find $f$ is a single $l$-class classification problem over the same sample space $\mathcal{X}$, given by a distribution $\tilde{P}$, and here we are interested in finding a classifier $\tilde{h} : \mathcal{X} \to \mathbb{R}^l$ of the form $\tilde{h} = \tilde{g} \circ f$, where $\tilde{g} \in \tilde{\mathcal{G}} \subset \{g' : \mathbb{R}^p \to \mathbb{R}^l\}$ is a classifier over the feature space $f(\mathcal{X}) := \{f(x) : x \in \mathcal{X}\}$. Given a training dataset $\tilde{S} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{ml}$, the classifier $\tilde{h}$ is trained on $\tilde{S}$, with the goal of minimizing the source test error, $L_{\tilde{P}}(\tilde{h}_{\tilde{S}})$.

Similarly to the target task, we assume that $\tilde{P}$ is balanced (i.e., $\tilde{P}[y = c] = 1/l$) with class-conditionals $\{\tilde{P}_c\}_{c=1}^l$, and that the dataset $\tilde{S} = \cup_{c=1}^l \{(\tilde{x}_{ci}, c)\}_{i=1}^m$ is also balanced with $\tilde{S}_c = \{\tilde{x}_{ci}\}_{i=1}^m \sim \tilde{P}_c^m$. Finally, we assume that the classes in the source and target tasks are selected randomly: that is, the sets of class-conditionals $\tilde{\mathcal{P}} = \{\tilde{P}_c\}_{c=1}^l$ and $\mathcal{P} = \{P_c\}_{c=1}^k$ for the source and target tasks are sampled i.i.d. from a distribution $\mathcal{D}$ over a set of class-conditional distributions $\mathcal{E}$. This completes the definition of the distribution of $P$ and $\tilde{P}$ (which are random themselves).

In a typical setting, the classifier $\tilde{h}$ is a deep neural network, $f$ is the representation in the last internal layer of the network (i.e., the penultimate, a.k.a. *embedding* layer), and $\tilde{g}$, the last layer of the network, is a linear map; similarly, $g$ in the target problem is often taken to be linear. The learned feature map $f$ is called a *foundation model* (Bommasani et al., 2021) when it can be effectively used in a wide range of target tasks. In particular, its effectiveness in dealing with downstream tasks (for any $f$) can be measured by its expected *transfer error*,

$$\mathcal{L}_{\mathcal{D}}(f) \;:=\; \mathbb{E}_P \mathbb{E}_S[L_P(h_{f,S})], \tag{2}$$

where the expectation is taken over the random choice of $P$ (i.e., averaging over randomly selected target tasks) and the training data $S$ for the target class.

Notice that while the feature map $f$ is evaluated on the distribution of target tasks $P$ determined by $\mathcal{D}$, the training of $f$ in a foundation model, as described above, is fully agnostic of this target. In the rest of the paper we analyze this setting, and provide an explanation through the recent concept of neural collapse.

**Notation.** $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. For a distribution $Q$ over $\mathcal{X} \subset \mathbb{R}^d$ and $u : \mathcal{X} \to \mathbb{R}^p$, the mean and variance of $u(x)$ for $x \sim Q$ are denoted by $\mu_u(Q) := \mathbb{E}_{x\sim Q}[u(x)]$ and by $\mathrm{Var}_u(Q) := \mathbb{E}_{x\sim Q}[\|u(x) - \mu_u(Q)\|^2]$. For a finite set $A = \{a_i\}_{i=1}^n$, we denote $\mathrm{Avg}_{i=1}^n[a_i] := \frac{1}{n}\sum_{i=1}^n a_i$ and by $U[A]$ the uniform distribution over $A$. For two functions $f, g : X \to [0, \infty)$, we denote $f \overset{\times}{\leq} g$ if there exists a constant $C > 0$, such that, $\forall x \in X :\; f(x) \leq C \cdot g(x)$.

## 3. Neural Collapse

Neural collapse (NC) is a recently discovered phenomenon in deep learning (Papyan et al., 2020): it has been observed that during the training of deep networks for standard classification tasks, the features (the output of the penultimate, a.k.a. embedding layer) associated with training samples belonging to the same class concentrate around the mean feature value for the same class, also satisfying some additional conditions (such as the mean feature vectors being orthogonal to each other).

From these we focus on the class-features variance collapse (called NC1 by Papyan et al., 2020), and used a simplified version inroduced by Galanti et al. (2022): For a feature map $f : \mathbb{R}^d \to \mathbb{R}^p$ and two distributions $Q_1, Q_2$ over $\mathcal{X} \subset \mathbb{R}^d$, we define their *class-distance normalized variance* (CDNV) as

$$V_f(Q_1, Q_2) \;:=\; \frac{\mathrm{Var}_f(Q_1)}{\|\mu_f(Q_1) - \mu_f(Q_2)\|^2}$$

(the definition can be extended to finite sets $S_1, S_2 \subset \mathcal{X}$ by defining $V_f(S_1, S_2) = V_f(U[S_1], U[S_2])$). This quantity essentially measures to what extent the feature vectors of samples from $Q_1$ and $Q_2$ are clustered in space. We say that a training process satisfies *neural collapse* if

$$\lim_{t\to\infty} \mathrm{Avg}_{i\neq j\in[l]}[V_{f^{(t)}}(\tilde{S}_i, \tilde{S}_j)] = 0, \tag{3}$$

where $f^{(t)}$ is the penultimate layer of a neural network $h^{(t)} = g^{(t)} \circ f^{(t)}$ that is trained to fit $\tilde{S}$ for $t$ steps. It has been observed (Papyan et al., 2020; Galanti et al., 2022) that for several problems the left-hand side of (3) indeed reduces to some small value during training.

## 4. Generalization Bounds for Transfer Learning

Galanti et al. (2022) introduced generalization bounds for transfer learning in the setting of Section 2. Their analysis takes three main steps; first, they show that if the CDNV on the training data, $\mathrm{Avg}_{i\neq j\in[l]}[V_f(\tilde{S}_i,\tilde{S}_j)]$, is small at the end of training, then we expect that the CDNV over unseen samples, $\mathrm{Avg}_{i\neq j\in[l]}[V_f(\tilde{P}_i,\tilde{P}_j)]$, would also be small for the same classes (see Lemma 4.1 below). As a second step, they bound the expected CDNV between two new target classes, $\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]$, using the CDNV between the source classes, $\mathrm{Avg}_{i\neq j\in[l]}[V_f(\tilde{P}_i,\tilde{P}_j)]$. Finally, they prove that $\mathcal{L}_{\mathcal{D}}(f) \overset{\times}{\leq} (k+k/n)\cdot\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]$, with $h_{f,S} := \arg\min_{c\in[C]} \|f(x_j)-\mu_f(S_c)\|$ being the *nearest class-center (NCC)* classifier.

However, there are multiple limitations to their arguments. First, the term $\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]$ may be very large even when $\mathcal{L}_{\mathcal{D}}(f)$ is very small. For example, if there is at least one anomalous pair of class-conditionals $P_1,P_2\in\mathcal{E}$ for which $\mu_f(P_1)=\mu_f(P_2)$, then, $\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]=\infty$ (as long as $\mathcal{D}[P_1],\mathcal{D}[P_2]>0$) even if for all other pairs $P_1',P_2'\in\mathcal{E}$, $V_f(P_1',P_2')$ is tiny. In fact, if the support $\mathcal{E}$ of $\mathcal{D}$ is infinite and $f$ is bounded, then clearly $\inf_{P_1\neq P_2\in\mathcal{E}}\|\mu_f(P_1)-\mu_f(P_2)\|=0$ and $\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]=\infty$. Second, the bound of Galanti et al. (2022) on $\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]$ scales with $(\inf_{f\in\mathcal{F}}\inf_{P_1\neq P_2\in\mathcal{E}}\|\mu_f(P_1)-\mu_f(P_2)\|)^{-1}$, which may be very large even if $\mathcal{E}$ is finite (e.g., if $\mathcal{F}$ contains a constant function).

In this paper we aim to circumvent these issue and provide an upper bound on the error of $h$ which does not suffer from these limitations. Instead of bounding the term $\mathbb{E}_{P_c\neq P_{c'}}[V_f(P_c,P_{c'})]$, we bound $\mathcal{L}_{\mathcal{D}}(f)$ in terms of the averaged margin error of the NCC classifier between pairs of source classes. As a next step, in Lemma 4.3, we show that each of these error terms can be bounded using the CDNV between pairs of source classes. Finally, in Theorem 4.4 we combine Lemmas 4.1-4.3 and obtain an upper bound on the transfer error of $f$ in terms of $\mathrm{Avg}_{i\neq j\in[l]}[V_f(\tilde{S}_i,\tilde{S}_j)]$ which is typically small (see Section 3).

### 4.1. Neural Collapse Generalizes to New Samples

We start by recalling Proposition 1 of Galanti et al. (2022). This proposition provides an upper bound on $V_f(\tilde{P}_i,\tilde{P}_j)$, decomposed into $V_f(\tilde{S}_i,\tilde{S}_j)$ and additional generalization gap terms bounding the difference between expectations and empirical averages for $f$ and its variants, where $f$ is the output of the learning algorithm with access to $\tilde{S}$.

For any $\delta\in(0,1)$, we let $\epsilon_1^c(\delta)$ and $\epsilon_2^c(\delta)$ be the smallest positive values such that with probability at least $1-\delta$, the learning algorithm returns a function $f\in\mathcal{F}$ that satisfies

$$\begin{aligned}\left\|\mathbb{E}_{x\sim\tilde{P}_c}[f(x)]-\mathrm{Avg}_{x\in\tilde{S}_c}[f(x)]\right\| &\leq \epsilon_1^c(\delta);\\ \left|\mathbb{E}_{x\sim\tilde{P}_c}[\|f(x)\|^2]-\mathrm{Avg}_{x\in\tilde{S}_c}[\|f(x)\|^2]\right| &\leq \epsilon_2^c(\delta).\end{aligned}$$

These quantities are typically bounded using Rademacher complexities (Bartlett & Mendelson, 2002) related to $\mathcal{F}$, scaling usually as $\mathcal{O}_m(1/\sqrt{m})$ (as also shown by Galanti et al., 2022). Next, we present their bound on the CDNV of the source distributions.

**Lemma 4.1** (Galanti et al., 2022, Proposition 1)**.** *Fix two source classes, $i$ and $j$ with distributions $\tilde{P}_i$ and $\tilde{P}_j$, and let $\delta\in(0,1)$. Let $\tilde{S}_c\sim\tilde{P}_c^m$ for $c\in\{i,j\}$. Let*

$$A_{ij}(\delta) := \frac{\epsilon_1^i(\delta/4)+\epsilon_1^j(\delta)}{\|\mu_f(\tilde{P}_i)-\mu_f(\tilde{P}_j)\|}$$

$$B_{ij}(\delta) := \frac{\mathrm{Avg}_{c\in\{i,j\}}\left[\epsilon_2^c(\delta)+2\|\mu_f(\tilde{P}_c)\|\cdot\epsilon_1^c(\delta)+\epsilon_1^c(\delta)^2\right]}{\|\mu_f(\tilde{S}_i)-\mu_f(\tilde{S}_j)\|^2}.$$

*Then, with probability at least $1-\delta$ over the selection of $\tilde{S}_i,\tilde{S}_j$, we have $V_f(\tilde{P}_i,\tilde{P}_j)\leq(V_f(\tilde{S}_i,\tilde{S}_j)+B_{ij}(\delta/4))\cdot(1+A_{ij}(\delta/4))^2$.*

The bound in the lemma has several terms. The first one is the empirical CDNV, $V_f(\tilde{S}_i,\tilde{S}_j)$, which is assumed to be implicitly minimized by the training algorithm (see Section 3). The rest of the terms are proportional to the generalization gaps $\epsilon_1^c(\delta/4)$ and $\epsilon_2^c(\delta/4)$ — as discussed above, typically we expect these terms to scale as $\mathcal{O}_m(1/\sqrt{m})$. As discussed in (Galanti et al., 2022), the denominators in $A_{ij}(\delta/4)$ and $B_{ij}(\delta/4)$ are expected to be lower bounded by positive constants under reasonable conditions.

Therefore, for a pair $i\neq j$, if the CDNV $V_f(\tilde{S}_i,\tilde{S}_j)$ is small and $m$ is large, then we expect $V_f(\tilde{P}_i,\tilde{P}_j)$ to be small, as well. That is, if neural collapse emerges in the training data of two source classes, we should also expect it to emerge in unseen samples of the same classes (for large $m$). For an extensive empirical validation of this argument, see the works of Galanti et al. (2022); Galanti & Galanti (2022).

### 4.2. Bounding the Transfer Error

In this section we bound the transfer error of $f$ using the averaged margin error of the NCC classifier between pairs of source classes. In the analysis, similarly to Galanti et al. (2022), we treat class-conditional distributions as data points on which the feature map $f$ is trained. We apply standard techniques to derive generalization bounds to new data

points, which, in this case, are class-conditional distributions. Recall that, in line with this view, we have already assumed that the source and target class-conditionals $\{\tilde{P}_c\}_{c=1}^l$ and $\{P_c\}_{c=1}^k$ are i.i.d. samples from $\mathcal{D}$.

For simplicity, we focus on the case where $\mathcal{F}$, the class of feature maps is a set of ReLU neural networks of the form $f(\cdot) = W^q\sigma(W^{q-1}\dots\sigma(W^1\cdot)) : \mathbb{R}^d \to \mathbb{R}^p$, where $\sigma(x) := \max\{0, x\}$ is the element-wise ReLU function, and $W^i \in \mathbb{R}^{d_{i+1}\times d_i}$ for $i \in [q]$, where $d_1 = d$ and $d_{q+1} = p$. Our bounds depend on the spectral complexity of a network $f$, which is defined as $\mathcal{C}(f) := \max_{j\in[p]} \|W_j^q\| \cdot \prod_{r=1}^{q-1} \|W^r\|$ and $\hat{\mathcal{C}}(f) := \mathcal{C}(f) + 1$. This quantity upper bounds the Lipschitz constant of $f$ and is similar in fashion to other (slightly different) notions of spectral complexity for neural networks (Golowich et al., 2017; Bartlett et al., 2017). We also denote $B := \sup_{x\in\mathcal{X}} \|x\|$.

We start by a generic result that allows us to upper bound the transfer error of the NCC classifier (on top of the feature map $f$) by the margin-error over the source classes (defined below), plus some additional terms, depending on the complexity of $f$ and the number of classes. The proof, presented in Appendix A, is partially based on Corollary 3 of Maurer & Pontil (2019) and the analysis of Bartlett et al. (2017).

**Lemma 4.2.** *Let $\delta \in (0, 1)$, $l \geq 2$ and let $\mathcal{F}$ be a class of ReLU neural networks of depth $q$. Then, with probability at least $1 - \delta$ over the random selection of the class-conditionals $\tilde{\mathcal{P}} = \{\tilde{P}_c\}_{c=1}^l$ of the source classes, for any $f \in \mathcal{F}$ and $\Delta > 0$, we have*

$$\mathcal{L}_{\mathcal{D}}(f) \leq (k-1)\cdot\left[\mathrm{Avg}_{i\neq j\in[l]}[L_{ij}^{2\Delta}(f)] + \delta\right] + \frac{\sqrt{q}kpB\cdot\hat{\mathcal{C}}(f)}{\Delta\sqrt{l}}\cdot\mathrm{polylog}(k, \delta^{-1}, \hat{\mathcal{C}}(f), |\log(\tfrac{\Delta}{B})|, l)$$

*where $L_{ij}^{\Delta}(f) := \Pr_{\hat{x}_i,\hat{S}_i,\hat{S}_j}[\|f(\hat{x}_i) - \mu_f(\hat{S}_j)\| \leq \|f(\hat{x}_i) - \mu_f(\hat{S}_i)\| + \Delta]$, where $\hat{S}_i \sim \tilde{P}_i^n, \hat{S}_j \sim \tilde{P}_j^n$ and $\hat{x}_i \sim \tilde{P}_i$.*

The bound above can be decomposed into several parts. The first term is the average expected $\Delta$-margin error of a NCC classifier, when averaged over all classification problems given by the pairs of source classes with $n$ samples each. For each pair $(i, j)$, the NCC classifier is defined with two datasets of size $n$, $\hat{S}_i$ and $\hat{S}_j$ (which are independent of the training datasets $\tilde{S}_i$ and $\tilde{S}_j$), and its performance is evaluated over a new test sample $\hat{x}_i \sim \tilde{P}_i$. The second term, $\delta$, is a free parameter and can be selected to be very small as the third term scales polylogarithmically with respect to $1/\delta$. The third term is proportional to the ratio between the complexity of the selected function $f$, captured by $\sqrt{q}pkB\cdot\hat{\mathcal{C}}(f)$ and $\Delta\sqrt{l}$. Therefore, as long as we increase the number of source classes $l$, we can expect the generalization to new classes to improve.

The next lemma is an extension of Proposition 5 of Galanti et al. (2022) and relates the margin-error to the CDNV.

**Lemma 4.3.** *Let $\{\tilde{P}_i, \tilde{P}_j\}$ be a pair of class-conditional distributions. Let $f : \mathbb{R}^d \to \mathbb{R}^p$ be any feature map and denote $\mu_c = \mu_f(\tilde{P}_c)$. Let $\Delta \leq 0.12s(f, \tilde{\mathcal{P}})^{-1/2}\cdot\|\mu_f(\tilde{P}_i) - \mu_f(\tilde{P}_j)\|$. Then, we have, for some universal constant $C > 0$,*

$$\tfrac{1}{2}(L_{ij}^{\Delta}(f) + L_{ji}^{\Delta}(f)) \leq C\left[\tfrac{1}{s(f,\tilde{\mathcal{P}})} + \tfrac{1}{n}\right]\cdot V_f(\tilde{P}_i, \tilde{P}_j),$$

*where $s(f, \tilde{\mathcal{P}}) = p$ if $\{f\circ\tilde{P}_c\}_{c=1}^l$ are spherically symmetric and $s(f, \tilde{\mathcal{P}}) = 1$ otherwise.*

Informally, the expected margin error $L_{ij}^{\Delta}(f)$ of $f$ is upper bounded by $V_f(\tilde{P}_i, \tilde{P}_j)$, and when $f\circ\tilde{P}_i$ and $f\circ\tilde{P}_j$ are spherically symmetric then by $(1/p + 1/n)\cdot V_f(\tilde{P}_i, \tilde{P}_j)$. Therefore, in case of neural collapse (i.e., when the $V_f(\tilde{P}_i, \tilde{P}_j)$ are small on average), $L_{ij}^{\Delta}(f)$ is small, explaining the success of foundation models in the low-data regime (such as in few-shot learning problems) in the presence of neural collapse. Finally, we also derive a much better bound (see Lemma B.2) which is exponentially small in $p/V_f(\tilde{P}_1, \tilde{P}_2)$ if $f\circ\tilde{P}_c$ are (assumed to be) centered Gaussian distributions as also assumed by Papyan et al. (2020).

Putting together Lemmas 4.1-4.3, we obtain the following generalization bound.

**Theorem 4.4.** *Let $\delta \in (0, 1)$, $l \geq 2$ and let $\mathcal{F}$ be a class of ReLU neural networks of depth $q$. Then, with probability at least $1-\delta$ over the selection of the source class-conditionals $\tilde{\mathcal{P}} = \{\tilde{P}_c\}_{c=1}^l$ and the corresponding source training data $\tilde{S}_1\dots,\tilde{S}_l$, for any $f \in \mathcal{F}$ and $\Delta = 0.12s(f, \tilde{\mathcal{P}})^{-1/2}\cdot\min_{i\neq j\in[l]}\|\mu_f(\tilde{P}_i) - \mu_f(\tilde{P}_j)\|$, we have*

$$\mathcal{L}_{\mathcal{D}}(f) \leq \tfrac{k\delta}{2} + Ck\cdot\left[\tfrac{1}{s(f;\tilde{\mathcal{P}})} + \tfrac{1}{n}\right]\cdot\mathrm{Avg}_{i\neq j\in[l]}\left[(V_f(\tilde{S}_i, \tilde{S}_j) + B_{ij}(\tfrac{\delta}{8l^2}))\cdot(1 + A_{ij}(\tfrac{\delta}{8l^2}))^2\right] + \frac{\sqrt{q}pkB\cdot\hat{\mathcal{C}}(f)}{\Delta\sqrt{l}}\cdot\mathrm{polylog}(k, \delta^{-1}, \hat{\mathcal{C}}(f), |\log(\tfrac{\Delta}{B})|, l)$$

*for some universal constant $C > 0$.*

Namely, the transfer error of $f$ is bounded by the sum of the averaged CDNV over the source training data, $V_f(\tilde{S}_i, \tilde{S}_j)$, combinations of the terms $A_{ij}(\frac{\delta}{8l^2})$ and $B_{ij}(\frac{\delta}{8l^2})$, where $\epsilon_1^c(\frac{\delta}{8l^2}) = \tilde{\mathcal{O}}\left(\frac{p\sqrt{q}\cdot B\cdot\hat{\mathcal{C}}(f)\cdot\log(l/\delta)}{\sqrt{m}}\right)$ and $\epsilon_2^c(\frac{\delta}{8l^2}) = \tilde{\mathcal{O}}\left(\frac{p\sqrt{q}\cdot B\cdot\hat{\mathcal{C}}(f)^2\cdot\log(l/\delta)}{\sqrt{m}}\right)$ (see Galanti et al., 2022) and a term that scales as $\tilde{\mathcal{O}}\left(\frac{\sqrt{q}pkB\cdot\hat{\mathcal{C}}(f)}{\Delta\sqrt{l}}\right)$. Therefore, if $l$ and $m$ are large and we have neural collapse across the source training samples (i.e., $\mathrm{Avg}_{i\neq j\in[l]}[V_f(\tilde{S}_i, \tilde{S}_j)]$ is small) and $\hat{\mathcal{C}}(f)$ is bounded (as a function of $m, l$), then the transfer error $\mathcal{L}_{\mathcal{D}}(f)$ is also small. Interestingly, in the presence of neural collapse in the source data, the dependence on $n$ is very weak.

## 5. Conclusions

Using pretrained models for transfer learning is a successful approach in the low-data regime. Recently, Galanti et al. (2022) suggested a new perspective on analysing this problem via the newly discovered phenomenon of neural collapse. However, their theoretical analysis relied on several unrealistic assumptions. In this work we provided better theoretical analysis by significantly relaxing their assumptions.

## 6. Acknowledgements

Tomer Galanti was supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216 and by NSF award 213418.

## References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6241–6250, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pp. 17–36, Bellevue, Washington, USA, 02 Jul 2012. PMLR.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Caruana, R. Learning many related tasks at the same time with backpropagation. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995.

Dhillon, G. S., Chaudhari, P., Ravichandran, A., and Soatto, S. A baseline for few-shot image classification. In *International Conference on Learning Representations*, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017.

Galanti, T. and Galanti, L. On the implicit bias towards minimal depth of deep neural networks, 2022.

Galanti, T., György, A., and Hutter, M. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SwIp410B6aQ.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9, 12 2017. doi: 10.1093/imaiai/iaz007.

Lee, K., Maji, S., Ravichandran, A., and Soatto, S. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

Maurer, A. and Pontil, M. Uniform concentration and symmetrization for weak interactions. In *COLT*, 2019.

Papyan, V., Han, X. Y., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: A good embedding is all you need? In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12359 of *Lecture Notes in Computer Science*, pp. 266–282. Springer, 2020.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

# A. Proof of Lemma 4.2

To prove Lemma 4.2, we apply Thm. 2 of Maurer & Pontil (2019). In order to do so, we use the following definitions (their Definition 1): Let $\|\cdot\|$ be the Euclidean norm, $\mathcal{U} \subset \mathbb{R}^d$ be any set and $A : \mathcal{U}^l \to \mathbb{R}$. For $\mathbf{z} = (z_1, \dots, z_l)$, $j \in \{1, \dots, l\}$ and $z'_j \in \mathcal{U}$, we define the $j$-th partial difference operator on $A$ as

$$D^j_{z'_j} A(\mathbf{z}) \;:=\; A(\dots, z_{j-1}, z_j, z_{j+1} \dots) - A(\dots, z_{j-1}, z'_j, z_{j+1}, \dots).$$

In addition, we denote

$$M(A) \;:=\; \max_{j \in [l]} \sup_{\substack{\mathbf{z} \in \mathcal{U}^l \\ z'_j \in \mathcal{U}:\, z'_j \neq z_j}} \frac{D^j_{z'_j} A(\mathbf{z})}{\|z_j - z'_j\|},$$

and

$$J(A) \;:=\; l \cdot \max_{\substack{r,j \in [l] \\ r \neq j}} \sup_{\substack{\mathbf{z} \in \mathcal{U}^l \\ z'_j \in \mathcal{U}:\, z'_j \neq z_j \\ z'_r \in \mathcal{U}:\, z'_r \neq z_r}} \frac{D^r_{z'_r} D^j_{z'_j} A(\mathbf{z})}{\|z_j - z'_j\|},$$

and the maximal difference

$$K(A) \;:=\; \max_{j \in [l]} \sup_{\substack{\mathbf{z} \in \mathcal{U}^l \\ z'_j \in \mathcal{U}:\, z'_j \neq z_j}} D^j_{z'_j} A(\mathbf{z}).$$

Now we are ready to proceed with the proof.

**Lemma A.1.** *Let $\delta \in (0,1)$, $l \geq 2$ and let $\mathcal{F}$ be a class of ReLU neural networks of depth $q$. Then, with probability at least $1-\delta$ over the random selection of the class-conditionals $\tilde{\mathcal{P}} = \{\tilde{P}_c\}_{c=1}^l$ of the source classes, for any $f \in \mathcal{F}$ and $\Delta > 0$, we have*

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(f) \;\leq\; & (k-1) \cdot \left[\mathrm{Avg}_{i \neq j \in [l]} [L^{2\Delta}_{ij}(f)] + \delta\right] \\ & + (k-1) \frac{\mathcal{C}(f) \cdot pB}{\Delta \sqrt{l}} \left( 128(3+\sqrt{2}) \sqrt{q\pi \log(l)} + (8+6\sqrt{2}) \sqrt{\log\left( \frac{\sqrt{3}(\mathcal{C}(f)+2)(|\log(\frac{\Delta}{B})|+2)}{\sqrt{\delta}} \right)} \right),\end{aligned}$$

*where $L^{\Delta}_{ij}(f) := \Pr_{\hat{x}_i, \hat{S}_i, \hat{S}_j}\left[\|f(\hat{x}_i) - \mu_f(\hat{S}_j)\| \leq \|f(\hat{x}_i) - \mu_f(\hat{S}_i)\| + \Delta\right]$, where $\hat{S}_i \sim \tilde{P}_i^n, \hat{S}_j \sim \tilde{P}_j^n$ and $\hat{x}_i \sim \tilde{P}_i$.*

*Proof.* In addition, we let $U_1 = (x_1, S_1)$ and $U_2 = (x_2, S_2)$, where $x_i \sim P_i$ and $S_i \sim P_i^n$ (for $i = 1, 2$).

We start the proof by reducing the error bound to a two-class classification problem:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(f) &= \mathbb{E}_{P_1,\dots,P_k} \mathbb{E}_{S_1,\dots,S_k} \mathbb{E}_{(x,y)\sim P} \left[\mathbb{I}[h_{f,S}(x) \neq y]\right] \\ &= \mathbb{E}_{P_1,\dots,P_k} \mathbb{E}_{S_1,\dots,S_k} \mathbb{E}_{(x,y)\sim P} \left[\mathbb{I}\left[\exists i \neq y :\; \|f(x) - \mu_f(P_i)\| \leq \|f(x) - \mu_f(P_y)\|\right]\right] \\ &\leq \mathbb{E}_{P_1,\dots,P_k} \mathbb{E}_{S_1,\dots,S_k} \mathbb{E}_{(x,y)\sim P} \left[\sum_{i \neq y} \mathbb{I}\left[\|f(x) - \mu_f(P_i)\| \leq \|f(x) - \mu_f(P_y)\|\right]\right] \\ &= \mathbb{E}_{P_1,\dots,P_k} \mathbb{E}_{S_1,\dots,S_k} \mathbb{E}_{y\sim U[k]} \mathbb{E}_{x\sim P_y} \left[\sum_{i \neq y} \mathbb{I}\left[\|f(x) - \mu_f(P_i)\| < \|f(x) - \mu_f(P_y)\|\right]\right] \\ &= \mathbb{E}_{y\sim U[k]} \sum_{i \neq y} \mathbb{E}_{P_i,P_y} \mathbb{E}_{S_i,S_y} \mathbb{E}_{x\sim P_y} \left[\mathbb{I}\left[\|f(x) - \mu_f(P_y)\| < \|f(x) - \mu_f(P_i)\|\right]\right] \\ &= (k-1) \cdot \mathbb{E}_{P_1,P_2} \mathbb{E}_{S_1,S_2} \mathbb{E}_{x_1\sim P_1} \left[\mathbb{I}\left[\|f(x) - \mu_f(P_2)\| < \|f(x) - \mu_f(P_1)\|\right]\right]\end{aligned}$$

To bound the right-hand side above further, we introduce a "soft-margin" relaxation $\ell_\Delta : \mathbb{R} \to [0, \infty)$ of the zero-one loss with margin $\Delta$ as

$$\ell_\Delta(r) \;:=\; \begin{cases} 0 & \text{if } r < -\Delta; \\ 1 + \frac{r}{\Delta} & \text{if } r \in [-\Delta, 0]; \\ 1 & \text{if } r > 0. \end{cases}$$

Then, clearly $\ell_\Delta(-r) \geq \mathbb{I}[r < 0]$ for all $\Delta > 0$ and $r \in \mathbb{R}$, and so

$$\mathbb{I}\left[\|f(x) - \mu_f(P_2)\| < \|f(x) - \mu_f(P_1)\|\right] \;\leq\; \ell_\Delta\left(\|f(x) - \mu_f(P_2)\| - \|f(x) - \mu_f(P_1)\|\right),$$

implying

$$\mathcal{L}_{\mathcal{D}}(f) \;\leq\; (k-1) \cdot \mathbb{E}_{P_1,P_2}\mathbb{E}_{S_1,S_2}\mathbb{E}_{x_1\sim P_1}\left[\ell_\Delta\left(\|f(x) - \mu_f(P_2)\| - \|f(x) - \mu_f(P_1)\|\right)\right]. \tag{4}$$

By the random selection of $P_1, P_2, \tilde{P}_c, c \in [l]$, the main term on the right-hand side above can be expressed as

$$\mathrm{Avg}_{i\neq j\in[l]}\, \mathbb{E}_{\tilde{P}_i,\tilde{P}_j}\mathbb{E}_{\hat{S}_i,\hat{S}_j}\mathbb{E}_{\hat{x}_i\sim\tilde{P}_i}\left[\ell_\Delta\left(\|f(\hat{x}_i) - \mu_f(\tilde{P}_j)\| - \|f(\hat{x}_i) - \mu_f(\tilde{P}_i)\|\right)\right]. \tag{5}$$

To write this averaging succinctly, we introduce some notation. Let $\hat{U}_c := (\hat{x}_c, \hat{S}_c)$ for $c \in [l]$ where $\hat{x}$ and $\hat{S}_c$ are defined in the statement of the lemma ($\hat{x}_c \sim \tilde{P}_c$ and $\hat{S}_c \sim \tilde{P}_c^n$, independently for all $c$), and and $\hat{\mathcal{U}} := \{\hat{U}_c\}_{c=1}^l$. Furthermore, let $\mathcal{Q} := \{U = (x, S) \mid x \in \mathcal{X},\ S \in \mathcal{X}^n\}$ denote the set of all point-sample pairs (note that $\hat{U}_c \in \mathcal{Q}$). For any $f \in \mathcal{F}$, define $H_f : \mathcal{Q} \to \mathbb{R}$ as $H_f(x, S) = (f(x), \mu_f(S))$ and let $H_{\mathcal{F}}(\mathcal{Q}) = \{H_f(x, S) : (x, S) \in \mathcal{Q}, f \in \mathcal{F}\}$. Finally define the aggregation function $A_\Delta : H_{\mathcal{F}}(\mathcal{Q})^l \to \mathbb{R}$ as

$$A_\Delta(\mathbf{z}) \;:=\; \mathrm{Avg}_{i\neq j\in[l]}[\ell_\Delta(\|u_i - v_j\| - \|u_i - v_i\|)],$$

where $\mathbf{z} = (z_1, \ldots, z_l)$ such that $z_i = (u_i, v_i) \in H_{\mathcal{F}}(\mathcal{Q})$ for all $i \in [l]$. Then

$$A_\Delta(H_f(\hat{\mathcal{U}})) \;=\; \mathrm{Avg}_{i\neq j\in[l]}\left[\ell_\Delta(\|f(\hat{x}_i) - \mu_f(\hat{S}_j)\| - \|f(\hat{x}_i) - \mu_f(\hat{S}_i)\|)\right]$$

by definition, and (5) can be rewritten as $\mathbb{E}_{\hat{\mathcal{U}}}\left[A_\Delta(H_f(\hat{\mathcal{U}}))\right]$ and from (4) we get

$$\mathcal{L}_{\mathcal{D}}(f) \;\leq\; (k-1) \cdot \mathbb{E}_{\hat{\mathcal{U}}}\left[A_\Delta(H_f(\hat{\mathcal{U}}))\right]. \tag{6}$$

Our next step is to bound the expectation on the right-hand side above with the randomly selected samples of class-conditional distributions $\tilde{P}_c, c \in [l]$ for a fixed $\Delta$. To this end, we will use the notion of Rademacher complexity: the Rademacher complexity of a set of vectors $Y \subset \mathbb{R}^s$ is defined as $R(Y) := \mathbb{E}_\epsilon \sup_{y\in Y}[\langle \epsilon, y\rangle]$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_s) \sim U[\{\pm 1\}^s]$. Let $\delta' \in (0, 1)$ and $\mathcal{F}' \subset \mathcal{F}$ with $\alpha = \sup_{f\in\mathcal{F}'} \mathcal{C}(f)$. Then, by Corollary 3 of Maurer & Pontil (2019)[1], with probability at least $1 - \delta'$ over the selection of $\hat{\mathcal{U}}$, for all $f \in \mathcal{F}'$, we have

$$\begin{aligned} \mathbb{E}_{\hat{\mathcal{U}}}\left[A_\Delta(H_f(\hat{\mathcal{U}}))\right] \;\leq\; & A_\Delta(H_f(\hat{\mathcal{U}})) + 2\sqrt{2\pi\log(l)}\big(2\alpha(A_\Delta) + J(A_\Delta)\big) \cdot \mathbb{E}_{\hat{\mathcal{U}}}[R(H_{\mathcal{F}'}(\hat{\mathcal{U}}))] \\ & + K(A_\Delta) \cdot \sqrt{l\log\left(\tfrac{1}{\delta'}\right)} \end{aligned} \tag{7}$$

Next we bound separately the terms $K(A_\Delta)$, $M(A_\Delta)$, $J(A_\Delta)$.

**Bounding $K(A_\Delta)$, $M(A_\Delta)$ and $J(A_\Delta)$.** Let $\Delta > 0$, $j \in [l]$ and $z_i = (u_i, v_i) \in \mathbb{R}^p \times \mathbb{R}^p$ for $i \in [l]$. Let $\mathbf{z} = (z_1, \ldots, z_l)$ and $\mathbf{z} = (z_1', \ldots, z_l')$, where $z_i = z_i'$ for all $i \neq j \in [l]$. We have

$$\begin{aligned} D^j_{z_j'} A_\Delta(\mathbf{z}) &= A_\Delta(\mathbf{z}) - A_\Delta(\mathbf{z}') \\ &= \frac{1}{l(l-1)} \sum_{i\in[l]:\ i\neq j} \big(\ell_\Delta(\|u_i - v_j\| - \|u_i - v_i\|) - \ell_\Delta(\|u_i - v_j'\| - \|u_i - v_i\|)\big) \\ &\quad + \frac{1}{l(l-1)} \sum_{i\in[l]:\ i\neq j} \big(\ell_\Delta(\|u_j - v_i\| - \|u_j - v_j\|) - \ell_\Delta(\|u_j' - v_i\| - \|u_j' - v_j'\|)\big). \end{aligned} \tag{8}$$

[1] The original statement makes use of the Gaussian complexity of instead of its Rademacher complexity. The Gaussian complexity of a set $A \subset \mathbb{R}^l$ is upper bounded by $2\sqrt{\log(l)}R(A)$.

Since $\ell_\Delta$ is $\Delta^{-1}$-Lipschitz and by the triangle inequality

$$\begin{aligned}
&\ell_\Delta(\|u_i - v_j\| - \|u_i - v_i\|) - \ell_\Delta(\|u_i - v'_j\| - \|u_i - v_i\|) \\
&\leq \Delta^{-1}\Big|\|u_i - v_j\| - \|u_i - v_i\| - \|u_i - v'_j\| + \|u_i - v_i\|\Big| \\
&= \Delta^{-1}\Big|\|u_i - v_j\| - \|u_i - v'_j\|\Big| \\
&\leq \Delta^{-1}\|v_j - v'_j\| \qquad (9)\\
&\leq \Delta^{-1}\|(u_j, v_j) - (u'_j, v'_j)\| \\
&= \Delta^{-1}\|z_j - z'_j\|. \qquad (10)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\ell_\Delta(\|u_j - v_i\| - \|u_j - v_j\|) - \ell_\Delta(\|u'_j - v_i\| - \|u'_j - v'_j\|) \\
&\leq \Delta^{-1}\Big|\|u_j - v_i\| - \|u_j - v_j\| - \|u'_j - v_i\| + \|u'_j - v'_j\|\Big| \\
&\leq \Delta^{-1}\left(\Big|\|u_j - v_j\| - \|u'_j - v'_j\|\Big| + \Big|\|u_j - v_i\| - \|u'_j - v_i\|\Big|\right) \\
&\leq \Delta^{-1}\left(\|u_j - u'_j\| + \|v_j - v'_j\| + \|u_j - u'_j\|\right) \\
&\leq \Delta^{-1}\left(\|u_j - u'_j\| + \|v_j - v'_j\| + \|(u_j, v_j) - (u'_j, v'_j)\|\right) \\
&\leq (1+\sqrt{2})\Delta^{-1}\|(u_j, v_j) - (u'_j, v'_j)\| \qquad (11)\\
&= (1+\sqrt{2})\Delta^{-1}\|z_j - z'_j\|, \qquad (12)
\end{aligned}$$

where the last inequality follows from the arithmetic mean–quadratic mean inequality. Combining (8), (10) and (12), we obtain that

$$M(A_\Delta) \;=\; \max_{j\in[l]} \sup_{\substack{\mathbf{z}\in\mathcal{U}^l \\ z'_j\in\mathcal{U}:\, z'_j\neq z_j}} \frac{D^j_{z'_j} A(\mathbf{z})}{\|z_j - z'_j\|} \;\leq\; \frac{2+\sqrt{2}}{\Delta l}\,. \qquad (13)$$

Using that $\|u_j - u'_j\|$ and $\|v_j - v'_j\|$ are both bounded by $2\max_{x\in\mathcal{X}, f\in\mathcal{F}'}\|f(x)\| \leq 2p\alpha B$ for any $\mathbf{z}, \mathbf{z}'$, we obtain from (8), (9) and (11) that

$$D^j_{z'_j} A_\Delta(\mathbf{z}) \leq \frac{6+4\sqrt{2}}{\Delta l} \max_{x\in\mathcal{X}, f\in\mathcal{F}'}\|f(x)\| \leq \frac{(6+4\sqrt{2})p\alpha B}{\Delta l}\,.$$

Hence,

$$K(A_\Delta) \leq \frac{(6+4\sqrt{2})p\alpha B}{\Delta l}\,. \qquad (14)$$

Next, we bound $J(A_\Delta)$. Let $j, r \in [l]$ with $j \neq r$ and $\mathbf{z} = (z_1, \ldots, z_l) \in H_{\mathcal{F}}(\mathcal{Q})^l$. We have

$$\begin{aligned}
D^r_{z'_r} D^j_{z'_j} A_\Delta(\mathbf{z}) &= \frac{1}{l(l-1)} \sum_{i\in[l]:\, i\neq j} D^r_{z'_r}\left(\ell_\Delta(\|u_i - v_j\| - \|u_i - v_i\|) - \ell_\Delta(\|u_i - v'_j\| - \|u_i - v_i\|)\right) \\
&\quad + \frac{1}{l(l-1)} \sum_{i\in[l]:\, i\neq j} D^r_{z'_r}\left(\ell_\Delta(\|u_j - v_i\| - \|u_j - v_j\|) - \ell_\Delta(\|u'_j - v_i\| - \|u'_j - v'_j\|)\right) \\
&= \frac{1}{l(l-1)} \cdot D^r_{z'_r}\left(\ell_\Delta(\|u_r - v_j\| - \|u_r - v_r\|) - \ell_\Delta(\|u_r - v'_j\| - \|u_r - v_r\|)\right) \\
&\quad + \frac{1}{l(l-1)} \cdot D^r_{z'_r}\left(\ell_\Delta(\|u_j - v_r\| - \|u_j - v_j\|) - \ell_\Delta(\|u'_j - v_r\| - \|u'_j - v'_j\|)\right), \qquad (15)
\end{aligned}$$

where the second equality follows as all terms are 0 for which $i \neq r$, since the function in that case is constant in $u_r$ (since

$j \neq r$ by our definition). Using again that $\ell_\Delta$ is $\Delta^{-1}$-Lipschitz, we have

$$\begin{aligned}
&D^r_{z'_r}\left(\ell_\Delta(\|u_r - v_j\| - \|u_r - v_r\|) - \ell_\Delta(\|u_r - v'_j\| - \|u_r - v_r\|)\right) \\
&= \ell_\Delta(\|u_r - v_j\| - \|u_r - v_r\|) - \ell_\Delta(\|u_r - v'_j\| - \|u_r - v_r\|) \\
&\quad - \ell_\Delta(\|u'_r - v_j\| - \|u'_r - v'_r\|) + \ell_\Delta(\|u'_r - v'_j\| - \|u'_r - v'_r\|) \\
&\leq 2\Delta^{-1}\|v_j - v'_j\| \,. \qquad (16)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&D^r_{z'_r}\left(\ell_\Delta(\|u_j - v_r\| - \|u_j - v_j\|) - \ell_\Delta(\|u'_j - v_r\| - \|u'_j - v'_j\|)\right) \\
&= \ell_\Delta(\|u_j - v_r\| - \|u_j - v_j\|) - \ell_\Delta(\|u_j - v'_r\| - \|u_j - v_j\|) \\
&\quad - \ell_\Delta(\|u'_j - v_r\| - \|u'_j - v'_j\|) + \ell_\Delta(\|u'_j - v'_r\| - \|u'_j - v'_j\|) \\
&\leq |\ell_\Delta(\|u_j - v_r\| - \|u_j - v_j\|) - \ell_\Delta(\|u'_j - v_r\| - \|u'_j - v'_j\|)| \\
&\quad + |\ell_\Delta(\|u_j - v'_r\| - \|u_j - v_j\|) - \ell_\Delta(\|u'_j - v'_r\| - \|u'_j - v'_j\|)| \\
&\leq 2\Delta^{-1}\left(\|u_j - u'_j\| + \|u_j - u'_j + v'_j - v_j\|\right) \\
&\leq 2\Delta^{-1}\left(2\|u_j - u'_j\| + \|v'_j - v_j\|\right) \qquad (17)
\end{aligned}$$

Combining (15)–(17) with the arithmetic mean–quadratic mean inequality, we obtain

$$\frac{D^r_{z'_r} D^j_{z'_j} A_\Delta(\mathbf{z})}{\|z_j - z'_j\|} \leq \frac{4\left(\|u_j - u'_j\| + \|v_j - v'_j\|\right)}{l(l-1)\Delta\|z_j - z'_j\|} \leq \frac{4\sqrt{2}}{l(l-1)\Delta},$$

which gives

$$J(A_\Delta) \leq \frac{4\sqrt{2}}{\Delta(l-1)} \,. \qquad (18)$$

Substituting (13), (14), and (18) into (7), we get for and $f \in \mathcal{F}'$ with $\sup_{f\in\mathcal{F}'} \mathcal{C}(f) \leq \alpha$ that with probability at least $1-\delta'$ over the selection of $\hat{\mathcal{U}}$,

$$\mathbb{E}_{\hat{\mathcal{U}}}\left[A_\Delta(H_f(\hat{\mathcal{U}}))\right] \;\leq\; A_\Delta(H_f(\hat{\mathcal{U}})) + \frac{8(3+\sqrt{2})\sqrt{\pi\log(l)}\cdot\mathbb{E}_{\hat{\mathcal{U}}}[R(H_{\mathcal{F}'}(\hat{\mathcal{U}}))]}{\Delta(l-1)} + \frac{(6+4\sqrt{2})p\alpha B}{\Delta\sqrt{l}}\sqrt{\log\left(\tfrac{1}{\delta'}\right)} \,. \qquad (19)$$

To be able to adapt to the complexity $\mathcal{C}(f)$ of $f$ and prove a bound for all $\Delta$, we will apply the above to $\mathcal{F}' = \mathcal{F}^\alpha := \{f \in \mathcal{F} : \mathcal{C}(f) \leq \alpha\}$ with positive integers $\alpha$ and for an exponential grid for $\Delta$ defined as $\Delta_t = 2^{-t}B$ for any integer $t$ and $\delta' = \delta_{t,\alpha}$ where $\delta_{t,\alpha} := \frac{\delta}{3\alpha(\alpha+1)|t|(|t|+1)}$ for all $t \neq 0$ and $\delta_{0,\alpha} := \frac{\delta}{3\alpha(\alpha+1)}$. Note that by definition, $\sum_{t=-\infty}^{\infty}\sum_{\alpha=1}^{\infty}\delta_{t,\alpha} = \delta$.

**Bounding $\mathbb{E}_{\hat{\mathcal{U}}}[R(H_{\mathcal{F}^\alpha}(\hat{\mathcal{U}}))]$.** As a next step, we upper bound $\mathbb{E}_{\hat{\mathcal{U}}}[R(H_{\mathcal{F}^\alpha}(\hat{\mathcal{U}}))]$. Let $\epsilon$ be the aggregation of i.i.d. Rademacher random variables $\epsilon^1_{cj}$ and $\epsilon^2_{cj}$ for $c \in [l]$ and $j \in [p]$. Then

$$\begin{aligned}
\mathbb{E}_{\hat{\mathcal{U}}}[R(H_{\mathcal{F}^\alpha}(\hat{\mathcal{U}}))] \;&=\; \mathbb{E}_{\hat{\mathcal{U}}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}^\alpha}\sum_{c=1}^{l}\sum_{j=1}^{p} f(\hat{x}_c)_j\epsilon^1_{cj} + \mu_f(\hat{S}_c)_j\epsilon^2_{cj}\right] \\
&\leq\; \sum_{j=1}^{p}\mathbb{E}_{\hat{\mathcal{U}}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}^\alpha}\sum_{c=1}^{l} f(\hat{x}_c)_j\epsilon^1_{cj}\right] + \sum_{j=1}^{p}\mathbb{E}_{\hat{\mathcal{U}}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}^\alpha}\sum_{c=1}^{l}\mu_f(\hat{S}_c)_j\epsilon^2_{cj}\right] \\
&\leq\; \sum_{j=1}^{p}\mathbb{E}_{\hat{\mathcal{U}}}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}^\alpha}\sum_{c=1}^{l} f(\hat{x}_c)_j\epsilon^1_{cj}\right] + \sum_{j=1}^{p}\mathbb{E}_{\hat{\mathcal{U}}}\mathbb{E}_{\hat{x}_c\sim\hat{S}_c}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}^\alpha}\sum_{c=1}^{l}\mu_f(\hat{x}_c)_j\epsilon^2_{cj}\right] \\
&=\; 2\sum_{j=1}^{p}\mathbb{E}_{\hat{x}_1,\ldots,\hat{x}_c}\left[R(\mathcal{F}^\alpha_{|j}(\{\hat{x}_c\}_{c=1}^{l}))\right],
\end{aligned}$$

where $f(\hat{x})_j$ denotes the $j$th coordinate of $f(\hat{x})$, and $\mathcal{F}^{\alpha}_{|j} := \{f(x)_j \mid f \in \mathcal{F}\}$. Finally, by Theorem 1 of Golowich et al. (2017),[2] we have

$$\sum_{j=1}^{p} \mathbb{E}_{\hat{x}_1,\ldots,\hat{x}_c}\left[R(\mathcal{F}^{\alpha}_{|j}(\{\hat{x}_c\}_{c=1}^{l}))\right] \;\le\; 4\alpha p\sqrt{ql}\cdot \sup_{x\in\mathcal{X}}\|x\| \,.$$

In particular,

$$\mathbb{E}_{\hat{\mathcal{U}}}[R(H_{\mathcal{F}^{\alpha}}(\hat{\mathcal{U}}))] \;\le\; 8\alpha p\sqrt{ql}\cdot \sup_{x\in\mathcal{X}}\|x\| \le 8\alpha p\sqrt{ql}B \,. \tag{20}$$

Substituting this into (19) and using that $l-1 \ge l/2$ for $l \ge 2$ and the union bound, we obtain that with probability at least $1-\delta$, for all $\alpha \in \mathbb{N}$ and $t \in \mathbb{Z}$, and all $f \in \mathcal{F}^{\alpha}$,

$$\begin{aligned}
&\mathbb{E}_{\hat{\mathcal{U}}}\left[A_{\Delta_t}(H_f(\hat{\mathcal{U}}))\right] \\
&\le A_{\Delta_t}(H_f(\hat{\mathcal{U}})) + \frac{\alpha p B}{\Delta_t\sqrt{l}}\left(128(3+\sqrt{2})\sqrt{q\pi\log(l)} + (6+4\sqrt{2})\sqrt{\log\left(\frac{3(\alpha+1)^2(|t|+1)^2}{\delta}\right)}\right) \,.
\end{aligned}$$

Now, for any $\Delta > 0$ there exists a $t$ such that $\Delta \le \Delta_t \le 2\Delta$; in this case $|t| \le |\log(\frac{\Delta}{B})| + 1$. Now, since $\ell_{\Delta}$, and hence that of $A_{\Delta}$, monotonically increases in $\Delta$, the above inequality with $\alpha = \lceil \mathcal{C}(f) \rceil \le \mathcal{C}(f) + 1$ implies

$$\begin{aligned}
&\mathbb{E}_{\hat{\mathcal{U}}}\left[A_{\Delta}(H_f(\hat{\mathcal{U}}))\right] \\
&\le A_{2\Delta}(H_f(\hat{\mathcal{U}})) + \frac{(\mathcal{C}(f)+1)\cdot pB}{\Delta\sqrt{l}}\left(128(3+\sqrt{2})\sqrt{q\pi\log(l)} + (8+6\sqrt{2})\sqrt{\log\left(\frac{\sqrt{3}(\mathcal{C}(f)+2)(|\log(\frac{\Delta}{B})|+2)}{\sqrt{\delta}}\right)}\right) \,.
\end{aligned} \tag{21}$$

Taking expectation with respect to $\hat{\mathcal{U}}$ given $\tilde{\mathcal{P}}$, we have

$$\begin{aligned}
\mathbb{E}_{\hat{\mathcal{U}}}[A_{2\Delta}(H_f(\hat{\mathcal{U}})) \mid \tilde{\mathcal{P}}] &= \mathrm{Avg}_{i\neq j\in[l]}\,\mathbb{E}_{\hat{\mathcal{U}}}\left[\ell_{2\Delta}(\|f(x_i)-\mu_f(\hat{S}_j)\| - \|f(x_i)-\mu_f(\hat{S}_i)\|) \mid \tilde{\mathcal{P}}\right] \\
&\le \mathrm{Avg}_{i\neq j\in[l]}\left[\Pr_{\hat{x}_i,\hat{S}_i,\hat{S}_j}[\|f(x_i)-\mu_f(P_j)\| \le \|f(x_i)-\mu_f(P_j)\| + 2\Delta]\right] \\
&= \mathrm{Avg}_{i\neq j\in[l]}[L^{2\Delta}_{ij}(f)] \,.
\end{aligned}$$

Combining with (21) and also that $\mathbb{E}_{\hat{\mathcal{U}}}\left[A_{\Delta}(H_f(\hat{\mathcal{U}}))\right] \le 1$ then yields

$$\begin{aligned}
\mathbb{E}_{\hat{\mathcal{U}}}\left[A_{\Delta}(H_f(\hat{\mathcal{U}}))\right] \;\le\; &\mathrm{Avg}_{i\neq j\in[l]}[L^{2\Delta}_{ij}(f)] + \delta \\
&+ \frac{(\mathcal{C}(f)+1)\cdot pB}{\Delta\sqrt{l}}\left(128(3+\sqrt{2})\sqrt{q\pi\log(l)} + (8+6\sqrt{2})\sqrt{\log\left(\frac{\sqrt{3}(\mathcal{C}(f)+2)(|\log(\frac{\Delta}{B})|+2)}{\sqrt{\delta}}\right)}\right) \,.
\end{aligned}$$

The statement of the lemma now follows by combining this inequality with (6).

□

## B. Proof of Lemma 4.3

**Lemma B.1.** *Let $\{\tilde{P}_i, \tilde{P}_j\}$ be a pair of class-conditional distributions for some given $i, j \in [l], i \neq j$. Let $f : \mathbb{R}^d \to \mathbb{R}^p$ be any feature map, and let $\Delta \le 0.12 s(f,\tilde{\mathcal{P}})^{-1/2}\cdot\|\mu_f(\tilde{P}_i) - \mu_f(\tilde{P}_j)\|$. Then, we have*

$$L^{\Delta}_{ij}(f) \overset{\times}{\le} \left[\tfrac{1}{s(f,\tilde{\mathcal{P}})} + \tfrac{1}{n}\right]\cdot V_f(\tilde{P}_i, \tilde{P}_j),$$

*where $s(f,\tilde{\mathcal{P}}) = p$ if $\{f\circ\tilde{P}_c\}_{c=1}^{l}$ are spherically symmetric and $s(f,\tilde{\mathcal{P}}) = 1$ otherwise.*

[2]The original statement makes use of the Frobenius norms of the matrices, however, the proof is exactly the same if we replace the Frobenius norm with the spectral norm.

**Proof.** For $c \in \{i, j\}$, let $\hat{S}_c \sim \tilde{P}_c^n$ and $\hat{x}_c \sim \tilde{P}_c$, $\hat{u}_c = f(\hat{x}_c)$, $\mu_c = \mu_f(\tilde{P}_c)$, $\hat{\mu}_c = \mu_f(\hat{S}_c)$ and $\sigma_c^2 = \mathrm{Var}(\hat{u}_c)$. Let $d_{ij} := \|\mu_i - \mu_j\|$ denote the considered centers' true distance, which sets the overall scale, and $\hat{d}_{ij} := \|\hat{\mu}_i - \hat{\mu}_j\|$ its estimate. Let $s := s(f, \tilde{\mathcal{P}})$ and $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in (0, 1)$ be constants that satisfy $\gamma_4 - 2\gamma_3 - 4\gamma_1 - 2\gamma_2^2/s > 0$ (to be selected by the end of the proof). Moreover, let $\alpha_{ij} := d_{ij}^2 \gamma_1 / \Delta$, where $\Delta = d_{ij}\gamma_2/\sqrt{s}$.

The basic proof idea is graphically illustrated in Figure 1: Consider the error event $\mathrm{Err} := [\|\hat{u}_i - \hat{\mu}_j\| \leq \|\hat{u}_i - \hat{\mu}_i\| + \Delta]$, then $L_{ij}^{\Delta}(f) = \Pr[\mathrm{Err}]$. For $\Delta = 0$, Err is a $p$-dimensional half-space orthogonal to the unit vector $\hat{e}_{ij} := (\hat{\mu}_j - \hat{\mu}_i)/\hat{d}_{ij}$, and to determine its probability, we can simply project all quantities onto the line $\mathbb{R} \cdot \hat{e}_{ij}$. This makes the problem one-dimensional and improves the bound by a factor of $p^{-1}$ in the spherically symmetric case.

However, one complication is that $\hat{e}_{ij}$ itself is random, so in order to indeed achieve a scale of $1/p$, one has to carefully avoid using triangle inequalities involving $\hat{u}_i$, which complicates the analysis, and also gives an extra $1/n$ term. $\Delta > 0$ involves further complications: Consider the error set $\mathrm{ErrU}(\hat{\mu}_i, \hat{\mu}_i) := \{\hat{u}_i : \|\hat{u}_i - \hat{\mu}_j\| \leq \|\hat{u}_i - \hat{\mu}_i\| + \Delta\}$. Its boundary $\partial\mathrm{ErrU}$ is now (one sheet of an) elliptic hyperboloid, and direct projection onto $\hat{e}_{ij}$ does not work anymore. If we confine $\hat{u}_i$ to the hyperball $B := \{\hat{u}_i : \|\hat{u}_i - \hat{\mu}_i\| \leq \alpha_{ij}\}$, then $B$ cuts ErrU into a finite moon-shaped set, and the projection method works again, while the probability of $\hat{u}_i \notin B$ scales with $1/p$.

We derive a sequence of bounds (A)-(H) which we chain all together at the end.

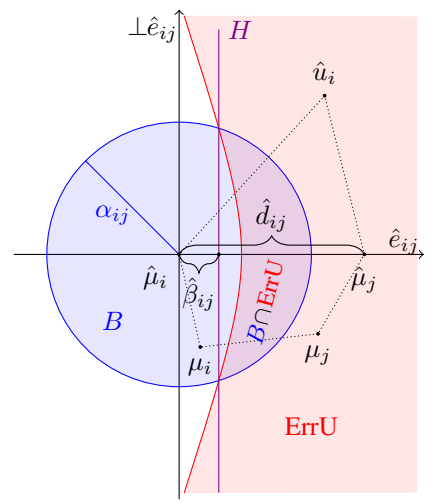


*Figure 1.* A geometric illustration of the various components in the proof of Lemma A.1. The blue section denotes the ball $B$ of radius $\alpha_{ij}$ and the red section denotes the set ErrU.

**(A)** As a first step we determine the halfspace (given $\hat{\mu}_i$ and $\hat{\mu}_j$) that covers the moon $B \cap \mathrm{ErrU}$. Assume that $\|\hat{u}_i - \hat{\mu}_j\| \leq \|\hat{u}_i - \hat{\mu}_i\| + \Delta$ and $\|\hat{u}_i - \hat{\mu}_i\| \leq \alpha_{ij}$. Then,

$$\begin{aligned}
\|\hat{u}_i - \hat{\mu}_i\|^2 + 2\alpha_{ij}\Delta + \Delta^2 &\geq \|\hat{u}_i - \hat{\mu}_i\|^2 + 2\|\hat{u}_i - \hat{\mu}_i\| \cdot \Delta + \Delta^2 \\
&= (\|\hat{u}_i - \hat{\mu}_i\| + \Delta)^2 \\
&\geq \|\hat{u}_i - \hat{\mu}_j\|^2 \\
&= \|(\hat{u}_i - \hat{\mu}_i) - (\hat{\mu}_j - \hat{\mu}_i)\|^2 \\
&= \|\hat{u}_i - \hat{\mu}_i\|^2 - 2(\hat{u}_i - \hat{\mu}_i)^\top(\hat{\mu}_j - \hat{\mu}_i) + \hat{d}_{ij}^2.
\end{aligned}$$

Therefore, we have $2(\hat{u}_i - \hat{\mu}_i)^\top(\hat{\mu}_j - \hat{\mu}_i) \geq \hat{d}_{ij}^2 - 2\alpha_{ij}\Delta - \Delta^2$, which can also be written as follows

$$(\hat{u}_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \hat{\beta}_{ij} := \tfrac{1}{2}\hat{d}_{ij} - \frac{2\alpha_{ij}\Delta + \Delta^2}{2\hat{d}_{ij}}.$$

This is the halfspace $H$ that covers $B \cap \mathrm{ErrU}$ most tightly.

**(B)** We can use this to bound the probability of the event Err in terms of $\alpha_{ij}$ and $\Delta$ by conditioning:

$$\begin{aligned}
\Pr[\mathrm{Err}] &= \Pr[\|\hat{u}_i - \hat{\mu}_j\| \leq \|\hat{u}_i - \hat{\mu}_i\| + \Delta] \\
&= \Pr[\|\hat{u}_i - \hat{\mu}_j\| \leq \|\hat{u}_i - \hat{\mu}_i\| + \Delta \ \wedge \ \|\hat{u}_i - \hat{\mu}_i\| \leq \alpha_{ij}] \\
&\quad + \Pr[\|\hat{u}_i - \hat{\mu}_j\| \leq \|\hat{u}_i - \hat{\mu}_i\| + \Delta \ \wedge \ \|\hat{u}_i - \hat{\mu}_i\| > \alpha_{ij}] \\
&\overset{(A)}{\leq} \Pr[(\hat{u}_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \hat{\beta}_{ij}] + \Pr[\|\hat{u}_i - \hat{\mu}_i\| > \alpha_{ij}] =: (C) + (F)
\end{aligned}$$

**(C)** Next we upper bound the term $\Pr[(\hat{u}_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \hat{\beta}_{ij}]$. Namely,

$$\begin{aligned}
&\Pr[(\hat{u}_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \hat{\beta}_{ij}] \\
&= \Pr[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} + (\mu_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \hat{\beta}_{ij}] \\
&\leq \Pr[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} + (\mu_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \gamma_3 d_{ij} \ \vee \ \hat{\beta}_{ij} \leq \gamma_3 d_{ij}] \\
&\leq \Pr[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} + (\mu_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \gamma_3 d_{ij}] \ + \ \Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij}] \ =: \ (D) \ + \ (H),
\end{aligned}$$

for $\gamma_3 \in (0, 1)$.

**(D)** By applying the union bound and use that a projection onto $\hat{e}_{ij}$ decreases the length of the vector $\mu_i - \hat{\mu}_i$, we get

$$\begin{aligned}
(D) \ &= \ \Pr\left[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} + (\mu_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \gamma_3 d_{ij}\right] \\
&\leq \ \Pr\left[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} \geq \gamma_3 d_{ij}/2 \ \vee \ (\mu_i - \hat{\mu}_i)^\top \hat{e}_{ij} \geq \gamma_3 d_{ij}/2\right] \\
&\leq \ \Pr\left[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} \geq \gamma_3 d_{ij}/2\right] + \Pr\left[\|\mu_i - \hat{\mu}_i\| \geq \gamma_3 d_{ij}/2\right] \ =: \ (E) \ + \ (G)
\end{aligned}$$

**(E)** Next we analyze each term separately. For a fixed pair $\hat{\mu}_i$ and $\hat{\mu}_j$, the variance of the random variable $(\hat{u}_i - \mu_i)^\top \hat{e}_{ij}$ is at most $\mathrm{Var}(\hat{u}_i)/s$, where we have exploited the projection in the spherically symmetric case. Therefore, by Markov's inequality,

$$\Pr\left[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} \geq d_{ij}\gamma_3/2\right] \ \leq \ \mathbb{E}_{\hat{\mu}_1, \hat{\mu}_2}\left[\Pr_{\hat{u}_i}\left[(\hat{u}_i - \mu_i)^\top \hat{e}_{ij} \geq d_{ij}\gamma_3/2\right]\right] \ \leq \ \mathbb{E}_{\hat{\mu}_1, \hat{\mu}_2}\left[\frac{4\gamma_3^{-2}\mathrm{Var}(\hat{u}_i)}{s \cdot d_{ij}^2}\right] \ = \ \frac{4\mathrm{Var}(\hat{u}_i)}{s \cdot \gamma_3^2 d_{ij}^2}.$$

**(F)** Again, by Markov's inequality, and using $\Delta \leq d_{ij}\gamma_2/\sqrt{s}$, hence $\alpha_{ij} = d_{ij}^2\gamma_1/\Delta \geq d_{ij}\sqrt{s}(\gamma_1/\gamma_2)$, we have

$$\begin{aligned}
\Pr\left[\|\hat{u}_i - \hat{\mu}_i\| \geq \alpha_{ij}\right] \ &\leq \ \Pr\left[\|\hat{\mu}_i - \mu_i\| \geq \alpha_{ij}/2\right] + \Pr\left[\|\hat{u}_i - \mu_i\| \geq \alpha_{ij}/2\right] \\
&\leq \ \frac{4\mathrm{Var}(\hat{\mu}_i)}{\alpha_{ij}^2} + \frac{4\mathrm{Var}(\hat{u}_i)}{\alpha_{ij}^2} \ \leq \ \frac{4\gamma_2^2\mathrm{Var}(\hat{\mu}_i)}{\gamma_1^2 \cdot s \cdot d_{ij}^2} + \frac{4\gamma_2^2\mathrm{Var}(\hat{u}_i)}{\gamma_1^2 \cdot s \cdot d_{ij}^2}
\end{aligned}$$

**(G)** and

$$\Pr\left[\|\hat{\mu}_i - \mu_i\| \geq \gamma_3 d_{ij}/2\right] \ \leq \ \frac{4\mathrm{Var}(\hat{\mu}_i)}{\gamma_3^2 d_{ij}^2} = \frac{4\mathrm{Var}(\hat{u}_i)}{n\gamma_3^2 d_{ij}^2}.$$

**(H)** Next, by conditioning,

$$\begin{aligned}
\Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij}] \ &= \ \Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij} \ \wedge \ \hat{d}_{ij} \leq \gamma_4 d_{ij}] \ + \ \Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij} \ \wedge \ \hat{d}_{ij} > \gamma_4 d_{ij}] \\
&\leq \ \Pr[\hat{d}_{ij} \leq \gamma_4 d_{ij}] \ + \ \Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij} \ \wedge \ \hat{d}_{ij} > \gamma_4 d_{ij}] \\
&\leq \ \Pr[\|\mu_j - \hat{\mu}_j\| + \|\mu_i - \hat{\mu}_i\| \geq (1 - \gamma_4) d_{ij}] \ + \ \Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij} \ \wedge \ \hat{d}_{ij} > \gamma_4 d_{ij}],
\end{aligned}$$

where the last inequality follows from solving the triangle inequality

$$d_{ij} \ \leq \ \|\mu_i - \hat{\mu}_i\| + \hat{d}_{ij} + \|\hat{\mu}_j - \mu_j\|$$

for $\hat{d}_{ij}$. We can bound the first term by

$$\begin{aligned}
\Pr[\|\mu_j - \hat{\mu}_j\| + \|\mu_i - \hat{\mu}_i\| \geq (1 - \gamma_4) d_{ij}] \ &\leq \ \Pr[\|\mu_j - \hat{\mu}_j\| \geq \gamma_4 d_{ij}/2] \ + \ \Pr[\|\mu_i - \hat{\mu}_i\| \geq (1 - \gamma_4) d_{ij}/2] \\
&\leq \ \frac{4(\mathrm{Var}(\hat{\mu}_j) + \mathrm{Var}(\hat{\mu}_i))}{(1 - \gamma_4)^2 d_{ij}^2}
\end{aligned}$$

We notice that if $\hat{\beta}_{ij} \leq \gamma_3 d_{ij}$ and $\hat{d}_{ij} > \gamma_4 d_{ij}$, then

$$\begin{aligned}\gamma_3 d_{ij} \;\geq\; \hat{\beta}_{ij} \;&>\; \frac{\gamma_4 d_{ij}}{2} - \frac{2\alpha_{ij}\Delta + \Delta^2}{d_{ij}} \\ &\geq\; \frac{\gamma_4 d_{ij}}{2} - \frac{2\gamma_1 d_{ij}^2 + \gamma_2^2 d_{ij}^2/p}{d_{ij}} \\ &\geq\; d_{ij}(\gamma_4/2 - 2\gamma_1 - \gamma_2^2/p)\end{aligned}$$

which is a contradiction to $\gamma_4 - 2\gamma_3 - 4\gamma_1 - 2\gamma_2^2/s > 0$ for appropriate choices of $\gamma_1, \ldots, \gamma_4 \in (0, 1)$. Hence

$$\Pr[\hat{\beta}_{ij} \leq \gamma_3 d_{ij}] \;\leq\; \frac{4(\mathrm{Var}(\hat{\mu}_j) + \mathrm{Var}(\hat{\mu}_i))}{(1-\gamma_4)^2 d_{ij}^2}$$

**Combining everything together** $(B) \leq (C) + (F) \leq (D) + (H) + (F) \leq (G) + (E) + (F) + (H)$ and using $\mathrm{Var}(\hat{\mu}_i) = \mathrm{Var}(\hat{u}_i)/n$ we finally obtain

$$\Pr[\mathrm{Err}] \;\leq\; \frac{4\mathrm{Var}(\hat{u}_i)}{n\gamma_3^2 d_{ij}^2} + \frac{4\mathrm{Var}(\hat{u}_i)}{s \cdot \gamma_3^2 d_{ij}^2} + \frac{4\gamma_2^2 \mathrm{Var}(\hat{u}_i)}{n\gamma_1^2 \cdot s \cdot d_{ij}^2} + \frac{4\gamma_2^2 \mathrm{Var}(\hat{u}_i)}{\gamma_1^2 \cdot s \cdot d_{ij}^2} + \frac{4(\mathrm{Var}(\hat{u}_j) + \mathrm{Var}(\hat{u}_i))}{n(1-\gamma_4)^2 d_{ij}^2}.$$

Finally, by choosing $\gamma_1 = \gamma_2 = 0.12$, $\gamma_3 = \gamma_4/2.5$ and $\gamma_4 = 0.65$ (satisfying $\gamma_4 - 2\gamma_3 - 4\gamma_1 - 2\gamma_2^2/p > 0$), we obtain

$$\Pr[\mathrm{Err}] \;\leq\; \frac{60\mathrm{Var}(\hat{u}_i)}{nd_{ij}^2} + \frac{60\mathrm{Var}(\hat{u}_i)}{s \cdot d_{ij}^2} + \frac{4\mathrm{Var}(\hat{u}_i)}{nd_{ij}^2} + \frac{4\mathrm{Var}(\hat{u}_i)}{s \cdot d_{ij}^2} + \frac{33(\mathrm{Var}(\hat{u}_j) + \mathrm{Var}(\hat{u}_i))}{nd_{ij}^2}.$$

□

**Lemma B.2.** *Let $\{\tilde{P}_i, \tilde{P}_j\}$ be a pair of class-conditional distributions. Let $f : \mathbb{R}^d \to \mathbb{R}^p$ be any feature map and denote $\mu_c = \mu_f(\tilde{P}_c)$. Let $\Delta < \|\mu_i - \mu_j\|$. Assume that $f \circ \tilde{P}_c$ are all spherical Gaussians with means $\mu_c$ (for all $c \in [l]$). Assume that $V_f^{ij} = \frac{\mathrm{Var}_f(\tilde{P}_i)}{(\|\mu_i - \mu_j\| - \Delta)^2} \leq 16$ for all $i \neq j \in [l]$. Then, we have*

$$L_{ij}^{\Delta}(f) \;\leq\; 3\frac{\exp\left(-p/32V_f^{ij}\right)}{(\mathrm{e} \cdot V_f^{ij})^{p/2}}.$$

*Proof.* Let $i \neq j \in [l]$ be a given pair. Let $d_{ij} := \|\mu_i - \mu_j\|$ be the considered centers' distance, which sets the overall scale. Let $\hat{S}_c \sim \tilde{P}^n$ and $\hat{x}_c \sim \tilde{P}_c$. Let $u_c = f(\hat{x}_c)$, $\mu_c = \mu_f(\tilde{P}_c)$, $\hat{\mu}_c = \mu_f(\hat{S}_c)$ and $\sigma_c^2 = \mathrm{Var}(u_c)$. By the triangle inequality,

$$\begin{aligned}\Pr\left[\|u_i - \hat{\mu}_j\| \leq \|u_i - \hat{\mu}_i\| + \Delta\right] \;&\leq\; \Pr\left[d_{ij} \leq 2\|u_i - \mu_i\| + \|\hat{\mu}_i - \hat{\mu}_i\| + \|\hat{\mu}_j - \mu_j\| + \Delta\right] \\ &\leq\; \Pr\left[(d_{ij} - \Delta)/4 \leq \|u_i - \mu_i\|\right] \\ &\quad + \Pr\left[(d_{ij} - \Delta)/4 \leq \|\hat{\mu}_i - \hat{\mu}_i\|\right] \\ &\quad + \Pr\left[(d_{ij} - \Delta)/4 \leq \|\hat{\mu}_j - \mu_j\|\right].\end{aligned} \tag{22}$$

Since $u_c \sim \mathcal{N}(\mu_c, \sigma_c^2 \cdot \mathbb{I}/p)$ and $\hat{\mu}_c \sim \mathcal{N}(\mu_c, \sigma_c^2 \cdot \mathbb{I}/(np))$, the random variables $E_i = \frac{\|u_i - \mu_i\|^2 p}{\sigma_i^2}$ and $U_i = \frac{\|\hat{\mu}_i - \mu_i\|^2 p}{n\sigma_i^2}$ are chi-squared random variables with $p$ components. We can write

$$\begin{aligned}\Pr\left[\|u_i - \mu_i\| \geq \frac{d_{ij} - \Delta}{4}\right] \;&=\; \Pr\left[E_i \geq \frac{(d_{ij} - \Delta)^2 p}{16\sigma_i^2}\right] \\ &\leq\; 1 - F\left(\frac{(d_{ij} - \Delta)^2 p}{16\sigma_i^2},\; p\right) \\ &=\; 1 - F\left(zp,\; p\right),\end{aligned}$$

where $F(x,p)$ is the CDF of the chi-squared distribution with $p$ components and $z = \left(\frac{d_{ij}-\Delta}{4\sigma_c}\right)^2$. By Chernoff's bound, if $z > 1$, we have

$$1 - F(zp,\ p) \ \leq\ (z\exp(1-z))^{p/2} \ \leq\ \frac{\exp\left(-p/32V_f^{ij}\right)}{(\mathrm{e}\cdot V_f^{ij})^{p/2}} \tag{23}$$

Similarly, we have

$$\Pr\left[\|\hat{\mu}_i - \mu_i\| \geq \frac{d_{ij}-\Delta}{4}\right] \ \leq\ \frac{\exp\left(-np/32V_f^{ij}\right)}{(\mathrm{e}\cdot V_f^{ij})^{p/2}} \ \leq\ \frac{\exp\left(-p/32V_f^{ij}\right)}{(\mathrm{e}\cdot V_f^{ij})^{p/2}}. \tag{24}$$

Finally, by combining (22)-(24) we obtain the desired inequality. □

## C. Proof of Theorem 4.4

**Theorem C.1.** *Let $\delta \in (0,1)$, $l \geq 2$ and let $\mathcal{F}$ be a class of ReLU neural networks of depth $q$. Then, with probability at least $1-\delta$ over the selection of the source class-conditionals $\tilde{\mathcal{P}} = \{\tilde{P}_c\}_{c=1}^l$ and the corresponding source training data $\tilde{S}_1 \ldots, \tilde{S}_l$, for any $f \in \mathcal{F}$ and $\Delta = 0.12 s(f,\tilde{\mathcal{P}})^{-1/2} \cdot \min_{i\neq j\in[l]} \|\mu_f(\tilde{P}_i) - \mu_f(\tilde{P}_j)\|$, we have*

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(f) \ \leq\ & Ck \cdot \left[\tfrac{1}{s(f;\tilde{\mathcal{P}})} + \tfrac{1}{n}\right] \cdot \mathrm{Avg}_{i\neq j\in[l]}\left[(V_f(\tilde{S}_i,\tilde{S}_j) + B_{ij}(\tfrac{\delta}{8l^2}))\cdot(1+A_{ij}(\tfrac{\delta}{8l^2}))^2\right] + \tfrac{k\delta}{2} \\ & + \frac{\sqrt{q}pkB\cdot\hat{\mathcal{C}}(f)}{\Delta\sqrt{l}} \cdot \mathrm{polylog}(k,\delta^{-1},\hat{\mathcal{C}}(f),|\log(\tfrac{\Delta}{B})|,l)\end{aligned}$$

*for some universal constant $C > 0$.*

*Proof.* By Lemma A.1, with probability at least $1-\frac{\delta}{2}$ over the selection of $\tilde{\mathcal{P}}$, for any $f \in \mathcal{F}$ and $\Delta > 0$, we have

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(f) \ \leq\ & (k-1)\,\mathrm{Avg}_{i\neq j\in[l]}[L_{ij}^{2\Delta}(f)] + (k-1)\cdot\tfrac{\delta}{2} \\ & + (k-1)\cdot\frac{(\mathcal{C}(f)+1)\cdot pB}{\Delta\sqrt{l}}\left(512\sqrt{2q\pi\log(l)} + \sqrt{2\log\left(\frac{\sqrt{6}(\mathcal{C}(f)+2)(|\log(\frac{\Delta}{B})|+2)}{\sqrt{\delta}}\right)}\right).\end{aligned} \tag{25}$$

In addition, by Lemma B.1, for any $f$ and for $\Delta = 0.12 s(f,\tilde{\mathcal{P}})^{-1/2} \cdot \min_{i\neq j\in[l]} \|\mu_f(\tilde{P}_i) - \mu_f(\tilde{P}_j)\|$, we have,

$$L_{ij}^{2\Delta}(f) \ \overset{\times}{\leq}\ \left[\frac{1}{s(f;\tilde{P})} + \frac{1}{n}\right]\cdot V_f(\tilde{P}_i,\tilde{P}_j).$$

By Lemma 4.1, for any pair $\tilde{P}_i, \tilde{P}_j$ with probability at least $1-\frac{\delta}{2l^2}$ over the selection of $\tilde{S}_i, \tilde{S}_j$, we have $V_f(\tilde{P}_i,\tilde{P}_j) \leq (V_f(\tilde{S}_i,\tilde{S}_j) + B_{ij}(\frac{\delta}{8l^2}))\cdot(1+A_{ij}(\frac{\delta}{8l^2}))^2$. Therefore, by union bound, with probability at least $1-\frac{\delta}{2}$ over the selection of $\tilde{S}_1,\ldots,\tilde{S}_l$, we have

$$\mathrm{Avg}_{i\neq j\in[l]}[L_{ij}^{2\Delta}(f)] \ \overset{\times}{\leq}\ \left[\frac{1}{s(f;\tilde{P})} + \frac{1}{n}\right]\cdot\mathrm{Avg}_{i\neq j\in[l]}\left[(V_f(\tilde{S}_i,\tilde{S}_j) + B_{ij}(\tfrac{\delta}{8l^2}))\cdot(1+A_{ij}(\tfrac{\delta}{8l^2}))^2\right]. \tag{26}$$

Again by union bound, with probability at least $1-\delta$, over the selection of $\tilde{\mathcal{P}} = \{\tilde{P}_c\}_{c=1}^l$ and $\{\tilde{S}_c\}_{c=1}^l$, we can combine (25) and (26) which gives us the desired bound. □