On the Trade-off of Intra-/Inter-class Diversity for Supervised Pre-training

Jieyu Zhang [*1] **Bohan Wang** [*2] **Zhengyu Hu** [3] **Pang Wei Koh** [1] **Alexander Ratner** [1]

Abstract

Pre-training datasets are critical for building state-of-the-art machine learning models, motivating rigorous study on their impact on downstream tasks. In this work, we study the impact of the trade-off between the intra-class diversity (the number of samples per class) and the inter-class diversity (the number of classes) of a supervised pre-training dataset. Empirically, we found that with the size of the pre-training dataset fixed, the best downstream performance comes with a balance on the intra-/inter-class diversity. To understand the underlying mechanism, we show theoretically that the downstream performance depends monotonically on both types of diversity. Notably, our theory reveals that the optimal class-to-sample ratio ($\frac{\#\text{classes}}{\#\text{samples per class}}$) is invariant to the size of the pre-training dataset, which motivates an application of predicting the optimal number of pre-training classes. We demonstrate the effectiveness of this application by an improvement of around 2 points on the downstream tasks when using ImageNet as the pre-training dataset.

1. Introduction

Many state-of-the-art deep neural network models involve a pre-training stage (Huh et al., 2016; Mahajan et al., 2018; Radford et al., 2021; Alayrac et al., 2022), where the model is typically pre-trained on a dataset different from the downstream tasks. Recently, researchers have identified the pre-training dataset as a key factor contributing to the success of pre-training (Entezari et al., 2023; Hashimoto, 2021; Jain et al., 2023; Gadre et al., 2023; Hong et al., 2023; Zhao et al., 2022). Despite the importance of the pre-training dataset and the countless effort required for data collection, how to design the pre-training dataset still remains underexplored.

In this work, we focus on supervised pre-training, one of the most popular pre-training paradigms, and study two key quantities of a supervised pre-training dataset: intra-class diversity (the number of different samples within each pre-training class) and inter-class diversity (the number of different pre-training classes). Intuitively, both diversities are beneficial for supervised pre-training (Huh et al., 2016). Yet when the size of the pre-training dataset is fixed, there exists a trade-off between both diversities, since one cannot increase both diversities simultaneously and increasing one diversity will decrease the other. We are then curious about the impact of this dataset diversity trade-off on the performance of the pre-trained models on downstream tasks, as well as how to balance both diversities in designing a supervised pre-training dataset for the best downstream performance.

Empirically, with ImageNet (Russakovsky et al., 2015) as the pre-training dataset and the pre-training dataset size fixed, we show that the optimal performance on the downstream tasks occurs when a balance on the intra-/inter-class diversity is achieved, indicating that the dataset diversity trade-off would result in a downstream performance trade-off. We then offer a theoretical explanation for this effect by demonstrating that downstream performance displays a convex relationship with respect to the class-to-sample ratio, i.e.,, the ratio of the number of pre-training classes to the number of samples per class, which can be seen as the ratio between inter-/intra-class diversity. The established analytical relationship between downstream performance and the class-to-sample ratio can serve as a guiding principle in designing a supervised pre-training dataset by estimating the optimal class-to-sample ratio rather than grid search.

Notably, our theory shows that the optimal class-to-sample ratio is invariant to the size of the pre-training dataset. Such an invariance uncovers a direct application: one may estimate the optimal class-to-sample ratio with small pre-training datasets and then leverage it to build a large-scale pre-training dataset. In particular, the optimal number of pre-training classes $\bar{K}$ is proportional to the square root of the size of the pre-training dataset N, i.e.,, $\bar{K} \propto \sqrt{N}$. We then empirically verify our theoretical finding on ImageNet (Russakovsky et al., 2015) and present the effectiveness of its application in predicting optimal number of classes for pre-training datasets with different sizes.

[*]Equal contribution [1]University of Washington, USA [2]University of Science and Technology of China, China [3]Mohamed bin Zayed University of Artificia l Intelligence, UAE. Correspondence to: Jieyu Zhang <jieyuz2@cs.washington.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Our major findings and contributions are as follows:

- In supervised pre-training, we found that when the pre-training dataset size is fixed, the intra-/inter-class diversity trade-off would cause a downstream performance trade-off, highlighting the importance of balancing both diversities in designing the pre-training dataset;
- We provide a theory which explains this effect by showing that the downstream performance displays a convex relationship with respect to the class-to-sample ratio, serving as a guiding principle in designing a supervised pre-training dataset.;
- Our theory also uncovers the invariance of the optimal class-to-sample ratio with respect to the size of the pre-training dataset, motivating an application in predicting the optimal number of classes when building a pre-training dataset for a specific downstream task.

2. Empirical Observations

2.1. Experimental setup

The goal of this work is to study the trade-off of intra-/inter-class diversity in a supervised pre-training dataset and its impact on the pre-trained model's performance on downstream tasks. Specifically, the inter-class diversity refers to the diversity of classes in pre-training dataset, i.e.,, how many different classes we have (K); while the intra-class diversity refers to the diversity of samples within each class, i.e.,, how many different samples in each class (n). When the size of the pre-training dataset is fixed, increasing either type of diversity will by definition decrease the other, leading to a dataset diversity trade-off. To study the impact of such dataset diversity trade-off, we experiment with pre-training datasets with varying number of classes and number of samples per class.

Evaluation protocol. Following common practice (Huh et al., 2016), we use the ImageNet (Russakovsky et al., 2015) as the dataset for supervised pre-training. In this work, we mainly use ResNet-18 (He et al., 2016) as the backbone model. For evaluating the performance of the pre-trained model on downstream tasks, we perform linear probing (tuning the head but freezing the lower layers). We repeat each individual experiment five times and report the averaged top-1 accuracy.

Downstream tasks. We adopt the following six datasets as the downstream classification tasks: Stanford40 dataset (Yao et al., 2011) for action recognition, Stanford-Dogs (Khosla et al., 2011) for fine-grained object recognition, MIT67 (Quattoni & Torralba, 2009) for scene classification, CIFAR10 (Krizhevsky et al., 2009) for image recognition datasets, Flowers102 (Nilsback & Zisserman, 2008) for image classification dataset, FGVCAircraft (Maji et al., 2013) for aircraft classification dataset.

While having all the other configurations fixed, during pre-training, we vary the number of classes and the number of samples per class. Specifically, given N as the size of the pre-training dataset and K as the number of classes, we randomly sample K classes from ImageNet and then uniformly sample $n = \frac{N}{K}$ samples from each class to compose the dataset. We experiment with following N and K values: $\{$1K, 2K, 5K, 10K, 20K, 50K, 100K$\}$ and $\{$2, 5, 10, 20, 50, 100, 200, 500, 1000$\}$ respectively. Note that with large N (e.g.,, $N = 10$K), the K must not be small (e.g.,, $K = 2$) since in ImageNet each class has at most 1300 samples.

2.2. Results and observations

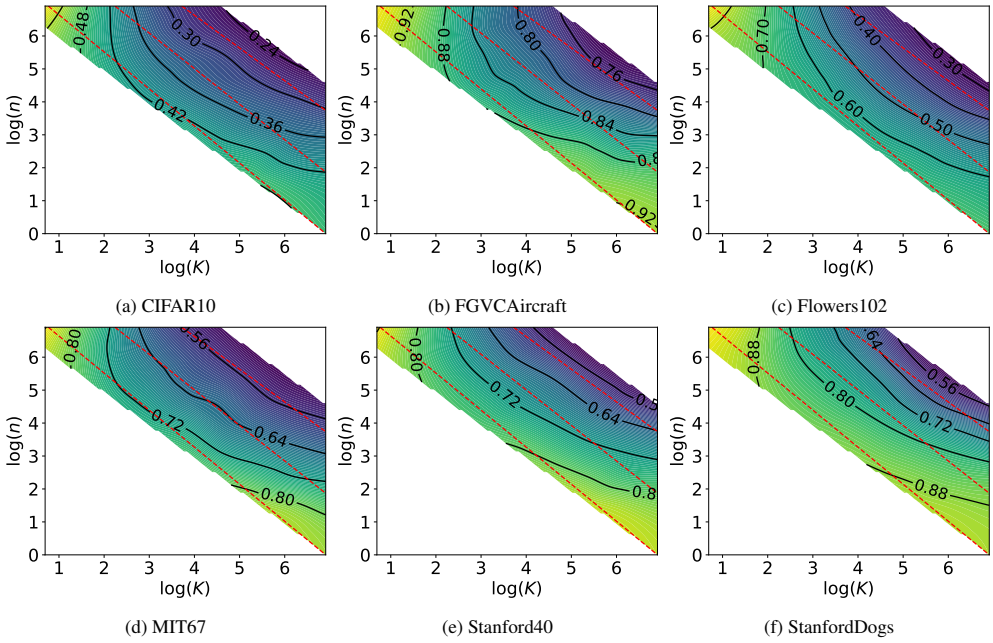
We visualize the results in Fig. 1. In the contour plot, the z-value is the error rate on the test set, thus lower is better. The x-axis and y-axis are inter-class diversity ($\log K$) and intra-class diversity ($\log n$) in the log space, respectively. The values on anti-diagonal lines ($y = -x + c$) share the same pre-training dataset size as $\log N = \log K + \log n$. From the results, we draw the following conclusions:

Both intra-/inter-class diversity are beneficial for downstream tasks. We can see that increasing either inter-class diversity (($\log K$) or intra-class diversity ($\log n$), given the other is fixed, would lead to a better test error rate. This is intuitive and as expected, since the size of the pre-training dataset N would increase accordingly, which is known to be beneficial for downstream tasks (Hernandez et al., 2021; Entezari et al., 2023; Huh et al., 2016).

A trade-off of intra-/inter-class diversity on downstream task performance. More importantly, by looking at the anti-diagonal lines where $\log N$ is fixed and equals $\log K + \log n$, we can see a trade-off between intra-/inter-class diversity on the test error rate of downstream tasks: either cases of 1) high inter-class diversity, low intra-class diversity and 2) low inter-class diversity and high intra-class diversity would not render the best performance. Instead, some point in the middle of the anti-diagonal line leads to the lowest test error rate.

3. Theoretical Understanding

In this section, we first present the theoretical setup and notations and then provide a theory on the impact of the pre-training dataset diversity on the downstream performance. We also show that the optimal class-to-sample ratio ($\frac{K}{n}$) is invariant to the size of the pre-training dataset, motivating an application of predicting the optimal number of pre-training



(a) CIFAR10 (b) FGVCAircraft (c) Flowers102

(d) MIT67 (e) Stanford40 (f) StanfordDogs

Figure 1. Test error rate (the darker color is better) across inter-class diversity ($\log(K)$) and intra-class diversity ($\log(n)$). The values on the red dashed anti-diagonal lines share the same pre-training dataset size ($\log(K) + \log(n) = \log(N)$ is constant), from which we can see that with fixed pre-training dataset size, the best downstream performance requires a balance on both diversities.

classes. Lastly, we present a contrastive scenario where the trade-off no longer exists and K dominates the downstream performance.

3.1. Setup and notations

Dataset . To be consistent with our experimental setup, we consider the supervised pre-training task. Specifically, we can access two datasets, one for the pre-training task (denoted as S^p) and another for the downstream task (denoted as S^d). Each example in the pre-training dataset consists of a data $x \in \mathcal{X} = \mathbb{R}^{d_1}$ (where d_1 is the dimension of data) and a label $y \in [K]$ (where K is the number of classes). Specifically, we denote $S^p = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, where N is the size of S^p, and assumes that S^p is sampled according to some underlying distribution $\mathcal{P}$. Similarly, every example in the downstream dataset consists of a data $\tilde{x} \in \mathcal{X}$ and a label $\tilde{y} \in [\tilde{K}]$ (note that $\tilde{K}$ does not necessarily equal to K), and is sampled according to another underlying distribution $\tilde{P}$. We denote $S^d = \{(\tilde{x}_1, \tilde{y}_1), \cdots, (\tilde{x}_{\tilde{N}}, \tilde{y}_{\tilde{N}})\}$.

Model. The models for both pre-training and downstream tasks consist of two components: the feature extractor and the linear classifier. Specifically, the model for the pre-training task is given as $f^p \circ h$, where $f^p : \mathbb{R}^{d_2} \to \mathbb{R}^K$ is the pre-training classifier (d_2 is the dimension of feature) and $h : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is the feature extractor. We denote the set of all possible f^p as $\mathcal{F}$, and the set of all possible h as $\mathcal{H}$. The model for the downstream task is given as $f^d \circ h$, where $f^d : \mathbb{R}^{d_2} \to \mathbb{R}^{\tilde{K}}$ is the downstream classifier and h is the feature extractor shared with the pre-training task. We set all possible f^p as $\tilde{\mathcal{F}}$.

Loss. To measure the correctness of model's predictions on the pre-training task, we use cross-entropy loss $\ell : \mathbb{R}^K \times [K] \to \mathbb{R}^{\geq 0}$. Specifically, given an example (x, y) and pre-training model $f^p \circ h$, the corresponding cross-entropy loss is defined as

$$\ell(f^p \circ h(x), y) = -\log \frac{e^{f^p_y \circ h(x)}}{\sum_{i=1}^K e^{f^p_i \circ h(x)}},$$

where $f^p_i \circ h(x)$ is the i-th coordinate of $f^p \circ h(x)$. We make the following assumption over the complexity of the model.

Assumption 3.1. There exist a positive constant M_ℓ, such that $\forall i \in [K], \forall f \in \mathcal{F}, h \in \mathcal{H}, x \in \mathcal{X}$,

$$\ell(f \circ h(x), i) \leq M_\ell.$$

Furthermore, define the gaussian complexity of function class $\mathcal{F} \circ \mathcal{H}$ over the marginal distribution $\mathcal{P}_x$ with sample number N as

$$G_N^{\mathcal{P}_x}(\mathcal{F} \circ \mathcal{H}) \triangleq \mathbb{E}_{(x_i)_{i=1}^N \sim \mathcal{P}_x^n} \mathbb{E}_{(\sigma_{i,j})_{i \in [N], j \in K} \sim \mathcal{N}(0, 1_{n \times K})} \sup_{f \in \mathcal{F}, h \in \mathcal{H}} \sum_{i=1}^N \sum_{j=1}^K \sigma_{i,j} f_j \circ h(x_i).$$

We assume that $G_N^{\mathcal{P}_x}(\mathcal{F}\circ\mathcal{H}) \le G\sqrt{N}$, where G is independent of $\mathcal{P}$ and N. [1]

Assumption 3.1 constrains the dependence of model complexity over sample size n, which is a standard assumption in generalization analysis (Mohri et al., 2018).

Optimal risk and solution. We first define the optimal feature extractor over the pre-training task as

$$\hat{h}_S \triangleq \arg\min_{h\in\mathcal{H}}\left(\min_{f\in\mathcal{F}}\sum_{i=1}^{N}\ell(f\circ h(x_i),y_i)\right).$$

We measure the performance of the extractor through its risk. Specifically, Given a feature extractor $h\in\mathcal{H}$, the risk of it over $\mathcal{P}$ is measured by

$$\mathcal{R}_p(h,\mathcal{P}) \triangleq \min_{f\in\mathcal{F}}\mathbb{E}_{S^p\sim\mathcal{P}}\frac{1}{N}\sum_{i=1}^{N}[\ell(f\circ h(x_i),y_i)].$$

The representation error of h over $\mathcal{P}$ is then defined as the gap between the risk of h and the possible smallest risk

$$\mathcal{E}_p(h,\mathcal{P}) \triangleq \mathcal{R}_p(h,\mathcal{P}) - \min_{\tilde{h}\in\mathcal{H}}\mathcal{R}_p(\tilde{h},\mathcal{P}).$$

Similarly, the risk and error over the downstream task are defined as

$$\mathcal{R}_d(h) \triangleq \min_{f\in\tilde{\mathcal{F}}}\mathbb{E}_{S^d\sim\tilde{\mathcal{P}}}\sum_{i=1}^{\tilde{N}}\ell(f\circ h(\tilde{x}_i),\tilde{y}_i),$$
$$\mathcal{E}_d(h) \triangleq \mathcal{R}_d(h) - \min_{\tilde{h}\in\mathcal{H}}\mathcal{R}_d(\tilde{h}).$$

3.2. The theory behind the trade-off of intra-/inter-class diversity

We start with a detailed characterization of the data generation process followed by assumptions and the main result.

Data generation process. We assume the pre-training data is generated through a two-step sampling process. Specifically, suppose that there is a distribution $\mathcal{D}$ over $\Delta(\mathcal{X})$ (the set consisting of all distributions on $\mathcal{X}$). Then, S^p is generated by the following procedure (and $\mathcal{P}$ is naturally induced)

- Sample K classes by i.i.d. sampling K distributions $\{\mathcal{P}_i\}_{i=1}^K$ according to $\mathcal{D}$. These are respectively the underlying distributions of K classes;
- For each $i\in[K]$, i.i.d. sample n data $\{x_{i,1},\cdots,x_{i,n}\}$ according to P_i and denote $S_i = \{(x_{i,1},i),\cdots,(x_{i,n},i)\}$. Note here $x_{i,j}$ does not contain the information of label, as its label information is already contained in i. The whole dataset is obtained by putting all S_i together, i.e., $S^p=\{S_1,\cdots,S_K\}$.

[1] This inequality holds for a wide range of models, including deep neural networks (Bartlett et al., 2017).

We make the following assumption on the correlation between the representation powers of the pre-training and the downstream task.

Assumption 3.2. Given $\mathcal{P}$, there exists non-negative constants $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$, such that $\forall h\in\mathcal{H}$,

$$\mathcal{E}_d(h) \le \nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\mathcal{E}_p(h,\mathcal{P}) + \nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}).$$

We further assume that $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$ are stable, that is, there exist two $\tilde{\mathcal{P}}$-dependent positive constants $M_0^{\tilde{\mathcal{P}}}$ and $M_1^{\tilde{\mathcal{P}}}$, for $\mathcal{P}=\Pi_{i=1}^K(\mathcal{P}_i,i)$ and $\mathcal{P}=\Pi_{i=1}^{K-1}(\mathcal{P}_i,i)\times(\mathcal{P}'_K,K)$ which differ by only one component, $|\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}) - \nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}')| \le \frac{M_0^{\tilde{\mathcal{P}}}}{K}$ and $|\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}) - \nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}')| \le \frac{M_1^{\tilde{\mathcal{P}}}}{K}$. Moreover, we assume that $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$ obeys concentration inequalities, i.e., there exist $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{D})$, $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{D})$, $C_0^{\tilde{\mathcal{P}}}$ and $C_1^{\tilde{\mathcal{P}}}$, such that $|\mathbb{E}_{\{\mathcal{P}_i\}_{i=1}^K\sim\mathcal{D}^K}\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}) - \nu_0^{\tilde{\mathcal{P}}}(\mathcal{D})| \le \frac{C_0^{\tilde{\mathcal{P}}}}{\sqrt{K}}$ and $|\mathbb{E}_{\{\mathcal{P}_i\}_{i=1}^K\sim\mathcal{D}^K}\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}) - \nu_1^{\tilde{\mathcal{P}}}(\mathcal{D})| \le \frac{C_1^{\tilde{\mathcal{P}}}}{\sqrt{K}}$.

Assumption 3.2 assumes that the pre-training representation error can bound the downstream representation error, which is a common assumption in existing works (Du et al., 2021; Zhao et al., 2022; Chen et al.). Also, as $\mathcal{P}$ is derived by sampling K distributions according to $\mathcal{D}$, we make mild assumptions that the coefficients $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$ is robust when changing the underlying distribution of only one class, and when K grows, the expectation of $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$ converge to some limits. These assumption holds when $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$ have some inner structure, for example, there exists $A^{\tilde{\mathcal{P}}}$, such that $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}) = \frac{1}{K(K-1)}\sum_{i\neq j}A^{\tilde{\mathcal{P}}}(\mathcal{P}_i,\mathcal{P}_j)$.

Theorem 3.3. *Let Assumptions 3.1 and 3.2 hold. Then, with probability at least* $1-\delta$,

$$\mathcal{E}_d(\hat{h}_S) \le \left(\nu_1^{\tilde{\mathcal{P}}}(\mathcal{D}) + M_1\sqrt{\frac{\log\frac{4}{\delta}}{2K}} + \frac{C_1}{\sqrt{K}}\right) \tag{1}$$
$$\times\left(5M_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + \frac{2GL_\ell}{\sqrt{n}}\right)$$
$$+\nu_0^{\tilde{\mathcal{P}}}(\mathcal{D}) + M_0\sqrt{\frac{\log\frac{4}{\delta}}{2K}} + \frac{C_0}{\sqrt{K}}. \tag{2}$$

Simplifying the Theorem 3.3. Denote the RHS of Equation 2 as U, we have

$$\begin{aligned} U =& \nu_1^{\tilde{\mathcal{P}}}(\mathcal{D}) \left(5M_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2}} + 2GL_\ell\right)\frac{1}{\sqrt{n}} \\ &+ \left(M_0\sqrt{\frac{\log\frac{4}{\delta}}{2}} + C_0\right)\frac{1}{\sqrt{K}} \\ &+ \left(M_1\sqrt{\frac{\log\frac{4}{\delta}}{2}} + C_1\right) \\ &\times \left(5M_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2}} + 2G\right)\frac{1}{\sqrt{N}} \\ &+ \nu_0^{\tilde{\mathcal{P}}}(\mathcal{D}). \end{aligned}$$

We can see that the above equation can be simplified as

$$U = \frac{A}{\sqrt{n}} + \frac{B}{\sqrt{K}} + \frac{C}{\sqrt{N}} + D, \tag{3}$$

where A, B, C, D do not depend on N, K, n, but instead depend on the downstream task since $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{D})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{D})$ depend on the downstream task.

Explaining downstream performance trade-off given a fixed N. From Equation 3 we can see that the performance on target task would increase when we increase 1) intra-class diversity n, 2) inter-class diversity K, and 3) the size of pre-training dataset N. When N is fixed, however, increasing either intra-class diversity or inter-class diversity would decrease the other (since $N = n \times K$) and therefore eventually leads to performance drop. Another way to see this is to parameterize U as a function of K without n:

$$U(K) = \frac{A\sqrt{K}}{\sqrt{N}} + \frac{B}{\sqrt{K}} + \frac{C}{\sqrt{N}} + D, \tag{4}$$

From this we can clearly see that both extremes of K (too large or too small) would not lead to optimal performance. Similar conclusion can be drawn regarding n when parametrizing U as a function of n without K.

3.3. Balancing intra-/inter-class diversity: the optimal class-to-sample ratio

When N is fixed, by leveraging the fact that $N = n \times K$, we can express U as

$$U = \frac{1}{N^{\frac{1}{4}}}\left(Ax^{\frac{1}{4}} + B\frac{1}{x^{\frac{1}{4}}}\right) + c, \tag{5}$$

where $c = \frac{C}{\sqrt{N}} + D$ is a constant and $x = \frac{K}{n}$ is the class-to-sample ratio. To minimize U, we have the optimal class-to-sample ratio $\bar{x} = \frac{B^2}{A^2}$. Notably, both A and B have no dependency on N. This indicates, **the optimal class-to-sample ratio for a specific downstream task is invariant to the size of the pre-training dataset.** Motivated by this, one could estimate the optimal class-to-sample ratio using a small N and then use it to predict the optimal number of classes for building a large pre-training dataset. In particular, given the optimal class-to-sample ratio $\bar{x}$, the optimal number of classes $\bar{K} = \frac{B}{A}\sqrt{N}$. And based on Equation 5, one only needs three ¡class-to-sample ratio, performance¿ tuples to estimate the constant A, B, and c with a fixed N for computing optimal class-to-sample ratio.

3.4. When there is no trade-off: a contrastive case

To further understand the performance trade-off related to the intra-/inter-class diversity, we show a contrastive case of data generation process, with which the trade-off no longer exists and, instead, K dominates the downstream performance. Concretely, the data generation process is given as follows:

Data generation process. The pre-training dataset is generated by first i.i.d. sampling data $\{x_1, x_2, \cdots, x_N\}$ according some distribution $\mathcal{P}_\mathcal{X} \in \Delta(\mathcal{X})$. We then obtain the label $y_i \in [K]$ of each x_i by performing clustering. The final pre-training dataset is given as $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ (and $\mathcal{P}$ is naturally induced).

The new data generation process is different from the one in Section 3.2. In particular, with different K, the new data generation process would not introduce new data samples as we manipulate the label by clustering the current sampled data, while for the data generation process in Section 3.2, the sampled data change according to the classes we picked.

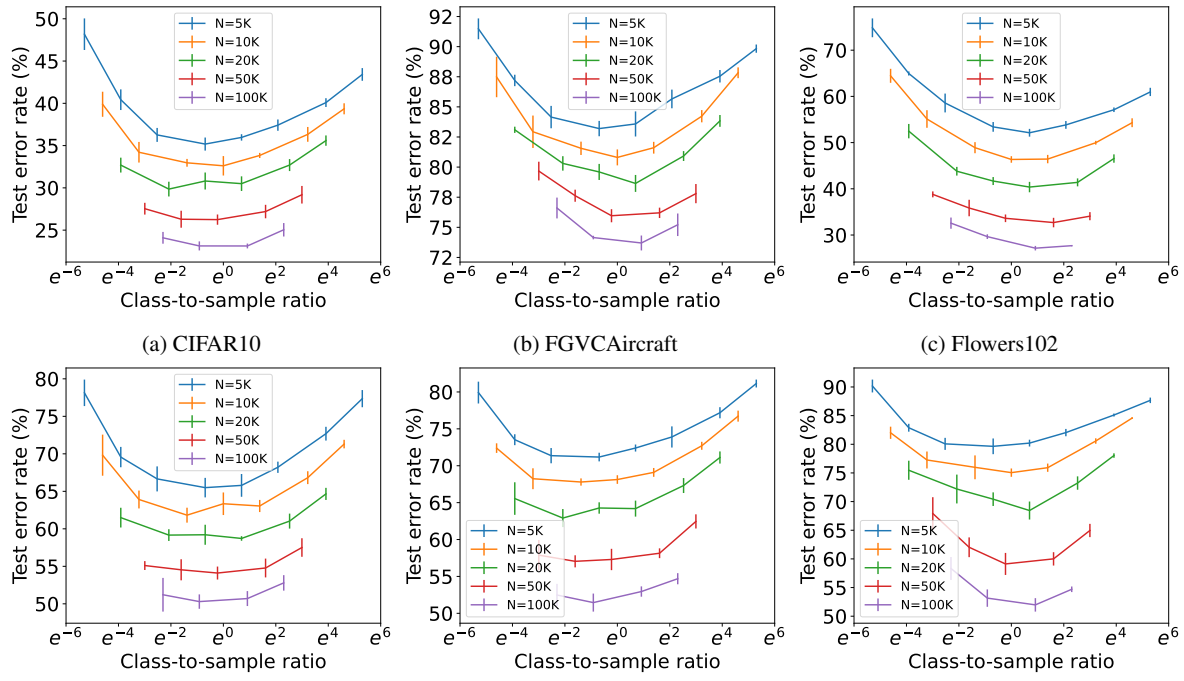
As an analogy to Assumption 3.2, we make the following assumption on the correlation between the representation powers of the pre-training and the downstream task, and then present the theorem.

Assumption 3.4. We assume the pre-training representation error can bound the downstream representation error. Specifically, there exists non-negative constants $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$, such that $\forall h \in \mathcal{H}$,

$$\mathcal{E}_d(h, \mathcal{P}) \le \nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\mathcal{E}_p(h) + \nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}).$$

We further assume that $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})$ and $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})$ obey concentration inequalities, i.e., there exist $\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}_\mathcal{X})$, $\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}_\mathcal{X})$, $C_0^{\tilde{\mathcal{P}}}$ and $C_1^{\tilde{\mathcal{P}}}$, such that $|\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}) - \nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}_\mathcal{X})| \le \frac{C_0^{\tilde{\mathcal{P}}}}{\sqrt{K}}$ and $|\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}) - \nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}_\mathcal{X})| \le \frac{C_1^{\tilde{\mathcal{P}}}}{\sqrt{K}}$.

Theorem 3.5. *Let Assumptions 3.1 and 3.4 hold. Then, with*



(a) CIFAR10 (b) FGVCAircraft (c) Flowers102

(d) MIT67 (e) Stanford40 (f) StanfordDogs

Figure 2. Test error rate across class-to-sample ratio. The vertical bar is standard deviation.

probability at least $1-\delta$,

$$\mathcal{E}_d(\hat{h}_S) \leq \left(\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}_\mathcal{X}) + \frac{C_0^{\tilde{\mathcal{P}}}}{\sqrt{K}}\right) \times \left(5M_\ell\sqrt{\frac{\log\frac{2}{\delta}}{2N}} + \frac{2GL_\ell}{\sqrt{N}}\right) + \nu_1^{\tilde{\mathcal{P}}}(\mathcal{P}_\mathcal{X}) + \frac{C_1^{\tilde{\mathcal{P}}}}{\sqrt{K}}.$$

According to Theorem 3.5, we can see that there is no longer a trade-off between the inter-class diversity K and the intra-class diversity n. Instead, the bound gets smaller when K is larger. To verify this, we conduct experiments under this setting (Figure 3 in the Appendix A.3), and observe a tendency of lower error rate with increasing K.

Discussion. The classes and data samples correspond to two dimensions of diversity. When the size of the pre-training dataset is fixed, in the case of Theorem 3.3, both diversities would vary in a see-saw-like way with different K, but for Theorem 3.5, since we fix the data samples used, only one dimension of diversity varies when K is different and therefore the downstream performance only depends on the varying diversity (K). This difference reveals the underlying cause of the downstream performance trade-off rather than relying solely on K and n values.

4. Justification and Application

In this section, we first verify our theoretical findings via empirical results. Then, we show the effectiveness of using the estimated optimal class-to-sample ratio to build the pre-training datasets.

4.1. Are the trade-off curves for different N aligned?

According to our findings in Section 3.3, the performance on downstream task is a convex function with respect to the class-to-sample ratio (Equation 5) and the optimal class-to-sample ratio $\bar{x} = \frac{B^2}{A^2}$ is invariant to the size of pre-training dataset N. To empirically verify this, we visualize the performance on downstream tasks as a function of the class-to-sample ratio with different N in Figure 2.

From the figures, we can see that the curves of different N for a specific downstream task are aligned, as well as the optimal class-to-ratios, which follows our theoretical findings. This indicates that empirically, one could extrapolate the optimal class-to-ratio estimated with a small N for building a large-scale pre-training dataset. Note that the rightmost point of each curve corresponds to using all the classes in ImageNet, i.e.,, $K = 1000$, and we can see that such a standard design choice does not lead to optimal downstream performance, especially with small pre-training datasets.

Table 1. Test error rate on downstream task (first row of each task, lower is better), and number of classes in pre-training dataset (second row). Underlined values are estimated performance rather than actual empirical performance.

N	Method	Target Dataset					
		CIFAR10	**FGVCAircraft**	**Flowers102**	**MIT67**	**Stanford40**	**StanfordDogs**
50K	Standard (K=1000)	29.19±0.14	77.80±0.28	34.08±0.35	57.51±0.18	62.45±0.36	64.96±0.38
	Grid Search	26.24±0.44	75.96±0.43	32.70±0.21	54.10±0.4	57.05±0.44	59.12±0.2
		(200)	(200)	(500)	(200)	(100)	(200)
	Theo-Optimal	26.13±0.1	75.95±0.4	32.96±0.48	53.99±0.43	56.72±0.16	58.87±0.35
		(169)	(293)	(415)	(161)	(138)	(260)
	Emp-Optimal	26.25±0.47	76.00±0.44	32.13±0.10	53.60±0.39	57.10±0.3	59.76 ±0.26
		(169)	(293)	(415)	(161)	(138)	(260)
	Extrapolation	26.27±0.21	76.18±0.14	32.60±0.06	53.01±0.23	57.25±0.14	60.15±0.27
		(190)	(168)	(296)	(163)	(134)	(158)
100K	Standard (K=1000)	25.04±0.26	75.21±0.48	27.69±0.45	52.79±0.36	54.69±0.39	54.70±0.45
	Grid Search	23.13±0.06	73.70±0.50	27.15±0.44	50.30±0.46	51.45±0.33	51.98±0.30
		(500)	(500)	(500)	(200)	(200)	(500)
	Theo-Optimal	22.84±0.40	73.52±0.41	27.33±0.40	50.15±0.43	51.74±0.36	51.69±0.23
		(276)	(372)	(655)	(249)	(207)	(392)
	Emp-Optimal	22.67±0.33	73.45±0.33	26.67±0.33	50.82±0.33	52.24±0.33	52.24±0.33
		(276)	(372)	(655)	(249)	(207)	(392)
	Extrapolation	23.12±0.13	73.30±0.17	26.98±0.41	50.42±0.25	52.32±0.06	53.17±0.26
		(269)	(238)	(418)	(231)	(190)	(233)

4.2. Predicting the optimal number of pre-training classes

As a direct application of our theoretical and empirical findings, one could estimate the optimal class-to-sample ratio with a small N and use it to decide the optimal number of classes when building a larger pre-training dataset. In particular, denoted by $\bar{x}$ the optimal class-to-sample ratio, the optimal number of classes for a given N is $\bar{K} = \sqrt{\bar{x}N}$. We refer to this approach as Extrapolation. In this section, we empirically compare it against the following methods of deciding the number of classes when building a pre-training dataset:

- **Standard:** the number of classes equals to 1000 as the standard design choice of ImageNet;
- **Grid Search:** the number of classes corresponding to the data point with the lowest error rate for each curve in Figure 2;
- **Theo-Optimal:** given the target size, we use the corresponding data points in Figure 2 to fit the theoretically-derived performance function (Equation 5), and then analytically calculate the optimal number of classes and performance on downstream tasks based on the fitted function;
- **Emp-Optimal:** the performance given by the Theo-Optimal is an estimated value instead of the actual performance, to remedy this, we use the number of classes given by Theo-Optimal to build a pre-training dataset and measure the performance of model trained with it.

We set the target size of pre-training dataset as $\{$50K, 100K$\}$ and round the number of classes to integer if needed. For our Extrapolation method, we only use three data points with N being much smaller than the target size to estimate the optimal class-to-sample ratio: $N = 5000$ and $K = \{10, 50, 200\}$. And we use the estimated optimal class-to-sample ratio for both target sizes. The results as well as the number of classes used by involved methods can be found in Table 1.

From the Table 1, we have several observations. First, although the ImageNet dataset is widely used, its number of classes ($K = 1000$) is not optimal for building a pre-training dataset of 50K/100K samples, since the Standard underperforms other methods. Second, the performance rendered by Emp-Optimal is close to that of Theo-Optimal, indicating that the performance function we derived (Equation 5) could faithfully predict the performance given the class-to-sample ratio. Third, methods except for the Standard all render similar test error rate even though their number of classes are different, which reveals that the performance is not sensitive to the number of classes as long as we pick a reasonable number. Thus, our Extrapolation method is superior to Grid Search and Emp-Optimal, since it needs much fewer samples to estimate the number of classes, while both Grid Search and Emp-Optimal require building the pre-training dataset of target size multiple times.

Table 2. Total number of samples used for building the pre-training dataset.

N	Method	Target Dataset					
		CIFAR10	FGVCAircraft	Flowers102	MIT67	Stanford40	StanfordDogs
50K	Standard (K=1000)	50K					
	Grid Search	150K (5 trials)					
	Emp-Optimal	158.164K	161.570K	159.403K	158.694K	159.268K	160.538K
	Extrapolation	55.418K	55.183K	55.830K	55.117K	54.598K	55.045K
100K	Standard (K=1000)	100K					
	Grid Search	260K (4 trials)					
	Emp-Optimal	272.336K	271.836K	268.164K	269.878K	261.981K	270.579K
	Extrapolation	103.937K	103.608K	104.517K	103.515K	103.049K	103.401K

4.3. Extra data needed for estimating the optimal class-to-sample ratio

One advantage of the Standard method is that it does not require extra data for estimating the optimal number of classes. In contrast, other methods would introduce extra data unused in the final pre-training dataset. For example, when estimating the optimal class-to-sample ratio, one may sample data from 1000 classes but eventually find the optimal number of classes is 600, then the data of the additional 400 classes would not be used in the final pre-training dataset. We are curious about how much data needed by different methods to build the final pre-training dataset. We list the total number of used samples in Table 2.

From the table, we can see that Grid Search and Emp-Optimal require more samples than the target size of the pre-training dataset, since they involve building the pre-training dataset of the target size multiple times. In contrast, the Extrapolation estimates the optimal number of classes using a small N of 5000. In addition, using Extrapolation, one only needs to estimate the optimal class-to-sample ratio once and then use it for building pre-training dataset with different size without re-estimation. Finally, the number of extra data needed by Extrapolation is relatively small compared to the target size of the pre-training dataset. As Extrapolation uses extra data for estimating the optimal class-to-sample ratio compared to Standard, to render a fair comparison, we compare it against Standard with N larger than the number of data used by Extrapolation (Appendix A.4) and still observe improvement in most cases.

5. Related Work

In this section, we briefly review recent studies on supervised pre-training from data-centric perspectives. First, on the composition of the pre-training dataset, Hashimoto (2021) presents a scaling law that predicts the test loss on downstream tasks under varying source dataset compositions, while Jain et al. (2023) studies the performance on downstream tasks when subsets of the pre-training dataset are removed. Second, on the label space of supervised pre-training, Zhao et al. (2022) offers a statistical analysis explaining pre-training techniques' success in NLP, showing that class diversity in pre-training tasks substantially enhances sample efficiency in downstream tasks, while the study by Hong et al. (2023) explores the impact of pre-training label granularity on downstream tasks, emphasizing the importance of selecting an appropriate level of label granularity. Lastly, Entezari et al. (2023) explores the impact of pre-training data distribution on transfer performance, finding the choice of the pre-training dataset to be crucial. In contrast, we dive into the trade-off of the intra-/inter-class diversity in the supervised pre-training dataset.

The study related the most to ours is (Huh et al., 2016), where the authors empirically examined the importance of pre-training data characteristics on downstream performance. While covering a wide range of pre-training data characteristics, this study takes a limited step to explore the trade-off of intra-/inter-class diversity in the pre-training dataset (Section 5.5 in (Huh et al., 2016)). Specifically, the authors only considered two different cases of intra-/inter-class diversity, i.e.,, $K = \{500, 1000\}$, which can hardly unfold the full picture of the impact of the intra-/inter-class diversity on the downstream performance to draw convincing conclusions. In contrast, we provide an in-depth study on the trade-off of intra-/inter-class diversity in the pre-training dataset. Both empirically and theoretically, we show that such a trade-off would transfer to the downstream performance and our theory uncovers a surprising property of the optimal class-to-sample ratio: it is invariant to the size of the pre-training dataset.

6. Conclusion

In this study, we explore the trade-off of the intra-/inter-class diversity in supervised pre-training datasets of fixed size. We discovered that the optimal downstream performance is achieved through a balance of intra-/inter-class diversity. Our theory demonstrates that downstream performance depends on both diversities, and the optimal class-to-sample ratio remains constant regardless of the dataset size. We apply this finding to predict the optimal number of classes in pre-training datasets and provide evidence of its effectiveness using ImageNet.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 2022.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Chen, S., Crammer, K., He, H., Roth, D., and Su, W. J. Weighted training for cross-task learning. In *International Conference on Learning Representations*.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

Entezari, R., Wortsman, M., Saukh, O., Shariatnia, M. M., Sedghi, H., and Schmidt, L. The role of pre-training data in transfer learning, 2023. URL https://openreview.net/forum?id=q_PkAzGFrmq.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Hashimoto, T. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hernandez, D., Kaplan, J., Henighan, T. J., and McCandlish, S. Scaling laws for transfer. *ArXiv*, abs/2102.01293, 2021.

Hong, G. Z., Cui, Y., Fuxman, A., Chan, S. H., and Luo, E. Towards understanding the effect of pretraining label granularity. *arXiv preprint arXiv:2303.16887*, 2023.

Huh, M., Agrawal, P., and Efros, A. A. What makes imagenet good for transfer learning? *ArXiv*, abs/1608.08614, 2016.

Jain, S., Salman, H., Khaddaj, A., Wong, E., Park, S. M., and Madry, A. A data-based perspective on transfer learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Mahajan, D., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pp. 185–201. Springer, 2018. doi: 10.1007/978-3-030-01216-8_12. URL https://doi.org/10.1007/978-3-030-01216-8_12.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. Technical report, 2013.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Quattoni, A. and Torralba, A. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 413–420. IEEE, 2009.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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. doi: 10.1007/s11263-015-0816-y.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, pp. 1331–1338. IEEE, 2011.

Zhao, Y., Chen, J., and Du, S. S. Blessing of class diversity in pre-training. *International Conference on Artificial Intelligence and Statistics*, 2022.

A. Appendix

A.1. Limitation and Potential Negative Social Impact

This work provides valuable insights into the construction of pre-training datasets for machine learning models. However, some potential negative societal impacts might arise from this research, such as:

Bias Amplification: When the research discusses the optimal class-to-sample ratio in pre-training datasets, it does not discuss how these classes are determined. There's a risk that the choice of classes and the samples within these classes could reflect and perpetuate existing biases in society. For instance, if the classes are determined by stereotypical or biased criteria, models trained on these datasets could amplify these biases in their predictions or recommendations.

Overemphasis on Quantity over Quality: This research might also create an overemphasis on the quantity (size and diversity) of the data at the expense of its quality. Poor data quality could lead to the development of inaccurate or unreliable machine learning models.

It's important to note that these potential negative impacts are not necessarily inherent to the research itself, but rather potential outcomes if the research is applied without careful consideration of these societal issues.

A.2. Experimental Details

A.2.1. Training Details

We build our code on Python and Pytorch. We fix the model to be the ResNet-18 (He et al., 2016). For pre-training, we set the number of epochs to be 100 and the batch size to be 64. We use the Adam optimizer for training with a learning rate of 0.1, a momentum of 0.9, and a weight decay of 1e-4. We repeat each experiment 3 times with different seeds and report the mean and variance of the results. All experiments ran on a machine with an Intel(R) Xeon(R) CPU E5-2678 v3 with a 512G memory and two 48G NVIDIA RTX A6000 GPUs.

A.2.2. Details of Downstream Tasks

Stanford Actions 40 (Yao et al., 2011). It contains images of humans performing 40 actions. There are about 180-300 images per class. We do not use bounding box and other annotation information for training. There are a total of 9,532 images, making it the smallest dataset in our benchmark experiments.

Stanford Dogs 120 (Khosla et al., 2011). It contains images of 120 breeds of dogs worldwide. There are precisely 100 examples per category in the training set. It is used for the task of fine-grained image categorization. We do not use the bounding box annotations. There are a total of 20,580 images.

MIT Indoors 67 (Quattoni & Torralba, 2009). It is a scene classification dataset containing 67 indoor scene categories, each consisting of 80 images for training and 20 for testing. Indoor scene recognition is challenging because spatial properties, background information, and object characters are expected to be extracted. There are 15,620 images in total.

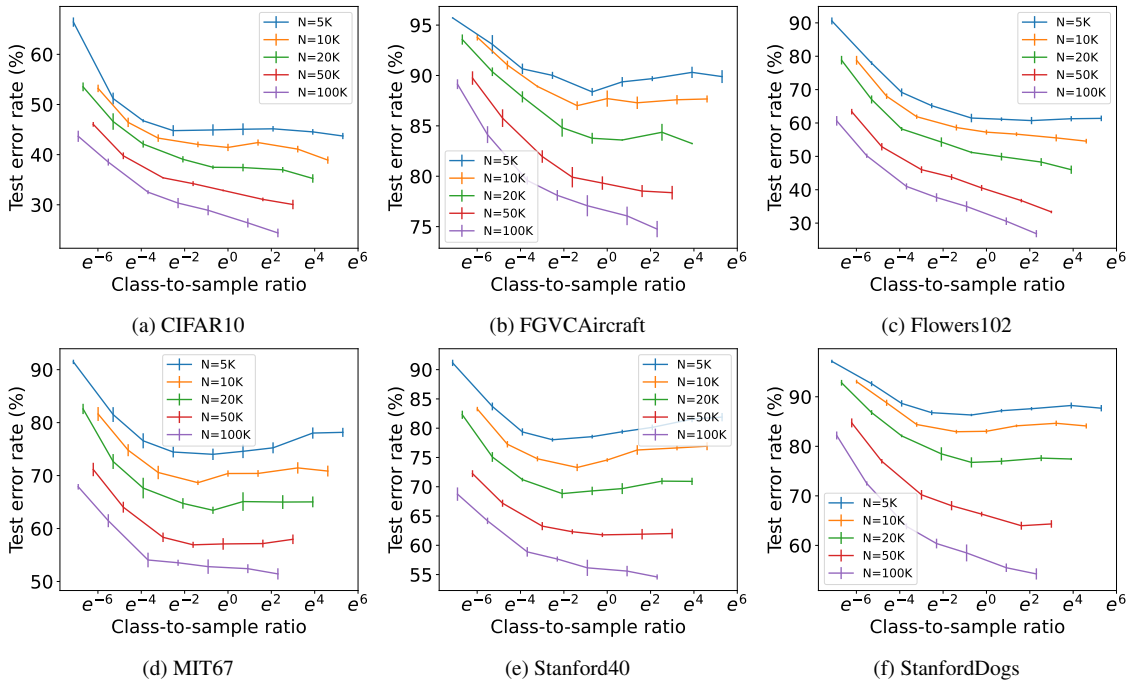
CIFAR10 (Krizhevsky et al., 2009). It is a collection of images commonly used to train machine learning and computer vision algorithms. It contains 60,000 32x32 color images in 10 different classes. The ten classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. There are 6,000 images of each class.

Flowers102 (Nilsback & Zisserman, 2008). It is an image classification dataset consisting of 102 flower categories. The flowers are chosen to be flowers commonly occurring in UK. Each class consists of between 40 and 258 images. The images have large scale, pose and light variations. In addition, some categories have significant variations within the category and several very similar categories.

FGVCAircraft (Maji et al., 2013). It contains 10,200 images of aircraft, with 100 images for each of 102 different aircraft model variants, most of which are airplanes. Each image's (main) aircraft is annotated with a tight bounding box and a hierarchical airplane model label. Aircraft models are organized in a four-level hierarchy.

A.3. Empirical Verification of Theorem 3.5

As shown in Figure 3, we can find that, across all datasets, the results generally showed a decreasing trend in test error rate as the class-to-sample ratio increased. This empirically supports the theoretical assertion that there is no downstream



(a) CIFAR10 (b) FGVCAircraft (c) Flowers102

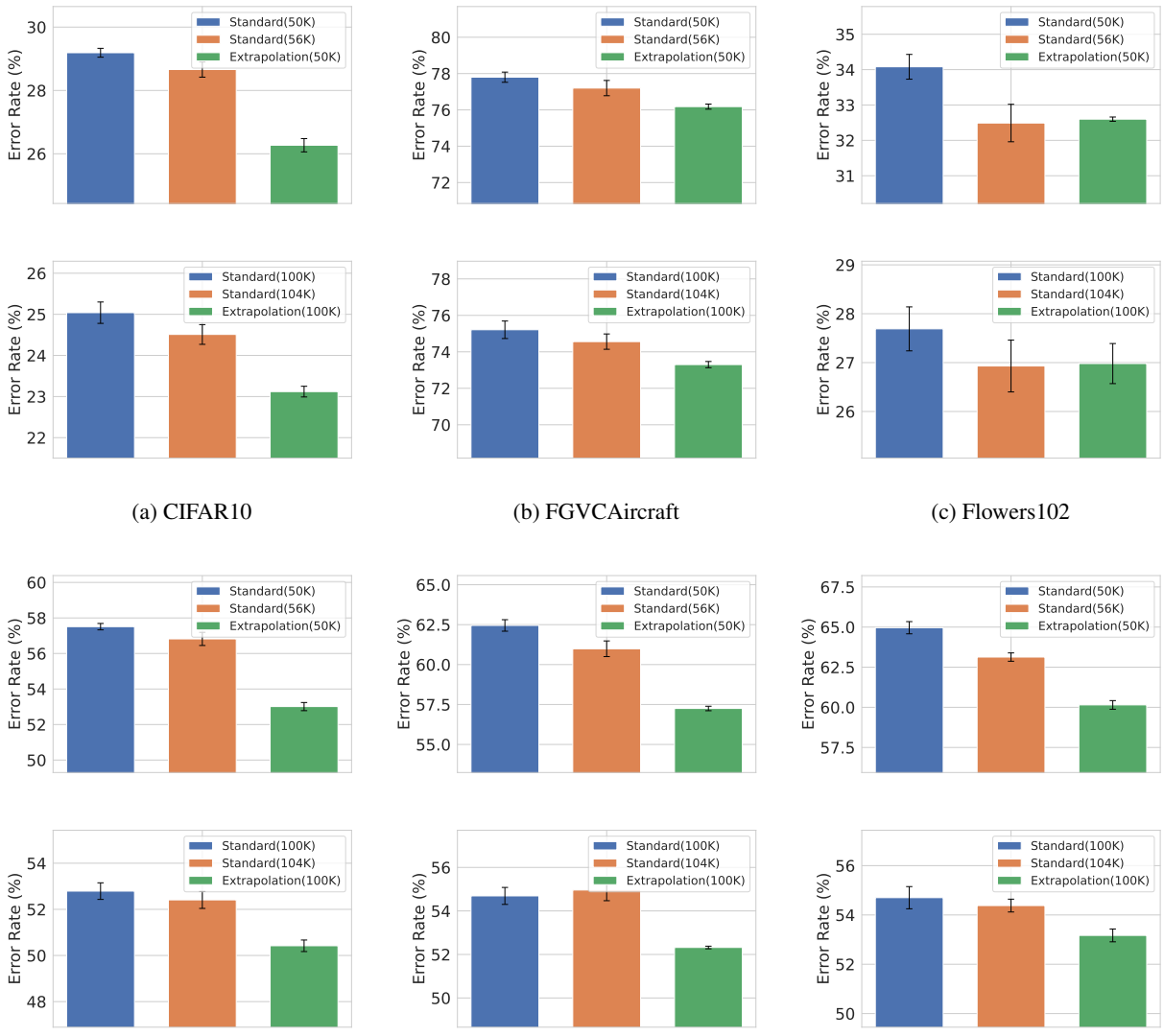
(d) MIT67 (e) Stanford40 (f) StanfordDogs

Figure 3. Test error rate across class-to-sample ratio. The vertical bar at each point is standard deviation.

performance trade-off caused by the inter-class diversity K and the intra-class diversity n. Instead, the downstream performance improves with increasing K, as suggested in Theorem 3.5.

A.4. Comparison of Extrapolation and Standard with Similar Budget

As the Extrapolation method requires extra data (not included in the final pre-training dataset) for estimating the optimal class-to-sample ratio, one might question whether incorporating the additional data into the final pre-training dataset using the Standard method ($K = 1000$) would result in better performance than the Extrapolation method. In this section, we present a comparative analysis of the Extrapolation method and the Standard method when the total number of data used is similar, to ensure a fair comparison. Figure 4 visualizes the error rates of different methods on various downstream tasks. Each bar plot represents a specific downstream task, and the number in parentheses indicates the size (N) of the corresponding pre-training dataset. We can observe that in most cases, the Extrapolation method demonstrates improved performance over the Standard method even when the latter uses a pre-training dataset larger than that of the Extrapolation plus the extra data needed.



(a) CIFAR10 (b) FGVCAircraft (c) Flowers102

(d) MIT67 (e) Stanford40 (f) StanfordDogs

Figure 4. As the Extrapolation method uses extra data for estimating the optimal class-to-sample ratio (see Table 2) compared to the Standard method, to render a fair comparison, we compare it against the Standard method with N larger than the total number of data used by Extrapolation. Each bar plot visualizes the error rates of different methods on a specific downstream task. The number in parentheses is N, i.e.,, the size of the corresponding pre-training dataset.

A.5. Proofs of Theoretical Results

A.5.1. Proof of Theorem 3.3

Proof. Define

$$h_{\mathcal{P}} \triangleq \arg\min_{h \in \mathcal{H}} \mathcal{R}_p(h, \mathcal{P}),$$
$$f_{\mathcal{P}} \triangleq \arg\min_{f} \mathbb{E}_{i \sim \mathrm{Unif}[K], x \sim \mathcal{P}_i}[\ell(f \circ h_{\mathcal{P}}(x), i)].$$

Step I. Bound $\mathcal{E}_p(\hat{h}_S, \mathcal{P})$.

Denote $\mathcal{E}_p(\hat{h}_S, \mathcal{P})$ can be decomposed into

$$\begin{aligned}\mathcal{E}_p(\hat{h}_S, \mathcal{P}) =& \mathcal{R}_p(\hat{h}_S, \mathcal{P}) - \min_{\tilde{h}\in\mathcal{H}} \mathcal{R}_p(\tilde{h}, \mathcal{P}) \\ =& \left(\mathcal{R}_p(\hat{h}_S, \mathcal{P}) - \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} \ell(\hat{f}_S \circ \hat{h}_S(x_{i,j}), i)\right) \\ &+ \left(\frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} \ell(\hat{f}_S \circ \hat{h}_S(x_{i,j}), i) - \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} (\ell(f_{\mathcal{P}} \circ h_{\mathcal{P}}(x_{i,j}), i))\right) \\ &+ \left(\frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} (\ell(f_{\mathcal{P}} \circ h_{\mathcal{P}}(x_{i,j}), i)) - \mathcal{R}_p(h_{\mathcal{P}}, \mathcal{P})\right).\end{aligned}$$

We tackle the three terms of the RHS of the above inequality respectively. As for the first term, since

$$\begin{aligned}\mathcal{R}_p(\hat{h}_S, \mathcal{P}) =& \min_{f\in\mathcal{F}} \mathbb{E}_{i\sim\mathrm{Unif}[K], x\sim\mathcal{P}_i}[\ell(f \circ \hat{h}_S(x), i)] \\ \leq& \mathbb{E}_{i\sim\mathrm{Unif}[K], x\sim\mathcal{P}_i}[\ell(\hat{f}_S \circ \hat{h}_S(x), i)],\end{aligned}$$

we have that the first term can be bounded by

$$\begin{aligned}&\mathcal{R}_p(\hat{h}_S, \mathcal{P}) - \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} \ell(\hat{f}_S \circ \hat{h}_S(x_{i,j}), i) \\ \leq& \mathbb{E}_{i\sim\mathrm{Unif}[K], x\sim\mathcal{P}_i}[\ell(\hat{f}_S \circ \hat{h}_S(x), i)] - \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} \ell(\hat{f}_S \circ \hat{h}_S(x_{i,j}), i) \\ \leq& \sup_{f\in\mathcal{F}, h\in\mathcal{H}} \left[\mathbb{E}_{i\sim\mathrm{Unif}[K], x\sim\mathcal{P}_i}[\ell(f \circ h(x), i)] - \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} \ell(f \circ h(x_{i,j}), i)\right].\end{aligned}$$

Applying gaussian complexity to $\frac{1}{n}\sum_{j=1}^{n}(\frac{1}{K}\sum_{i=1}^{K} \ell(f \circ h(x_{i,j}), i))$, we obtain that with probability at least $1-\delta$, the RHS of the above inequality is smaller than

$$\begin{aligned}&M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2n}} + 2\mathbb{E}_{((x_{i,j})_{i=1}^{K})_{j=1}^{n}\sim(\Pi_{i=1}^{k}\mathcal{P}_i)^K} \mathbb{E}_{(\sigma_i)_{i=1}^{n}\sim\mathcal{N}(0,1_{n\times n})} \sup_{f\in\mathcal{F}, h\in\mathcal{H}} \frac{1}{n}\sum_{j=1}^{n}\sigma_j \left(\frac{1}{K}\sum_{i=1}^{K} \ell(f \circ h(x_{i,j}), i)\right) \\ \leq& M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2n}} + \frac{2}{K}\sum_{i=1}^{K} \mathbb{E}_{(x_{i,j})_{j=1}^{n}\sim\mathcal{P}_i^K} \mathbb{E}_{(\sigma_i)_{i=1}^{n}\sim\mathcal{N}(0,1_{n\times n})} \sup_{f\in\mathcal{F}, h\in\mathcal{H}} \frac{1}{n}\sum_{j=1}^{n}\sigma_j \ell(f \circ h(x_{i,j}), i) \\ \leq& M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2n}} + \frac{2L_\ell}{K}\sum_{i=1}^{K} \mathbb{E}_{(x_{i,j})_{j=1}^{n}\sim\mathcal{P}_i^K} \mathbb{E}_{(\sigma_{j,l})_{j\in[n], l\in[\dim(\mathcal{Y})]}\sim\mathcal{N}(0,1_{n\dim(\mathcal{Y})})} \sup_{f\in\mathcal{F}, h\in\mathcal{H}} \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{\dim(\mathcal{Y})} \sigma_{j,l} f_l \circ h(x_{i,j}) \\ \leq& M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2n}} + 2GL_\ell\frac{1}{\sqrt{n}}.\end{aligned}$$

Here the second inequality is due to Slepian's Lemma, and the last inequality is due to Assumption 3.1. All in all, with probability at least $1-\delta$, the first term can be bounded as

$$\mathcal{R}_p(\hat{h}_S, \mathcal{P}) - \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} \ell(\hat{f}_S \circ \hat{h}_S(x_{i,j}), i) \leq M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2n}} + 2GL_\ell\frac{1}{\sqrt{n}}.$$

Meanwhile, the second term is non-positive due to the optimality of $\hat{f}_S$ and $\hat{h}_S$.

Finally, the third term can be bounded as

$$\begin{aligned}
&\left(\frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n}(\ell(f_{\mathcal{P}}\circ h_{\mathcal{P}}(x_{i,j}),i))-\mathcal{R}_p(h_{\mathcal{P}},\mathcal{P})\right)\\
=&\left(\frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n}(\ell(f_{\mathcal{P}}\circ h_{\mathcal{P}}(x_{i,j}),i))-\mathbb{E}_{i\sim\mathrm{Unif}[K],x\sim\mathcal{P}_i}[\ell(f_{\mathcal{P}}\circ h_{\mathcal{P}}(x),i)]\right)\\
\overset{w.p.1-\delta}{\leq}&M_\ell\sqrt{\frac{2\log\frac{1}{\delta}}{n}},
\end{aligned}$$

where the last inequality is due to Hoeffding's inequality. As a conclusion of Stage I, we obtain that with probability at least $1-2\delta$,

$$\mathcal{E}_p(\hat{h}_S,\mathcal{P})\leq 5M_\ell\sqrt{\frac{\log\frac{1}{\delta}}{2n}}+\frac{2GL_\ell}{\sqrt{n}}.$$

Step II. Bound $\mathcal{E}_d(\hat{h}_S)$. Applying Assumption 3.2, we obtain that with probability at least $1-2\delta$,

$$\mathcal{E}_d(\hat{h}_S)\leq\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\mathcal{E}_p(\hat{h}_S,\mathcal{P}_{i_1},\mathcal{P}_{i_2})+\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})\leq\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\left(5M_\ell\sqrt{\frac{\log\frac{1}{\delta}}{2n}}+\frac{2GL_\ell}{\sqrt{n}}\right)+\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}).$$

By McDiarmid's inequality, we obtain that with probability at least $1-\delta$,

$$\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})\leq\mathbb{E}_{\mathcal{P}\sim\mathcal{D}^K}\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})+M_0\sqrt{\frac{\log\frac{1}{\delta}}{2K}}\leq\nu_0^{\tilde{\mathcal{P}}}(\mathcal{D})+M_0\sqrt{\frac{\log\frac{1}{\delta}}{2K}}+\frac{C_0}{\sqrt{K}}.$$

Similarly, with probability at least $1-\delta$,

$$\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\leq\mathbb{E}_{\mathcal{P}\sim\mathcal{D}^K}\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})+M_1\sqrt{\frac{\log\frac{1}{\delta}}{2K}}\leq\nu_1^{\tilde{\mathcal{P}}}(\mathcal{D})+M_1\sqrt{\frac{\log\frac{1}{\delta}}{2K}}+\frac{C_1}{\sqrt{K}}.$$

As a conclusion, we obtain that with probability at least $1-4\delta$,

$$\mathcal{E}_d(\hat{h}_S)\leq\left(\nu_1^{\tilde{\mathcal{P}}}(\mathcal{D})+M_1\sqrt{\frac{\log\frac{1}{\delta}}{2K}}+\frac{C_1}{\sqrt{K}}\right)\left(5M_\ell\sqrt{\frac{\log\frac{1}{\delta}}{2n}}+\frac{2GL_\ell}{\sqrt{n}}\right)+\nu_0^{\tilde{\mathcal{P}}}(\mathcal{D})+M_0\sqrt{\frac{\log\frac{1}{\delta}}{2K}}+\frac{C_0}{\sqrt{K}}.$$

The proof is completed. □

A.6. Proof of Theorem 3.5

Proof. Denote

$$\begin{aligned}
h_{\mathcal{P}}&\triangleq\arg\min_{h\in\mathcal{H}}\mathcal{R}_p(h),\\
f_{\mathcal{P}}&\triangleq\arg\min_{f}\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f\circ h_{\mathcal{P}}(x),y)].
\end{aligned}$$

Step I. Bound $\mathcal{E}_p(\hat{h}_S)$.

Denote $\mathcal{E}_p(\hat{h}_S)$ can be decomposed into

$$\begin{aligned}\mathcal{E}_p(\hat{h}_S) =& \mathcal{R}_p(\hat{h}_S) - \min_{\tilde{h}\in\mathcal{H}} \mathcal{R}_p(\tilde{h}) \\ =& \left(\mathcal{R}_p(\hat{h}_S) - \frac{1}{N}\sum_{i=1}^N \ell(\hat{f}_S \circ \hat{h}_S(x_i), y_i)\right) \\ &+ \left(\frac{1}{N}\sum_{i=1}^N \ell(\hat{f}_S \circ \hat{h}_S(x_i), y_i) - \frac{1}{N}\sum_{i=1}^N \left(\ell(f_{\mathcal{P}} \circ h_{\mathcal{P}}(x_i), y_i)\right)\right) \\ &+ \left(\frac{1}{N}\sum_{i=1}^N \left(\ell(f_{\mathcal{P}} \circ h_{\mathcal{P}}(x_i), y_i)\right) - \mathcal{R}_p(h_{\mathcal{P}})\right).\end{aligned}$$

We tackle the three terms of the RHS of the above inequality respectively. As for the first term, since

$$\mathcal{R}_p(\hat{h}_S) = \min_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f \circ \hat{h}_S(x), y)] \le \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(\hat{f}_S \circ \hat{h}_S(x), y)],$$

we have that the first term can be bounded by

$$\begin{aligned}&\mathcal{R}_p(\hat{h}_S) - \frac{1}{N}\sum_{i=1}^N \ell(\hat{f}_S \circ \hat{h}_S(x_i), y_i) \\ \le& \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(\hat{f}_S \circ \hat{h}_S(x), i)] - \frac{1}{N}\sum_{i=1}^N \ell(\hat{f}_S \circ \hat{h}_S(x_i), y_i) \\ \le& \sup_{f\in\mathcal{F}, h\in\mathcal{H}} \left[\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f \circ h(x), i)] - \frac{1}{N}\sum_{i=1}^N \ell(f \circ h(x_i), y_i)\right].\end{aligned}$$

Applying gaussian complexity to $\frac{1}{N}\sum_{i=1}^N \ell(f \circ h(x_i), y_i)$, we obtain that with probability at least $1-\delta$, the RHS of the above inequality is smaller than

$$\begin{aligned}&M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2N}} + 2\mathbb{E}_{(x_i)_{i=1}^N\sim P^N}\mathbb{E}_{(\sigma_i)_{i=1}^N\sim\mathcal{N}(0,1_N)} \sup_{f\in\mathcal{F},h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^n \sigma_i \ell(f \circ h(x_i), y_i) \\ \le& M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2N}} + 2L_\ell\mathbb{E}_{(x_i)_{i=1}^N\sim P^N}\mathbb{E}_{(\sigma_{i,j})_{i\in[N],j\in[\dim(\mathcal{Y})]}\sim\mathcal{N}(0,1_{N\dim(\mathcal{Y})})} \sup_{f\in\mathcal{F},h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{\dim(\mathcal{Y})} \sigma_{i,j} f_j \circ h(x_i) \\ \le& M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2N}} + 2GL_\ell\frac{1}{\sqrt{N}}.\end{aligned}$$

Here the first inequality is due to Slepian's lemma, and the last inequality is due to Assumption 3.1. All in all, with probability at least $1-\delta$, the first term can be bounded as

$$\mathcal{R}_p(\hat{h}_S) - \frac{1}{N}\sum_{i=1}^N \ell(\hat{f}_S \circ \hat{h}_S(x_i), y_i) \le M_\ell\sqrt{\frac{9\log\frac{1}{\delta}}{2N}} + 2GL_\ell\frac{1}{\sqrt{N}}.$$

Meanwhile, the second term is non-positive due to the optimality of $\hat{f}_S$ and $\hat{h}_S$.

Finally, the third term can be bounded as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{P}}\circ h_{\mathcal{P}}(x_i),y_i)-\mathcal{R}_p(h_{\mathcal{P}})\\
=&\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathcal{P}}\circ h_{\mathcal{P}}(x_i),y_i)-\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f_{\mathcal{P}}\circ h_{\mathcal{P}}(x),y)]\\
\overset{w.p.1-\delta}{\leq}&M_\ell\sqrt{\frac{2\log\frac{1}{\delta}}{N}},
\end{aligned}$$

where the last inequality is due to Hoeffding's inequality. As a conclusion of Stage I, we obtain that with probability at least $1-2\delta$,

$$\mathcal{E}_p(\hat{h}_S,\mathcal{P})\leq 5M_\ell\sqrt{\frac{\log\frac{1}{\delta}}{2N}}+\frac{2GL_\ell}{\sqrt{N}}.$$

Step II. Bound $\mathcal{E}_d(\hat{h}_S)$. Applying Assumption 3.4, we obtain that with probability at least $1-2\delta$,

$$\mathcal{E}_d(\hat{h}_S)\leq\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\mathcal{E}_p(\hat{h}_S)+\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P})\leq\nu_1^{\tilde{\mathcal{P}}}(\mathcal{P})\left(5M_\ell\sqrt{\frac{\log\frac{1}{\delta}}{2N}}+\frac{2GL_\ell}{\sqrt{N}}\right)+\nu_0^{\tilde{\mathcal{P}}}(\mathcal{P}).$$

The proof is completed by applying Assumption 3.4. □