BLESSING OF CLASS DIVERSITY IN PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a new statistical analysis aiming to explain the recent superior achievements of the pre-training techniques in natural language processing (NLP). We prove that when the classes of the pre-training task (e.g., different words in masked language model task) are sufficiently diverse, in the sense that the least singular value of the last linear layer in pre-training is large, then pre-training can significantly improve the sample efficiency of downstream tasks. Inspired by our theory, we propose a new regularization technique that targets the multi-class pre-training: a *diversity regularizer only to the last linear layer* in the pre-training phase. Our empirical results show that this technique consistently boosts the performance of the pre-trained BERT model on different downstream tasks.

1 INTRODUCTION

Pre-training refers to training a model on a few or many tasks to help it learn parameters that can be used in other tasks. For example, in natural language processing (NLP), one first pre-trains a complex neural network model to predict masked words (masked language modeling), and then fine-tunes the model on downstream tasks, e.g., sentiment analysis (Devlin et al., 2019). Recently, pre-training technique has revolutionized natural language processing (NLP). Models based on this technique have dramatically improve the performance for a wide range of downstream tasks (Devlin et al., 2019; Radford et al., 2018; Yang et al., 2019; Clark et al., 2020; Lan et al., 2020; Liu et al., 2020).

Despite the large body of empirical work on pre-training, satisfactory theories are still lacking, especially theories that can explain the success of pre-training in NLP. Existing theories often rely on strong distributional assumptions (Lee et al., 2020), smoothness conditions (Robinson et al., 2020) or noise-robustness conditions (Bansal et al., 2021) to relate the pre-training task(s) to downstream tasks. These assumptions are often hard to verify.

A line of work studied *multi-task* pre-training (Caruana, 1997; Baxter, 2000; Maurer et al., 2016; Du et al., 2021; Tripuraneni et al., 2020a;b; Thekumparampil et al., 2021). In particular, recently, researchers have identified a new condition, the *diversity* of pre-training tasks, which has been shown to be crucial to allow pre-trained models to be useful for downstream tasks. See Section 2 for more detailed discussions on related work.

Unfortunately, this line of theory cannot be used to explain the success of pre-training in NLP. The theory of multi-task pre-training requires *a large number of diverse tasks*, e.g., the number of tasks needs to be larger than the last layer's input dimension (a.k.a. embedding dimension), which is typically 768, 1024 or 2048 (Devlin et al., 2019). However, in NLP pre-training, there are only a few, if not one, pre-training tasks. Therefore, we need a new theory that is applicable to this setting.

Furthermore, while existing theories are able to explain certain empirical phenomena, it remains unclear how to utilize the theory in practice. Ideally, we would like the theory to be useful to guide our training procedure or inspire new techniques to improve the performance of real-world models.

Since in NLP pre-training, we do not have multiple tasks, we propose to study *the blessing of multiple classes*. Concretely, consider the Masked Language Model (MLM) pre-training task in NLP. In such a pre-training task, we have a large collection of sentences (e.g. from Wikipedia). During the pre-training phase, we randomly mask a few words in each sentence, and predict the masked words using the remaining words in this sentence. This pre-training task is a *multi-class classification* *problem* where the number of classes is about 30K when using byte-pair-encoding (BPE) sub-word units.[1] Note that this number is much larger than the embedding dimension (768, 1024 or 2048).

In this paper, we develop a new statistical analysis aiming to explain the success of pre-training for NLP. The key notion of our theory is the *diversity of classes*, which serves as a similar role as the *diversity of tasks* in multi-task pre-training theory (Du et al., 2021; Tripuraneni et al., 2020a). Our theory is not only applicable to the real-world NLP pre-training setting, but also inspires new techniques to improve the practical performance. We summarize our contributions below.

1.1 Our Contributions

First, we define a new notion, *diversity of classes*, which is the least singular value of the last linear layer in pre-training. We prove finite-sample bounds to show that for the cross-entropy loss, if the diversity of classes is large, then pre-training on a single task provably improves the statistical efficiency of the downstream tasks. To our knowledge, this is the first set of theoretical results that demonstrates the statistical gain of the standard practice of NLP pre-training, without strong distributional or smoothness conditions.

Second, from a technical point of view, previous theoretical work on multi-task (Du et al., 2021; Tripuraneni et al., 2020b) builds on scalar output, and thus could not apply to multiclass tasks (e.g., cross entropy loss). We introduce a vector-form Radamacher complexity chain rule for disassembling composite function classes based on vector-form Rademacher contraction property (Maurer, 2016). This generalizes the scalar-form chain rule in Tripuraneni et al. (2020b). Furthermore, we adopt the modified self-concordance condition to show that the least singular value of the last linear layer serves as a diversity parameter for cross-entropy loss. We believe our techniques can be useful in other problems.

Third, inspired by our theory, we develop a new regularization technique to promote class diversity for multi-class pre-training. We apply the negative log determinant regularizer only to the last linear layer of the pre-training model in masked language modeling. Our empirical results on BERT-base show this technique can boost the performance of downstream tasks.

Organization This paper is organized as follows. In Section 2, we review the related work. In Section 3, we formally describe the problem setup and introduce the necessary definitions. In Section 4, we state our main Theorem 1 then instantiate it with several applications. In Section 5, we describe our new regularization technique and present the empirical results. We conclude in Section 6 and defer the proofs to Appendix.

2 Related Work

Here we mostly focus on the theoretical aspects of pre-training. While there is a long list of work demonstrating the empirical success of self-supervised learning, there are only a few papers that study its theoretical aspects. One line of work studied the theoretical properties of *contrastive learning* (Saunshi et al., 2019; Tosh et al., 2020), which is a different setting considered in this paper. The most relevant one is by Lee et al. (2020) which showed that if the input data and pre-training labels were independent (conditional on the downstream labels), then pre-training provably improved statistical efficiency. However, this conditional independence assumption rarely holds in practice. For example, in question-answering task, this assumption implies that given the answer, the question sentence and the masked word are independent. Robinson et al. (2020) assumed the Central Condition and a smoothness condition that related the pretraining task and the downstream task. Bansal et al. (2021) related generalization error of self-supervised learning to the noise-stability and rationality. However, it is difficult to verify the assumptions in these papers.

A recent line of work studied multitask pre-training (Du et al., 2021; Tripuraneni et al., 2020a;b; Thekumparampil et al., 2021) in which the notion, diversity, has been identified to be the key that enables pre-training to improve statistical efficiency. These works generally require a large number

[1]This is a standard setting in the BERT model (Devlin et al., 2019) and is widely adopted as a common practice. By breaking down English words into BPE sub-word units, it could drastically increase the coverage of English language by using a relatively small (32768) vocabulary.

of diverse tasks thus are not applicable to NLP, as we have mentioned. In comparison, we study single-task multi-class pre-training which is different from theirs. Du et al. (2021) noted that their results allowed an easy adaptation to multi-class settings in (Remark 6.2). However, they only focused on quadratic loss with one-hot labels for multi-class classification. Instead, we study the commonly used cross-entropy loss. While their analyses do not imply results in our setting, our theoretical analyses are inspired by this line of work.

Our paper uses a diversity regularizer proposed in Zou & Adams (2012) to improve the performance of pre-training. We note that there are other diversity regularizers (Xie et al., 2017; Mariet & Sra, 2015; Cogswell et al., 2015). These may also improve the performance as the one in Zou & Adams (2012). We leave it as a future work to investigate these regularizers.

3 PRELIMINARIES

In this section, we introduce the necessary notations, the problem setup, and several model-dependent quantities of pre-training.

3.1 NOTATION AND SETUP

Notations. Let $[n] = \{1, 2, \cdots, n\}$. We use $\|\cdot\|$ or $\|\cdot\|_2$ to denote the ℓ_2 norm of a vector. Let $\mathcal{N}(\mu, \sigma_2)$ be the one-dimensional Gaussian distribution. For a matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$, let $\|\mathbf{W}\|_{1,\infty} = \max_q(\sum_p |\mathbf{W}_{q,p}|)$ and $\|\mathbf{W}\|_{\infty\to 2}$ be the induced ∞-to-2 operator norm. We use the standard $O(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$ notation to hide universal constant factors, and use $\widetilde{O}(\cdot)$ to hide logarithmic factors. We also use $a \lesssim b$ to indicate $a = O(b)$.

Problem Setup. The procedure is divided into two stages: pre-training stage to find a representation function and the downstream training stage to obtain a predictor for the downstream task. In both stages, we use $\hat{R}$ to represent empirical risk and use R to represent expected loss.

In the first stage, we have one pre-training task with n samples, $\{x_i^{\text{pre}}, y_i^{\text{pre}}\}_{i=1}^n$, where $x_i^{\text{pre}} \in \mathcal{X}^{\text{pre}} \subset \mathbb{R}^d$ is the input and $y_i^{\text{pre}} \in \{0,1\}^{k-1}$ is the one-hot label for k-class classification (if y_i^{pre} is all-zero then it represents the k-th class).[2] For instance, in masked language modeling, the input of each sample is a sentence with one word masked out, and the label is the masked word.[3] k in this example is the size of the vocabulary ($\approx 30K$). We aim to obtain a good representation function $\hat{h}$ within a function class $\mathcal{H} \subset \{\mathbb{R}^d \to \mathbb{R}^r\}$ where r is the embedding dimension (often equals to 768, 1024, 2048 in NLP pre-training). For example, one popular choice of the representation function $\hat{h}$ in NLP applications is the Transformer model and its variants (Vaswani et al., 2017; Devlin et al., 2019). On top of the representation, we predict the labels using function f^{pre} within function class $\mathcal{F}^{\text{pre}} \subset \{\mathbb{R}^r \to \mathbb{R}^{k-1}\}$.

To train the representation function and predictor in pre-training stage, we consider the Empirical Risk Minimization (ERM) procedure

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \min_{f^{\text{pre}}\in\mathcal{F}^{\text{pre}}} \hat{R}_{\text{pre}}(f^{\text{pre}}, h) \triangleq \arg\min_{h\in\mathcal{H}} \min_{f^{\text{pre}}\in\mathcal{F}^{\text{pre}}} \frac{1}{n}\sum_{i=1}^{n} \ell(f^{\text{pre}} \circ h(x_i^{\text{pre}}), y_i^{\text{pre}}) \tag{1}$$

where ℓ is the loss function. We overload the notation for both the pre-training task and the downstream task, i.e., for pre-training, $\ell : \mathbb{R}^{k-1} \times \{0,1\}^{k-1} \to \mathbb{R}$ and for the downstream task, $\ell : \mathbb{R}^{k'-1} \times \{0,1\}^{k'-1} \to \mathbb{R}$. e.g., cross-entropy: $\ell(\hat{y}; y) = -y^\top \hat{y} + \log\left(1 + \sum_{s=1}^{k-1} \exp\left(\hat{y}_s\right)\right)$.

Now for the downstream task, we assume there are m samples $\{x_i^{\text{down}}, y_i^{\text{down}}\}_{i=1}^m$. Note that $x_i^{\text{down}} \in \mathcal{X}^{\text{down}} \subset \mathbb{R}^d$ is the input and $y_i^{\text{down}} \in \{0,1\}^{k'-1}$ is the one-hot label for k'-class classification.[4]

[2]We assume only one pre-training task for the ease of presentation. It is straightforward to generalize our results to multiple pre-training tasks.

[3]Here we say only one word being masked only for the ease of presentation. It is straightforward to generalize our results to the case where multiple words are masked out.

[4]For simplicity, we assume we only have one downstream task. Our theoretical results still apply if we have multiple downstream tasks.

Note that in most real-world applications, we have $n \gg m$ and $k \gg k'$. For example, in sentiment analysis, $k' = 2$ ("positive" or "negative"). A widely studied task SST-2 (Wang et al., 2019) has $m \approx 67K$, which is also generally much smaller than the pre-training corpus (e.g., $n > 100$M samples).

To train the classifier for the downstream task, we fix the representation function learned from the pre-training task and train the task-dependent predictor within function class $\mathcal{F}^{\text{down}} \subset \{\mathbb{R}^r \to \mathbb{R}^{k'-1}\}$:

$$\hat{f}^{\text{down}} = \underset{f^{\text{down}} \in \mathcal{F}^{\text{down}}}{\arg\min}\ \hat{R}_{\text{down}}(f^{\text{down}}, \hat{h}) = \underset{f^{\text{down}} \in \mathcal{F}^{\text{down}}}{\arg\min}\ \frac{1}{m} \sum_{i=1}^{m} \ell(f^{\text{down}} \circ \hat{h}(x_i^{\text{down}}), y_i^{\text{down}}). \quad (2)$$

Therefore, our predictor for the downstream task consists a pair $(\hat{f}^{\text{down}}, \hat{h})$. We use the Transfer Learning Risk defined below to measure the performance

$$\text{Transfer Learning Risk} \triangleq R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - \mathbb{E}_{x^{\text{down}}, y^{\text{down}}} \left[\ell\left(g^{\text{down}}\left(x^{\text{down}}\right), y^{\text{down}}\right)\right]$$

where $R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) \triangleq \mathbb{E}_{x^{\text{down}}, y^{\text{down}}} \left[\ell\left(\hat{f}^{\text{down}} \circ \hat{h}\left(x^{\text{down}}\right), y^{\text{down}}\right)\right]$ is the expected loss (the expectation is over the distribution of the downstream task), and $g^{\text{down}} = \arg\min_{g \in \{\mathbb{R}^d \to \mathbb{R}^{k'-1}\}} \mathbb{E}_{x^{\text{down}}, y^{\text{down}}} \left[\ell\left(g\left(x^{\text{down}}\right), y^{\text{down}}\right)\right]$ is the optimal predictor for the downstream task.

In our analysis, we also need to use the following term to characterize the quality of pre-training

$$\text{Pre-training Risk} = R_{\text{pre}}(\hat{f}^{\text{pre}}, \hat{h}) - \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}} \left[\ell\left(g^{\text{pre}}\left(x^{\text{pre}}\right), y^{\text{pre}}\right)\right]$$

where $R_{\text{pre}}(\hat{f}^{\text{pre}}, \hat{h}) \triangleq \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}} \left[\ell\left(\hat{f}^{\text{pre}} \circ \hat{h}\left(x^{\text{pre}}\right), y^{\text{pre}}\right)\right]$ is the expected loss, and $g^{\text{pre}} = \arg\min_{g \in \{\mathbb{R}^d \to \mathbb{R}^{k-1}\}} \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}} \left[\ell\left(g\left(x^{\text{pre}}\right), y^{\text{pre}}\right)\right]$ is the optimal predictor for the pre-training task.

Following the existing work on representation learning (Maurer et al., 2016; Du et al., 2021; Tripuraneni et al., 2020b), throughout the paper we make the following realizability assumption.

Assumption 1 (Realizability). *There exist $h \in \mathcal{H}$, $f^{\text{pre}} \in \mathcal{F}^{\text{pre}}$, $f^{\text{down}} \in \mathcal{F}^{\text{down}}$ such that $g^{\text{pre}} = f^{\text{pre}} \circ h$ and $g^{\text{down}} = f^{\text{down}} \circ h$.*

This assumption posits that the representation class and the task-dependent prediction classes are sufficiently expressive to contain the optimal functions. Importantly, the pre-training and downstream tasks share a *common* optimal representation function h. This assumption formalizes the intuition that pre-training learns a good representation that is also useful for downstream tasks.

3.2 Task Relatedness and Diversity

We use the following definitions, which are natural analogies of those in Tripuraneni et al. (2020b) for multi-task transfer learning. To measure the "closeness" between the learned representation and true underlying feature representation, we use the following metric, following Tripuraneni et al. (2020b):

Definition 1. *Let $h \in \mathcal{H}$ be the optimal representation function and $h' \in \mathcal{H}$ be any representation function. Let $f^{\text{pre}} \in \mathcal{F}$ be the optimal pre-training predictor on top of h. The pre-training representation difference is defined as:*

$$d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(h'; h) = \inf_{f' \in \mathcal{F}^{\text{pre}}} \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}} \left[\ell(f' \circ h'(x^{\text{pre}}), y^{\text{pre}}) - \ell(f^{\text{pre}} \circ h(x^{\text{pre}}), y^{\text{pre}})\right] \quad (3)$$

where the expectation is over the pre-training data distribution.

Intuitively, this measures the performance difference between the optimal predictor and the best possible predictor given a representation function h'.

For transfer learning, we also need to introduce a similar concept on the downstream task.

Definition 2. *Let $h \in \mathcal{H}$ be the optimal representation function and $h' \in \mathcal{H}$ be any representation function. On the downstream task, for a function class $\mathcal{F}^{\text{down}}$, let $f^{\text{down}} \in \mathcal{F}$ be the optimal* *pre-training predictor on top of h. We define the worst-case representation difference between representations h and $h' \in \mathcal{H}$ as:*

$$d_{\mathcal{F}^{\rm down}}(h'; h) = \sup_{f^{\rm down} \in \mathcal{F}^{\rm down}} \inf_{f' \in \mathcal{F}^{\rm down}} \mathbb{E}_{x^{\rm down}, y^{\rm down}} \left[\ell(f' \circ h'(x^{\rm down}), y^{\rm down}) - \ell(f^{\rm down} \circ h(x^{\rm down}), y^{\rm down}) \right]$$

where the expectation is over the data distribution of the downstream task.

We now introduce the key notion of *diversity*, which measure how well a learned representation, say h', from the pre-training task can be transferred to the downstream task.

Definition 3. *Let $h \in \mathcal{H}$ be the optimal representation function. Let $f^{\rm pre} \in \mathcal{F}^{\rm pre}$ be the optimal pre-training predictor on top of h. The **diversity parameter** $\nu > 0$ is the largest constant that satisfies*

$$d_{\mathcal{F}^{\rm down}}(h'; h) \leq \frac{d_{\mathcal{F}^{\rm pre}, f^{\rm pre}}(h'; h)}{\nu}, \forall h' \in \mathcal{H}. \tag{4}$$

While Definition 1- 3 are naturally defined from inspecting the pre-training procedure, it is not trivial to use these definitions to derive statistical guarantees. In particular, one of our key technical challenge is to show the least singular value of the last linear layer serves as a lower bound of the diversity parameter when $\mathcal{F}^{\rm pre}$ and $\mathcal{F}^{\rm down}$ are linear function classes.

3.3 Model Complexities

Lastly, we need to introduce some notions to measure the complexity of the function classes considered. In this paper, we consider Gaussian complexity which quantifies the extent to which the function in the class $\mathcal{Q}$ can be correlated with a noise sequence of length $n \times r$.

Definition 4 (Gaussian complexity)**.** *Let μ be a probability distribution on a set $\mathcal{X} \subset \mathbb{R}^d$ and suppose that $x_1, \cdots, x_n$ are independent samples selected according to μ. Let $\mathcal{Q}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}^r$. Define random variable*

$$\hat{G}_n(\mathcal{Q}) = \mathbb{E}\left[\sup_{q \in \mathcal{Q}} \frac{1}{n} \sum_{k=1}^{r} \sum_{i=1}^{n} g_{ki} q_k(x_i) \right] \tag{5}$$

as the empirical Rademacher complexity, where $q_k(x_i)$ is the k-th coordinate of the vector-valued function $q(x_i)$, g_{ki} ($k \in [r], i \in [n]$) are independent Gaussian $\mathcal{N}(0,1)$ random variables. The Gaussian complexity of $\mathcal{Q}$ is $G_n(\mathcal{Q}) = E_\mu \hat{G}_n(\mathcal{Q})$.

Our main results are stated in terms of the Gaussian complexity. In Section 4.3 and 4.4 we will plug in existing results of the Gaussian complexity of certain function classes to obtain concrete bounds.

We will need the following worst-case Gaussian complexity for the pre-training predictor within $\mathcal{F}^{\rm pre}$

$$\bar{G}_n(\mathcal{F}^{\rm pre}) = \max_{h(x_1), \cdots, h(x_n)} \hat{G}_n(\mathcal{F}^{\rm pre} | h \circ x^{\rm pre}), \tag{6}$$

here $h \in \mathcal{H}$ and $x^{\rm pre} = x_1, \cdots, x_n \in \mathcal{X}^{\rm pre}$. Similarly we define $\bar{G}_m(\mathcal{F}^{\rm down})$ as the worst-case Gaussian complexity for the downstream predictor within $\mathcal{F}^{\rm down}$.

We note that a closely related notion is Rademacher complexity. The empirical Rademacher complexity and Gaussian complexity only differ by a log factor (Ledoux & Talagrand, 1991).

4 Main Results

In this section we present our main theoretical results. In Section 4.1 we present an analysis in terms of the diversity parameter for general loss function under certain regularity conditions. In Section 4.2, we specialize the result to a setting that is most relevant to NLP pre-training applications, where $\mathcal{F}^{\rm pre}$ and $\mathcal{F}^{\rm down}$ are sets of linear functions and the loss is cross-entropy. In this particular case, our key result will show that one can use the singular value of the last linear to bound the diversity parameter. In Section 4.3 and 4.4 we instantiate our bounds on two concrete representation function classes: linear subspace and multi-layer network to showcase our main results.

4.1 Main Theorem

In this subsection, we present our generic end-to-end transfer learning guarantee for multi-class transfer learning problem. We do not impose any specific function class formulations. Throughout this subsection, we only make the following mild regularity assumptions to make our results general.

Assumption 2 (Regularity conditions)**.** *We assume the following regularity conditions hold:*

- *In pre-training, $\ell(\cdot,\cdot)$ is B^{pre}-bounded, and $\ell(\cdot,y)$ is L^{pre}-Lipschitz for all y.*
- *In downstream task, $\ell(\cdot,y)$ is B^{down}-bounded and L^{down}-Lipschitz for all y.*
- *Any predictor $f \in \mathcal{F}^{\mathrm{pre}}$ is $L(\mathcal{F}^{\mathrm{pre}})$-Lipschitz with respect to the Euclidean distance.*
- *Predictors are bounded: $\|f \circ h(x)\| \le D_{\mathcal{X}^{\mathrm{pre}}}$ for any $x \in \mathcal{X}^{\mathrm{pre}}, h \in \mathcal{H}, f \in \mathcal{F}^{\mathrm{pre}}$. Similarly $\|f \circ h(x)\| \le D_{\mathcal{X}^{\mathrm{down}}}$ for any $x \in \mathcal{X}^{\mathrm{down}}, h \in \mathcal{H}, f \in \mathcal{F}^{\mathrm{down}}$.*

Specifically, one can show that common task-dependent losses satisfy these conditions. For example, when ℓ is cross-entropy loss (cf. Appendix A.2) for $k-$class classification, we prove ℓ is $\sqrt{k-1}-$Lipschitz and $D_{\mathcal{X}}-$bounded where $\mathcal{X}$ denotes the input data domain.

Under these assumptions, we have the following guarantee.

Theorem 1. *Under Assumption 1 and 2, for a given fixed failure probability δ, with probability at least $1-\delta$ we have the Transfer Learning Risk upper bounded by:*

$$O\left(\frac{1}{\nu}\left\{L^{\mathrm{pre}}\left[\log(n)\cdot[L(\mathcal{F}^{\mathrm{pre}})\cdot G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\mathrm{pre}})] + \frac{\sqrt{k}D_{\mathcal{X}^{\mathrm{pre}}}}{n^2}\right] + B^{\mathrm{pre}}\sqrt{\frac{\log(1/\delta)}{n}}\right\}\right.$$
$$\left. + L^{\mathrm{down}}\cdot\bar{G}_m(\mathcal{F}^{\mathrm{down}}) + B^{\mathrm{down}}\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

The first line comes from pre-training ERM procedure and it accounts for the error of using an approximate optimal representation $\hat{h} \approx h$. The second line characterizes the statistical error of learning the downstream-task predictor f^{down} from m samples. Note the diversity parameter appears in the denominator, which relates the pre-training risk to the transfer learning risk. Theorem 1 shows the risk would be small if the Gaussian complexities are small. We will show concrete examples where $G_n(\mathcal{H})$ and $\bar{G}_n(\mathcal{F}^{\mathrm{pre}})$ are $O(\sqrt{1/n})$ and $\bar{G}_m$ scales as $O(\sqrt{1/m})$. We believe this theorem applies broadly beyond the concrete settings considered in this paper.

In comparison with previous results, transfer learning risk analyses in (Du et al., 2021; Tripuraneni et al., 2020b) focus on scalar output. Their results cannot be applied to multi-class transfer learning tasks. In Theorem 1, we generalize the analysis in (Tripuraneni et al., 2020b) to handle multi-class classification where the output is high dimensional (number of classes). Technically, in the proof, we introduce a vector-form Radamacher complexity chain rule for disassembling composite function classes by making use of the vector-form Rademacher contraction property (Maurer, 2016).

4.2 Main Results for Multi-class Classification with Cross-entropy Loss

Now we specialize the general result to the setting that is of most interest to NLP pre-training, where the loss function ℓ is cross-entropy and the $\mathcal{F}^{\mathrm{pre}}$ and $\mathcal{F}^{\mathrm{down}}$ are sets of linear functions. This choice is consistent with the NLP pre-training: e.g., BERT (Devlin et al., 2019) uses transformers as the representation learning function class $\mathcal{H}$ and uses word-embedding matrices as $\mathcal{F}^{\mathrm{pre}}$.

Formally we define

$$\mathcal{F}^{\mathrm{pre}} = \{f | f(z) = \alpha^\top z, \alpha \in \mathbb{R}^{r\times(k-1)}, \|\alpha_s\| \le c_1, \forall s \in [k-1]\} \tag{7}$$
$$\mathcal{F}^{\mathrm{down}} = \{f | f(z) = \alpha^\top z, \alpha \in \mathbb{R}^{r\times(k'-1)}, \|\alpha_s\| \le c_0, \forall s \in [k'-1]\} \tag{8}$$

where c_1 and c_0 are some universal constants. Then the regularity conditions are instantiated as:

- Pre-training loss $\ell(\cdot,y)$ is $\sqrt{k-1}$-Lipschitz and $B^{\mathrm{pre}} = D_{\mathcal{X}^{\mathrm{pre}}}$-bounded.

- Downstream loss is $\sqrt{k'-1}$-Lipschitz and $B^{\text{down}} = D_{\mathcal{X}^{\text{down}}}$-bounded.
- Any function $f \in \mathcal{F}^{\text{pre}}$ is $L(\mathcal{F}^{\text{pre}}) = c_1\sqrt{k-1}$-Lipschitz w.r.t. the l_2 distance.

See Appendix A.2 for detailed derivations. Next we discuss our main assumption that relates the diversity parameter to a concrete quantity of the last linear layer.

Assumption 3 (Lower Bounded Least Eigenvalue)**.** *Let the optimal linear predictor at the last layer for pre-training be* $\alpha^{\text{pre}} \in \mathbb{R}^{r\times(k-1)}$*. We assume* $\tilde{\nu} \triangleq \sigma_r(\alpha^{\text{pre}}\,(\alpha^{\text{pre}})^\top) > 0$.

Similar assumptions have been used in multi-task representation learning (Du et al., 2021; Tripuraneni et al., 2020a;b), and are shown to be necessary (Maurer et al., 2016; Du et al., 2021). Different from their versions, our assumption is tailored for the multi-class classification setting.

One of our key technical contributions is to show $\tilde{\nu}$ serves as a lower bound for the diversity parameter ν. See Lemma 3 in Appendix. Intuitively, this assumption ensures that the pre-training task matrix spans the entire r-dimensional space and thus covers the output of the optimal representation $h(\cdot) \in \mathbb{R}^r$. This is quantitatively captured by the $\sigma_r(\alpha^{\text{pre}}\,(\alpha^{\text{pre}})^\top)$, which measures how spread out these vectors are in $\mathbb{R}^r$. Technically, we apply a modified *self-concordance* condition for better characterizing multinomial logistic regression (Bach et al., 2010). See Appendix A.4 for proof.

We now state our theorem for this specific setting.

Theorem 2. *Under Assumptions 1 and 3, with probability at least* $1-\delta$ *we have the Transfer Learning Risk upper bounded by:*

$$O\left(\frac{1}{\tilde{\nu}}\left\{\sqrt{k}\left[\log(n)[\sqrt{k}\cdot G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\text{pre}})] + \frac{\sqrt{k}D_{\mathcal{X}^{\text{pre}}}}{n^2}\right] + D_{\mathcal{X}^{\text{pre}}}\sqrt{\frac{\log(1/\delta)}{n}}\right\}\right.$$
$$\left. + \sqrt{k'}\cdot\mathbb{E}_{\mathcal{X}^{\text{down}}}\hat{G}_m(\mathcal{F}^{\text{down}}|\hat{h}\circ x^{\text{down}}) + \sigma\sqrt{\frac{\log(1/\delta)}{m}} + D_{\mathcal{X}^{\text{down}}}\sqrt{\frac{\log(1/\delta)}{m}}\right)$$

Here $\mathbb{E}_{\mathcal{X}^{\text{down}}}\hat{G}_m(\mathcal{F}^{\text{down}}|\hat{h}\circ x^{\text{down}})$ *is Gaussian complexity of embeddings*

$$\hat{h}\circ x^{\text{down}} = \{\hat{h}(x_1),\cdots,\hat{h}(x_m)|x^{\text{down}} = x_1,\cdots,x_m \in \mathcal{X}^{\text{down}}\} \tag{9}$$

where the expectation is over $\mathcal{X}^{\text{down}}$*, and* $\sigma^2 = \frac{1}{m}\sup_{f\in\mathcal{F}^{\text{down}}}\sum_{i=1}^m Var(\ell(f\circ\hat{h}(x_i^{\text{down}}), y_i^{\text{down}}))$ *is the maximal variance over* $\mathcal{F}^{\text{down}}$.

We remark that in Theorem 2, since we specialize to the case where $\mathcal{F}^{\text{pre}}$ and $\mathcal{F}^{\text{down}}$ are sets of linear functions, we can replace the term $L^{\text{down}}\cdot\bar{G}_m(\mathcal{F}^{\text{down}})$ in Theorem 1 by $\left(\sqrt{k'}\cdot\mathbb{E}_{\mathcal{X}^{\text{down}}}\hat{G}_m(\mathcal{F}^{\text{down}}|\hat{h}\circ x^{\text{down}}) + \sigma\sqrt{\frac{\log(1/\delta)}{m}}\right)$ by utilizing the functional Bernstein inequality. This improvement can help us obtain Theorem 3. See Appendix A.2 for the full proof.

Now we discuss the interpretation of Theorem 2. Typically, $\bar{G}_n(\mathcal{F}^{\text{pre}})$ is much smaller than $G_n(\mathcal{H})$ because $G_n(\mathcal{H})$ represents the complexity of the representation function, which is often complex. In practice, n is often large. Therefore, in the benign case where $\tilde{\nu} = \Theta(k)$ (when the condition number of α^{pre} is $O(1)$), then the dominating term will be $\bar{G}_n(\mathcal{F}^{\text{pre}})$. As we will show in the following subsections, this term typically scales as $O(\sqrt{1/n})$. Together, this theorem clearly shows when 1) the number of pre-training data is large, and 2) the least singular value of the last linear layer for pre-training is large, the transfer learning risk is small. On the other hand, if $\tilde{\nu}$ is small, then the bound becomes loose. This is consistent with prior counter examples on multi-task pre-training (Maurer et al., 2016; Du et al., 2021) where the diversity is shown to be necessary.

4.3 Linear Subspace Representation

We instantiate this setting by assuming the underlying representation to be a projection onto a low-dimensional subspace. For $r \ll d$, we let the function class be

$$\mathcal{H} = \{h|h(x) = B^\top x, B \in \mathbb{R}^{d\times r}, \text{B is a matrix with orthonormal columns}\} \tag{10}$$

We require some additional regularity conditions. Following prior work (Du et al., 2021; Tripuraneni et al., 2020b), we assume that $\|x\| \le D$ and input data distribution satisfies the following condition.

Definition 5. *We say the covariate distribution $P_x(\cdot)$ is Σ-sub-gaussian if for all $v \in \mathbb{R}^d$, $\mathbb{E}[\exp(v^\top x)] \leq \exp\left(\frac{\|\Sigma^{1/2}v\|^2}{2}\right)$ where the covariance Σ further satisfies $\sigma_{max}(\Sigma) \leq C$ and $\sigma_{min}(\Sigma) \geq c \geq 0$ for universal constants c, C.*

We have the following theorem that guarantees the performance of transfer learning.

Theorem 3. *Suppose Assumption 1 and 3 hold. For a sufficiently large constant c_3, we assume $n \geq c_3 d, m \geq c_3 r, D \leq c_3(\min(\sqrt{dr^2}, \sqrt{rm}))$. Then with probability at least $1-\delta$, we have the Transfer Learning Risk upper bounded by:*

$$O\left(\frac{1}{\tilde{\nu}}\left[\sqrt{k}\log(n)\left(\sqrt{\frac{kdr^2}{n}} + k\sqrt{\frac{r}{n}}\right) + \sqrt{k}D\left(\frac{k}{n^2} + \sqrt{\frac{\log(1/\delta)}{n}}\right)\right]\right.$$
$$\left. + (k')^{\frac{3}{2}}\sqrt{\frac{r}{m}} + k'\sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{k'}D\sqrt{\frac{\log(1/\delta)}{m}}\right)$$

To interpret this bound, consider the practically relevant scenario where $k' = O(1)$ (e.g., sentiment analysis), $m \ll n$, $k \ll n$ and $r \ll d$, in the benign case $\tilde{\nu} = \Omega(k)$, we have the transfer learning risk $\widetilde{O}\left(\sqrt{\frac{dr^2}{n}} + \sqrt{\frac{r}{m}}\right)$. Note that this is exactly the desired theoretical guarantee because the first term accounts for using all pre-training data to learn the representation function and the second term accounts for using the downstream data to learn the last linear layer. This is significantly better than not using pre-training, in which case the risk scales $O\left(\sqrt{\frac{d}{m}}\right)$.

4.4 Deep Neural Network Representation

In this subsection, we assume the underlying representation to be a practical $\sigma = tanh$-activated neural network. Predictors are still required to be linear functions at the interest of NLP pre-training, i.e.,

$$\mathcal{H} = \{h | h(x) = W_K\sigma\left(W_{K-1}\sigma\left(\cdots\sigma\left(W_1 x\right)\right)\right)\}, \tag{11}$$
$$\mathcal{F}^{\text{pre}} = \{f | f(z) = \alpha^\top z, \alpha \in \mathbb{R}^{r\times(k-1)}, \|\alpha_s\| \leq c_1 M(K)^2, \forall s \in [k-1]\}, \tag{12}$$
$$\mathcal{F}^{\text{down}} = \{f | f(z) = \alpha^\top z, \alpha \in \mathbb{R}^{r\times(k'-1)}, \|\alpha_s\| \leq c_0 M(K)^2, \forall s \in [k'-1]\}. \tag{13}$$

We further assume 1) for each $p \in [K]$, $\|W_p\|_{1,\infty} \leq M(p)$, and 2) $\|W_K\|_{\infty\to 2} \leq M(K)$. Adapt Gaussian complexity results in Golowich et al. (2018) we have

$$G_n(\mathcal{H}) \leq \widetilde{\mathcal{O}}\left(\frac{rM(K)\cdot D\sqrt{K}\cdot\Pi_{p=1}^{K-1}M(p)}{\sqrt{n}}\right), \quad \bar{G}_n(\mathcal{F}^{\text{pre}}) \leq \mathcal{O}\left(\frac{(k-1)M(K)^3}{\sqrt{n}}\right). \tag{14}$$

Now we are ready to state our theorem for this practical setting of NLP pre-training.

Theorem 4. *Under Assumption 1 and 2, assume $M(K) \geq c_3$ for a universal constant c_3. Then with probability at least $1-\delta$, Transfer Learning Risk is upper bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{krM(K)^3\cdot D\sqrt{K}\cdot\Pi_{p=1}^{K-1}M(p)}{\tilde{\nu}\sqrt{n}} + \frac{kM(K)^3}{\tilde{\nu}\sqrt{n}} + \frac{k'^{\frac{3}{2}}M(K)^3}{\sqrt{m}}\right).$$

5 Experiments

Our theoretical analysis in previous sections implies that the diversity of the model parameter matrix at the linear output layer in pre-training has a significant impact on the transfer capability, in the sense that the larger ν (diversity parameter of f^{pre}), the smaller the risk. Therefore, we could *explicitly add a diversity regularizer* to the linear output layer to increase diversity. Motivated by this, we propose to add the following diversity regularizer to the original BERT pre-training loss so that it becomes:

$$L'(\Theta) = L(\Theta) - \lambda \cdot \ln\det(\alpha^{\text{pre}}\left(\alpha^{\text{pre}}\right)^\top), \tag{15}$$

Table 1: **Performance of diversity-regularized BERT pre-training with different values of diversity factor λ.** We finetune the pretrained model on 8 downstream tasks from GLUE benchmark and evaluate them on their dev sets. All results are "mean (std)" from 5 runs with different random seeds. For MNLI, we average the accuracies on its matched and mismatched dev sets. For MRPC and QQP, we average their accuracy and F1 scores. For STS-B, we average Pearson's correlation and Spearman's correlation. All other tasks uses accuracy as the metric. The better-than-baseline numbers are underlined, and the best numbers are highlighted in boldface.

Model	MNLI	MRPC	SST-2	CoLA	QQP	QNLI	RTE	STS-B
BERT-base ($\lambda = 0.005$)	**84.17** (0.23)	87.16 (1.81)	92.48 (0.19)	59.99 (0.28)	89.42 (0.08)	**88.11** (0.54)	67.28 (3.43)	89.33 (0.07)
BERT-base ($\lambda = 0.05$)	84.01 (0.10)	86.35 (5.15)	**93.00** (0.16)	**62.66** (1.07)	**89.46** (0.03)	87.64 (0.44)	60.64 (6.08)	**89.57** (0.13)
BERT-base ($\lambda = 0.5$)	84.00 (0.20)	**89.42** (0.51)	92.93 (0.24)	60.76 (0.71)	89.33 (0.12)	88.01 (0.23)	**67.93** (1.18)	89.22 (0.23)
BERT-base (reproduced)	83.96 (0.08)	86.14 (4.64)	92.64 (0.20)	61.46 (0.74)	89.28 (0.09)	88.10 (0.27)	63.64 (6.64)	89.19 (0.07)

where Θ denotes the set of all model parameters, λ is a hyper-parameter that controls the magnitude of the diversity regularization, $\det(\cdot)$ denote the determinant of a matrix, and α^{pre} is the model parameter matrix at the output linear layer. This type of diversity regularizer was proposed in Zou & Adams (2012). This regularization technique is different from prior work because it is specifically designed for multi-class pre-training: we only add the diversity regularizer to the last linear layer.

We use the above diversity-regularized loss (along with the original ℓ_2-regularization) to pretrain BERT-base models under different values of diversity factor λ. Then we fine-tune them on 7 classification tasks and 1 regression task from the GLUE benchmark (Wang et al., 2019) to evaluate their transfer performance.[5] Our pre-training and finetuning implementations are based on the opensource code released by Nvidia.[6] We use the same pre-training data as the original BERT (i.e., English Wikipedia + TorontoBookCorpus).[7] Our detailed pre-training and finetuning hyper-parameters along with other experimental details are reported in Appendix A.6.2.

In Table 5, we report our performance on the dev sets of the 8 downstream tasks. All the experiments are repeated 5 times with different random seeds, and we report their mean values along with the standard deviations. The complete experiment results (including full MNLI, QQP, and MRPC results) can be found in Appendix A.6. From Table 5, we note that adding the diversity regularization could generally improve the performance on these downstream tasks. In particular, when $\lambda = 0.5$, our pretrained model outperforms the original BERT-base on 6 out of 8 tasks (with 3 of them being significant), while achieving comparable performance on the other 2 tasks. Although our model is slightly behind the original BERT on CoLA and QNLI, such a performance gap is not statistically significant. Besides, we also see that our model with $\lambda = 0.5$ achieves a much more stable performance (i.e., smaller std) on tasks with scarce finetuning data ($<$ 4K samples in MRPC and RTE). Our results, albeit still preliminary, demonstrate the potential of such a simple diversity-regularizer. It could be an effective and simple performance booster for any of the existing pre-trained NLP models (e.g., XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2020), ALBERT (Lan et al., 2020), etc) with negligible computation and implementation cost. We leave the development of the more advanced diversity regularizer as a future work.

6 Conclusion

We theoretically prove the benefit of multi-class pre-training using the notion of class diversity. Inspired by the theory, we further propose a new regularization technique specially designed for pre-training. Our experiments show potential of this diversity regularizer. An interesting direction is to further investigate the impact of different diversity regularizers on larger pre-training models. Lastly, we only studied multi-class pre-training in this paper, it is interesting to develop a theory for other pre-training techniques such as those based on sequence reconstruction (Lewis et al., 2019).

[5]We do not report the WNLI (classification) task due to its reported issues of the task in Devlin et al. (2019).

[6]Distributed under Apache License: `https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT`

[7]Collected and pre-processed using the code and script included in the open-source code: `https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT`

REFERENCES

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

Francis Bach et al. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Yamini Bansal, Gal Kaplun, and Boaz Barak. For self-supervised learning, Rationality implies generalization, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Srmggo3b3X6.

Jonathan Baxter. A Model of Inductive Bias Learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Rich Caruana. Multitask Learning. *Machine learning*, 28(1):41–75, 1997.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing Overfitting in Deep Networks by Decorrelating Representations. *arXiv preprint arXiv:1511.06068*, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent Sample Complexity of Neural Networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991. ISBN 9783540520139. URL https://books.google.com.hk/books?id=cyKYDfvxRjsC.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2020. URL https://openreview.net/forum?id=SyxS0T4tvS.

Zelda Mariet and Suvrit Sra. Diversity Networks: Neural Network Compression Using Determinantal Point Processes. *arXiv preprint arXiv:1511.05077*, 2015.

Pascal Massart. About the Constants in Talagrand's Concentration Inequalities for Empirical Processes. *Annals of Probability*, pp. 863–884, 2000.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The Benefit of Multitask Representation Learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.

Joshua Robinson, Stefanie Jegelka, and Suvrit Sra. Strength from weakness: Fast learning using weak supervision. In *International Conference on Machine Learning*, pp. 8127–8136. PMLR, 2020.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *International Conference on Machine Learning*, pp. 5628–5637, 2019.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample Efficient Linear Meta-Learning by Alternating Minimization. *arXiv preprint arXiv:2105.08306*, 2021.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. *arXiv preprint arXiv:2008.10150*, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable Meta-Learning of Linear Representations. *arXiv preprint arXiv:2002.11684*, 2020a.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the Theory of Transfer Learning: The Importance of Task Diversity. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL `https://proceedings.neurips.cc/paper/2020/hash/59587bffec1c7846f3e34230141556ae-Abstract.html`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJ4km2R5t7`.

Pengtao Xie, Aarti Singh, and Eric P Xing. Uncorrelation and Evenness: a New Diversity-Promoting Regularizer. In *International Conference on Machine Learning*, pp. 3811–3820. PMLR, 2017.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html`.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

James Y Zou and Ryan P Adams. Priors for Diversity in Generative Latent Variable Models. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2996–3004, 2012.

A APPENDIX

In Section 3, we have introduced Gaussian complexity. Let us restate for clarity.

Let μ be a probability distribution on a set $\mathcal{X} \subset \mathbb{R}^d$ and suppose that $x_1, \cdots, x_n$ are independent samples selected according to μ. Let $\mathcal{Q}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}^r$. Define random variable

$$\hat{G}_n(\mathcal{Q}) = \mathbb{E}\left[\sup_{q\in\mathcal{Q}} \frac{1}{n}\sum_{k=1}^{r}\sum_{i=1}^{n} g_{ki} q_k(x_i)\right] \tag{16}$$

as the empirical Rademacher complexity, where $q_k(x_i)$ is the k-th coordinate of the vector-valued function $q(x_i)$, g_{ki} $(k \in [r], i \in [n])$ are independent Gaussian $\mathcal{N}(0,1)$ random variables. The Gaussian complexity of $\mathcal{Q}$ is $G_n(\mathcal{Q}) = E_\mu \hat{G}_n(\mathcal{Q})$.

Analogously to the above we can define the empirical Rademacher complexity for vector-valued functions as

$$\hat{R}_n(\mathcal{Q}) = \mathbb{E}\left[\sup_{q\in\mathcal{Q}} \frac{1}{N}\sum_{k=1}^{r}\sum_{i=1}^{N} \epsilon_{ki} q_k(x_i)\right] \tag{17}$$

where $\epsilon_{ki}(k \in [r], i \in [n])$ are independent Rademacher Rad$(\frac{1}{2})$ random variables. Its population counterpart is defined as $R_n(\mathcal{Q}) = E_\mu[\hat{R}_n(\mathcal{Q})]$. Note that the superscripts existing in $\hat{G}$ and $\hat{R}$ imply that they are empirical measures.

A.1 PROOFS FOR THEOREM 1

We illustrate Theorem 1 in two stages. First we show pre-training representation difference can be upper bounded by constants and function class complexities. Then we transfer it to the downstream task through the diversity parameter.

Pre-training

Theorem 5. *In pre-training, with probability at least $1-\delta$, it holds that:*

$$\begin{aligned} & d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(h'; h) \\ & \le 4\sqrt{\pi} L^{\text{pre}} G_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) + 4B^{\text{pre}}\sqrt{\frac{\log(2/\delta)}{n}} \\ & \le 4096 L^{\text{pre}}\left[\frac{\sqrt{k-1}D_{\mathcal{X}^{\text{pre}}}}{n^2} + \log(n)[L(\mathcal{F}^{\text{pre}})G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\text{pre}})]\right] + 4B^{\text{pre}}\sqrt{\frac{\log(2/\delta)}{n}}. \end{aligned}$$

Proof. We begin with

$$d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(h'; h) \le 2 \sup_{f\in\mathcal{F}^{\text{pre}}, h\in\mathcal{H}} |R_{\text{pre}}(f^{\text{pre}}, h) - \hat{R}_{\text{pre}}(f^{\text{pre}}, h)|.$$

From the definition of Rademacher complexity (Wainwright, 2019, Theorem 4.12), with probability at least $1-2\delta$ we have

$$\sup_{f^{\text{pre}}\in\mathcal{F}^{\text{pre}}, h\in\mathcal{H}} |R_{\text{pre}}(f^{\text{pre}}, h) - \hat{R}_{\text{pre}}(f^{\text{pre}}, h)| \le 2R_n(\ell(\mathcal{F}^{\text{pre}} \circ \mathcal{H})) + 2B^{\text{pre}}\sqrt{\frac{\log(1/\delta)}{n}}.$$

Next, we apply the vector contraction inequality (Maurer, 2016). For function class $\mathcal{F}$ whose output is in $\mathbb{R}^K$ with component $f_k(\cdot)$, and the function (h_i)s are some L-Lipschitz functions: $\mathbb{R}^K \mapsto \mathbb{R}$, we have

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \epsilon_i h_i(f(x_i)) \le \sqrt{2} L \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{i=1}^{n}\sum_{k=1}^{K} \epsilon_{ik} f_k(x_i). \tag{18}$$

Hence for loss function ℓ satisfying $|\ell(x)-\ell(y)| \leq L^{\text{pre}}\|x-y\|_2, \forall x, y \in \mathbb{R}^{k-1}$, the f takes value in $\mathbb{R}^{k-1}$ with component functions $f_s(\cdot), s \in [k-1]$, we have that population Rademacher complexity can be bounded by

$$\begin{aligned}
R_n(l(\mathcal{F}^{\text{pre}} \circ \mathcal{H})) &= \mathbb{E}_{X^{\text{pre}}} \frac{1}{n} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}^{\text{pre}}, h \in \mathcal{H}} \sum_{i=1}^n \epsilon_i \ell(f \circ h(x_i^{\text{pre}})) \\
&\leq \mathbb{E}_{X^{\text{pre}}} \frac{1}{n} \sqrt{2} L^{\text{pre}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}^{\text{pre}}, h \in \mathcal{F}} \sum_{i=1}^n \sum_{s=1}^{k-1} \epsilon_{is} f_s(h(x_i^{\text{pre}})) \\
&= \sqrt{2} L^{\text{pre}} R_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) \\
&\leq \sqrt{\pi} L^{\text{pre}} G_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H})
\end{aligned}$$

where the last line uses the fact that Rademacher complexity is upper bounded by Gaussian complexity: $R_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) \leq \sqrt{\frac{\pi}{2}} G_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H})$. Therefore we have, with probability at least $1-\delta$,

$$\begin{aligned}
&d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(h'; h) \\
&\leq 4\sqrt{\pi} L^{\text{pre}} G_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) + 4B^{\text{pre}} \sqrt{\frac{\log(2/\delta)}{n}} \\
&\leq 4096 L^{\text{pre}} \left[\frac{\sqrt{k-1} D_{\mathcal{X}^{\text{pre}}}}{n^2} + \log(n)[L(\mathcal{F}^{\text{pre}}) G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\text{pre}})] \right] + 4B^{\text{pre}} \sqrt{\frac{\log(2/\delta)}{n}},
\end{aligned}$$

where the last line uses decomposition of $G_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H})$ into the individual Gaussian complexities of $\mathcal{H}$ and $\mathcal{F}^{\text{pre}}$, leverages an expectation version of novel chain rule for Gaussian complexities (Lemma 1). □

In the spirit of Gaussian complexity decomposition theorem (Tripuraneni et al., 2020b, Theorem 7), we introduce the following decomposition result upon vector-form Gaussian complexities.

Lemma 1. *We have the following vector form Gaussian complexity decomposition:*

$$\hat{G}_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) \leq \frac{8\sqrt{k-1} D_{\mathcal{X}^{\text{pre}}}}{n^2} + 512 C(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) \cdot \log(n) \tag{19}$$

where we use $C(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) = L(\mathcal{F}^{\text{pre}}) \cdot \hat{G}_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\text{pre}})$ to represent the complexity measure of the composite function class.

Proof. Our proof extends (Tripuraneni et al., 2020b, Theorem 7), which focuses on a multi-task scalar formulation. We further extend it to multi-class vector formulation. Specifically, on top of the representation class $\mathcal{H}$, they need to handle $\mathcal{F}^{\otimes t}$ (t is the number of tasks) while our objective is a single function class $\mathcal{F}^{\text{pre}}$ of higher dimension ($\mathcal{F}^{\text{pre}}$ is $(k-1)$-dimensional for a k-class classification task). We note that our proof technique and that of previous works (Tripuraneni et al., 2020b; Maurer et al., 2016) both hinge on several properties of Gaussian processes.

To bound the empirical composite function class $\mathcal{F}^{\text{pre}}(\mathcal{H})$, note that vector-form Gaussian complexity is defined as

$$\begin{aligned}
\hat{G}_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) &= \mathbb{E}\left[\frac{1}{n} \hat{\sup}_{f(h) \in \mathcal{F}^{\text{pre}}(\mathcal{H})} \sum_{s=1}^{k-1} \sum_{i=1}^n g_{is} f_s(h(x_i^{\text{pre}})) \right] \\
&= \frac{1}{\sqrt{n}} \mathbb{E}[\sup_{f(h) \in \mathcal{F}^{\text{pre}}(\mathcal{H})} Z_{f(h)}]
\end{aligned}$$

where we define mean-zero process $Z_{f(h)} = \frac{1}{\sqrt{n}} \sum_{s=1}^{k-1} \sum_{i=1}^n g_{is} f_s(h(x_i^{\text{pre}}))$, then $\mathbb{E} \sup_{f(h)} Z_{f(h)} = \mathbb{E} \sup_{f(h)} Z_{f(h)} - Z_{f'(h')} \leq \mathbb{E} \sup_{f(h), f'(h')} Z_{f(h)} - Z_{f'(h')}$. We further notice that $Z_{f(h)} - Z_{f'(h')}$ is a sub-gaussian random variable parameter

$$\begin{aligned}
d^2(f(h), f'(h') | x^{\text{pre}}) &= \frac{1}{n} \sum_{i=1}^n \|f(h(x_i^{\text{pre}})) - f'(h'(x_i^{\text{pre}}))\|^2 \\
&= \frac{1}{n} \sum_{s=1}^{k-1} \sum_{i=1}^n (f_s(h(x_i^{\text{pre}})) - f'_s(h'(x_i^{\text{pre}})))^2
\end{aligned}$$

Dudley's entropy integral bound (Wainwright, 2019, Theorem 5.22) shows

$$\begin{aligned}
&\mathbb{E} \sup_{f(h), f'(h')} Z_{f(h)} - Z_{f'(h')} \\
&\leq 2\mathbb{E} \sup_{d(f(h), f'(h')|x^{\text{pre}}) \leq \delta} Z_{f(h)} - Z_{f'(h')} + 32\mathcal{J}(\frac{\delta}{4}, D_{\mathcal{X}^{\text{pre}}}) \\
&= 2\mathbb{E} \sup_{d(f(h), f'(h')|x^{\text{pre}}) \leq \delta} Z_{f(h)} - Z_{f'(h')} + 32 \int_{\frac{\delta}{4}}^{D_{\mathcal{X}^{\text{pre}}}} \sqrt{\log N(u; \mathcal{F}^{\text{pre}}(\mathcal{H})|x^{\text{pre}})} du.
\end{aligned}$$

It is straightforward to see the first term follows:

$$\mathbb{E} \sup_{d(f(h), f'(h')|x^{\text{pre}}) \leq \delta} Z_{f(h)} - Z_{f'(h')} \leq \mathbb{E}[\|g\|]\delta \leq \sqrt{n(k-1)}\delta$$

We now turn to bound the second term by decomposing the distance metric into a distance over $\mathcal{F}^{\text{pre}}$ and a distance over $\mathcal{H}$. We claim that, for arbitrary $h \in \mathcal{H}, f \in \mathcal{F}^{\text{pre}}$, let h' be ϵ_1-close to h in empirical l_2-norm w.r.t inputs $x_1^{\text{pre}}, x_2^{\text{pre}} \cdots, x_n^{\text{pre}}$. Given h', let f' be ϵ_2-close to f in empirical l_2 loss w.r.t $h'(x^{\text{pre}})$. Using the triangle inequality we have that

$$\begin{aligned}
d(f(h), f'(h')|x^{\text{pre}}) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|f(h(x_i^{\text{pre}})) - f'(h'(x_i^{\text{pre}}))\|} \\
&\leq d(f(h), f(h')|x^{\text{pre}}) + d(f(h'), f'(h')|x^{\text{pre}}) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|f(h(x_i^{\text{pre}})) - f(h'(x_i^{\text{pre}}))\|^2} + \epsilon_2 \\
&\leq L(\mathcal{F}^{\text{pre}})\sqrt{\frac{1}{n}\sum_{i=1}^{n} \|h(x_i^{\text{pre}}) - h'(x_i^{\text{pre}})\|^2} + \epsilon_2 \\
&= L(\mathcal{F}^{\text{pre}}) \cdot \epsilon_1 + \epsilon_2,
\end{aligned}$$

where we have used that $\|f(x) - f(y)\| \leq L(\mathcal{F}^{\text{pre}})\|x - y\|$ for any $f \in \mathcal{F}^{\text{pre}}$.

As for the cardinality of the covering $C_{\mathcal{F}^{\text{pre}}(\mathcal{H})}$. Observe $|C_{\mathcal{F}^{\text{pre}}(\mathcal{H})}| = \sum_{h \in C_{\mathcal{H}(x^{\text{pre}})}} |C_{\mathcal{F}_h^{\text{pre}}}| \leq |C_{\mathcal{H}(x^{\text{pre}})}| \cdot \max_{h \in \mathcal{H}(x^{\text{pre}})} |C_{\mathcal{F}^{\text{pre}}_{h(x^{\text{pre}})}}|$. This provides a bound on the metric entropy of

$$\log N(\epsilon_1 \cdot L(\mathcal{F}^{\text{pre}}) + \epsilon_2; \mathcal{F}^{\text{pre}}(\mathcal{H})|x^{\text{pre}}) \leq \log N(\epsilon_1; \mathcal{H}|x^{\text{pre}}) + \max_{h(x^{\text{pre}})} N(\epsilon_2; \mathcal{F}^{\text{pre}}|h \circ x^{\text{pre}}).$$

Applying the covering number upper bound with $\epsilon_1 = \frac{\epsilon}{2 \cdot L(\mathcal{F}^{\text{pre}})}, \epsilon_2 = \frac{\epsilon}{2}$ gives a bound of entropy integral ofa ,

$$\begin{aligned}
&\int_{\frac{\delta}{4}}^{D_{\mathcal{X}^{\text{pre}}}} \sqrt{\log N(u; \mathcal{F}^{\text{pre}}(\mathcal{H})|x^{\text{pre}})} du \\
&\leq \int_{\frac{\delta}{4}}^{D_{\mathcal{X}^{\text{pre}}}} \sqrt{\log N\left(\frac{u}{2L(\mathcal{F}^{\text{pre}})}; \mathcal{H}\middle|x^{\text{pre}}\right)} du + \int_{\frac{\delta}{4}}^{D_{\mathcal{X}^{\text{pre}}}} \max_{h \circ x^{\text{pre}}} \sqrt{\log N\left(\frac{u}{2}; \mathcal{F}^{\text{pre}}\middle|h \circ x^{\text{pre}}\right)} du
\end{aligned}$$

From the Sudakov minoration theorem (Wainwright, 2019, Theorem 5.30) for Gaussian processes and the fact that packing numbers at scale u upper bounds the covering number at scale $\forall u > 0$ we find:

$$\log N(u; \mathcal{H}|x^{\text{pre}}) \leq 4\left(\frac{\sqrt{n}\hat{G}_n(\mathcal{H})}{u}\right)^2, \quad \log N(u; \mathcal{F}^{\text{pre}}|h(x^{\text{pre}})) \leq 4\left(\frac{\sqrt{n}\hat{G}_n(\mathcal{F}^{\text{pre}}|h \circ x^{\text{pre}})}{u}\right)^2.$$

Combining the definition of worst-case Gaussian complexity with the aforementioned results we have

$$\hat{G}_n(\mathcal{F}^{\text{pre}} \circ \mathcal{H}) \leq 2\sqrt{k-1}\delta + 256 \log \frac{4D_{\mathcal{X}^{\text{pre}}}}{\delta} \left(L(\mathcal{F}^{\text{pre}})\hat{G}_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\text{pre}})\right),$$

substitute δ with $\frac{4D_{\mathcal{X}^{\text{pre}}}}{n^2}$, proof is completed. □

Downstream learning Next we turn to the second stage and come up with theoretical guarantees by using inexact $\hat{h}$ learned from the first stage.

Theorem 6. *In the downstream task, we have that with probability at least* $1-\delta$,

$$R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h) \leq d_{\mathcal{F}^{\text{down}}}(\hat{h}; h) + 4\sqrt{\pi} L^{\text{down}} \cdot \bar{G}_m(\mathcal{F}^{\text{down}}) + 4B^{\text{down}}\sqrt{\frac{\log(2/\delta)}{m}}$$

Proof. Assumption 1 implies

$$\mathbb{E}_{x^{\text{down}}, y^{\text{down}}}\left[\ell\left(g^{\text{down}}\left(x^{\text{down}}\right), y^{\text{down}}\right)\right] = R_{\text{down}}(f^{\text{down}}, h).$$

To start, let $\tilde{f} = \arg\min_{f\in\mathcal{F}^{\text{down}}} R_{\text{down}}(f, \hat{h})$ and $R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h)$ equals

$$\left[R_{\text{down}}(\tilde{f}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h)\right] + \left[R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - R_{\text{down}}(\tilde{f}, \hat{h})\right]$$

where the first term satisfies

$$\begin{aligned}
&\inf_{\tilde{f}\in\mathcal{F}^{\text{down}}} \left[R_{\text{down}}(\tilde{f}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h)\right] \\
&\leq \sup_{f^{\text{down}}\in\mathcal{F}^{\text{down}}} \inf_{\tilde{f}\in\mathcal{F}^{\text{down}}} \left[R_{\text{down}}(\tilde{f}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h)\right] \\
&= d_{\mathcal{F}^{\text{down}}}(\hat{h}, h)
\end{aligned}$$

The second term follows the similar lines of Theorem 5

$$R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - R_{\text{down}}(\tilde{f}, \hat{h}) \leq 4\sqrt{\pi} L^{\text{down}} \mathbb{E}_{\mathcal{X}^{\text{down}}} \hat{G}_m(\mathcal{F}^{\text{down}} | \hat{h} \circ x^{\text{down}}) + 4B^{\text{down}}\sqrt{\frac{\log(1/\delta)}{m}}$$

Again we make use of the worst-case argument

$$\mathbb{E}_{\mathcal{X}^{\text{down}}} \hat{G}_m(\mathcal{F}^{\text{down}} | \hat{h} \circ x^{\text{down}}) \leq \bar{G}_m(\mathcal{F}^{\text{down}}).$$

Combining the results gives the statement. □

Proof for main Theorem 1 Having introduced class diversity parameter, proof is directly completed via combination of Theorem 5 and Theorem 6.

A.2 Proofs for Theorem 2

We could provide a better dependence on the boundedness noise parameters in Theorem 6 using Bernstein inequality. We present the following corollary which has data-dependence in the Gaussian complexities.

Corollary 1. *Presuming Assumption 1 holds, we have that then with probability at least* $1-\delta$,

$$\begin{aligned}
&R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h) \\
&\leq d_{\mathcal{F}^{\text{down}}}(\hat{h}; h) + 4\sqrt{\pi} L^{\text{down}} \cdot \mathbb{E}_{\mathcal{X}^{\text{down}}} \hat{G}_m(\mathcal{F}^{\text{down}} | \hat{h} \circ x^{\text{down}}) + 4\sigma\sqrt{\frac{\log(2/\delta)}{m}} + 50B^{\text{down}}\frac{\log(2/\delta)}{m}
\end{aligned}$$

Proof. Denote $Z = \sup_f |\hat{R}_{\text{down}}(f, \hat{h}) - R_{\text{down}}(f, \hat{h})|$, we apply the functional Bernstein inequality (Massart, 2000, Theorem 3) to control the fluctuations. With probability at lest $1-\delta$, we have

$$Z \leq 2\mathbb{E}[Z] + 4\frac{\sigma}{\sqrt{m}}\sqrt{\log(\frac{1}{\delta})} + 35\frac{B^{\text{down}}}{m}\log(\frac{1}{\delta}), \tag{20}$$

where $\sigma^2 = \frac{1}{m}\sup_f \sum_{i=1}^m Var(\ell(f\circ \hat{h}(x_i^{\text{down}}), y_i^{\text{down}}))$. Thus

$$\begin{aligned}
\mathbb{E}[Z] &\leq 2\mathbb{E}_{\mathcal{X}^{\text{down}}} \hat{R}_m(l(\mathcal{F}^{\text{down}}) | \hat{h} \circ x^{\text{down}}) \\
&\leq 2\mathbb{E}_{\mathcal{X}^{\text{down}}} \sqrt{2} L^{\text{down}} \hat{R}_m(\mathcal{F}^{\text{down}} | \hat{h} \circ x^{\text{down}}) \\
&\leq 2\mathbb{E}_{\mathcal{X}^{\text{down}}} \sqrt{\pi} L^{\text{down}} \hat{G}_m(\mathcal{F}^{\text{down}} | \hat{h} \circ x^{\text{down}}),
\end{aligned}$$

where the second line uses vector-based contraction principle, the last line upper bounds the empirical Rademacher complexity by Gaussian counterparts. □

Proof for Theorem 2 Observe that

$$\ell(\eta; y) = -y^\top \eta + \log(1 + \sum_{s=1}^{k-1} e^{\eta_s}), \ell(\eta; y) \le \|\eta\|$$

and

$$\left|\frac{\partial \ell(\eta; y)}{\partial \eta_i}\right| = \left|y_i - \frac{e^{\eta_i}}{1 + \sum_{s=1}^{k-1} e^{\eta_s}}\right|,$$

$$|\nabla_\eta \ell(\eta; y)| \le \sqrt{k-1},$$

so it is $L^{\text{pre}} = \sqrt{k-1}$–Lipschitz. By definition the class $\mathcal{F}^{\text{pre}}$ with parameters $\|\alpha_s\|_2 \le \mathcal{O}(1), s \in [k-1]$, we obtain that $L(\mathcal{F}^{\text{pre}}) = \mathcal{O}\left(\sqrt{k-1}\right)$ since for any $x, y \in \mathbb{R}^r$, any $f \in \mathcal{F}^{\text{pre}}$ we have

$$\begin{aligned}
\|f(x) - f(y)\|^2 &= \|\alpha^\top x - \alpha^\top y\|^2 \\
&\le \sum_{s=1}^{k-1} \left(\langle \alpha_s, x - y\rangle\right)^2 \\
&\le \sum_{s=1}^{k-1} \|\alpha_s\|^2 \|x - y\|^2 \\
&\le c_1^2 (k-1) \|x - y\|^2
\end{aligned}$$

In conclusion we have

- Pre-training loss $\ell(\cdot, y^{\text{pre}})$ is $\sqrt{k-1}$-Lipschitz and $B^{\text{pre}} = D_{\mathcal{X}^{\text{pre}}}$-bounded.
- Downstream loss $\ell(\cdot, y^{\text{down}})$ is $\sqrt{k'-1}$-Lipschitz and $B^{\text{down}} = D_{\mathcal{X}^{\text{down}}}$-bounded.
- Linear layer f is $L(\mathcal{F}^{\text{pre}}) = O\left(\sqrt{k-1}\right)$-Lipschitz.

Consider task-specific function classes for characterizing class-diversity parameters. From Lemma 2 and Lemma 3 we know that

$$\nu = \Omega(\tilde{\nu}), \quad \tilde{\nu} = \sigma_r(\alpha_1 \alpha_1^\top).$$

Combining these pieces of results then the proof is completed.

A.3 WHAT IS *diversity parameter* FOR LINEAR LAYERS?

In this subsection, we lower bound the diversity parameter ν by $\tilde{\nu}$. We note the proof of this part is very different from the multi-task setting studied in previous works (Du et al., 2021; Tripuraneni et al., 2020b).

First, we demonstrate cross entropy loss between learned representation class and true underlying representation could be bounded by quadratic loss,

Lemma 2. *Suppose ℓ is a $(k-1)$-class cross entropy loss and y^{pre} is $(k-1)$-onehot label. The cross entropy can be bounded from both sides with quadratic loss,*

$$\begin{aligned}
&c_0 \mathbb{E}_{x^{\text{pre}}} \left[\exp(-10 \max(\|\alpha'^\top \hat{h}(x^{\text{pre}})\|, \|\alpha^\top h(x^{\text{pre}})\|)) \left\|\alpha'^\top \hat{h}(x^{\text{pre}}) - \alpha^\top h(x^{\text{pre}})\right\|^2\right] \\
&\le \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}}[\ell(\hat{f}^{\text{pre}} \circ \hat{h}(x^{\text{pre}})) - \ell(f^{\text{pre}} \circ h(x^{\text{pre}}), y^{\text{pre}})] \\
&\le \frac{1}{2} \mathbb{E}_{x^{\text{pre}}} \left\|\alpha'^\top \hat{h}(x^{\text{pre}}) - \alpha^\top h(x^{\text{pre}})\right\|^2,
\end{aligned}$$

where $c_0 = \frac{1}{2}\lambda_{min}(\Phi''(\alpha^\top h(x^{\text{pre}})))$ denote second-order derivative of softmax function, α' and α are parameters for $\hat{f}^{\text{pre}}$ and f^{pre} respectively.

Proof. The conditional distribution of $(k-1)$-class classification task is defined as:

$$\begin{aligned}
\mathcal{P}(\cdot | \alpha^\top h(x)) &= e^{y^\top \alpha^\top h(x) - \Phi(\alpha^\top h(x))} \\
&= e^{y^\top \alpha^\top h(x) - \log(1 + \sum_{s=1}^{k-1} e^{(\alpha^\top h(x))_s})}
\end{aligned}$$

Taylor's expansion tells that

$$\Phi(\alpha'^\top \hat{h}(x)) = \Phi(\alpha^\top h(x)) + \nabla\Phi(\alpha^\top h(x))^\top (\alpha'^\top \hat{h}(x) - \alpha^\top h(x)) + o\left(\|\alpha'^\top \hat{h}(x) - \alpha^\top h(x)\|^2\right)$$

For generalized linear models: $\mathcal{P}(\cdot|\alpha^\top h(x)) = \exp\left(y^\top \alpha^\top h(x) - \Phi(\alpha^\top h(x))\right)$, it is known that

$$\begin{aligned} & KL\left[\mathcal{P}(\cdot|\alpha^\top h(x)), \mathcal{P}(\cdot|\alpha'^\top \hat{h}(x))\right] \\ &= \Phi(\alpha'^\top \hat{h}(x)) - \Phi(\alpha^\top h(x)) - \nabla\Phi(\alpha^\top h(x))^\top (\alpha'^\top \hat{h}(x) - \alpha^\top h(x)) \end{aligned}$$

Our goal is to bound $KL-$divergence with quadratic distance,

$$\frac{\mu}{2}\|\alpha'^\top \hat{h}(x) - \alpha^\top h(x)\|^2 \leq KL\left[\mathcal{P}(\cdot|\alpha^\top h(x)), \mathcal{P}(\cdot|\alpha'^\top \hat{h}(x))\right] \leq \frac{L}{2}\|\alpha'^\top \hat{h}(x) - \alpha^\top h(x)\|^2. \tag{21}$$

Here $L = \sup_{x'} \lambda_{max} \nabla^2\Phi(x')$ for intermediate $x' \in \left[\alpha'^\top \hat{h}(x), \alpha^\top h(x)\right]$. μ is some constant related to $\alpha'^\top \hat{h}(x)$ and $\alpha^\top h(x)$ which requires more investigations.

First we show L which could be considered as upper bound of the maximum eigenvalue of multinomial logistic regression. Restate this sub-problem for clarification: for $\Phi(x) = \log\left(1 + \sum_{s=1}^{k-1} e^{x_s}\right)$, its gradient at i^{th}-coordinate is $\frac{\partial \Phi}{\partial x_i} = e^{x_i}/1+\sum_s e^{x_s}$, so Hessian matrix is computed as

$$\frac{\partial^2 \Phi}{\partial x_i \partial x_j} = \begin{cases} \frac{e^{x_i}\cdot(1+\sum_{s\neq i} e^{x_s})}{(1+\sum_s e^{x_s})^2}, & i = j \\ \\ \frac{-e^{x_i}e^{x_j}}{(1+\sum_s e^{x_s})^2}, & i \neq j. \end{cases} \tag{22}$$

Our goal is to bound the smallest and largest eigenvalues of this Hessian matrix while $x \in [x_1, x_2]$. Next we show its largest eigenvalue is upper bounded by 1. Denote $\sigma(x) = \frac{1}{1+\sum_s e^{x_s}}\left[e^{x_1}, \cdots, e^{x_{k-1}}\right]^\top$, the Hessian matrix can be restate as

$$\nabla^2\Phi = diag(\sigma(x)) - \sigma(x)\sigma(x)^\top.$$

For any non-zero vector y, we have

$$\begin{aligned} y^\top \nabla^2\Phi y &= \sum_i \sigma(x)_i y_i^2 - \left(\sigma(x)^\top y\right)^2 \\ &\leq \max(\sigma(x)_i)\|y\|^2 \\ &\leq \|y\|^2, \end{aligned}$$

which implies its largest eigenvalue is no bigger than 1. Also, this Hessian is strictly-convex, from Cauchy-Schwarz inequality we know

$$\begin{aligned} \left(\sigma(x)^\top y\right)^2 &\leq \sum_i \sigma(x)_i y_i^2 \cdot \sum_i \sigma(x)_i \\ &< \sum_i \sigma(x)_i y_i^2, \end{aligned}$$

so this matrix is positive definite.

The derivations of bounding second order remaining terms of second-order Taylor's expansion from below are more complicated. Note that, logistic regression is not globally strongly-convex (so as multinomial logistic regression). However, when landscape is considered locally (e.g., here $x' \in \left[\alpha'^\top \hat{h}(x), \alpha^\top h(x)\right]$), a class of convex functions called self-concordant functions would present benign local properties. Next we introduce *self-concordant* functions and related useful properties.

Definition 6 (Modified self-concordance). *F: $\mathbb{R}^p \mapsto \mathbb{R}$ is three times differentiable convex function such that for some $R > 0$, for all $u, v \in \mathbb{R}^p$, the function $g : t \mapsto F(u + tv)$ satisfies for all $t \in \mathbb{R}$*

$$|g'''(t)| \leq R\|v\|_2 \times g''(t) \tag{23}$$

Proposition 1. *Multinomial logistic regression satisfies* self-concordance *condition, namely:* $\Phi(x) = \log(1+\sum_{s=1}^{k-1} e^{x_s}), x \in \mathbb{R}^{k-1}$ *is convex, for all* $u, v \in \mathbb{R}^{k-1}$*, the function* $g : t \mapsto \Phi(u+tv)$ *satisfies*

$$|g'''(t)| \leq 5\|v\|_2 g''(t) \tag{24}$$

Proof. Let $P(t; v^0) = 1 + \sum_s e^{u_s+tv_s}$ and $P(t; v^i) = \sum_s v_s^i e^{u_s+tv_s}, i > 1$. Then we use multinomials P to represent derivatives of $g(t)$

$$\begin{aligned} g(t) &= \log(P(T; V^0)) \\ g'(t) &= \frac{P(t; v^1)}{P(t; v^0)} \\ g''(t) &= \frac{P(t; v^2)P(t; v^0)}{P(t; v^0)^2} \\ g'''(t) &= \frac{P(t; v^3)P(t; v^0)^2 - 3P(t; v^2)P(t; v^1)P(t; v^0) + 2P(t; v^1)^3}{P(t; v^0)^3} \end{aligned}$$

□

Let $r_s = e^{u_s+v_s t}$, hence

$$\begin{aligned} g''(t) &= \frac{(\sum_s v_s^2 r_s) \cdot (1 + \sum_s r_s) - (\sum_s v_s r_s)^2}{(1 + \sum_s r_s)^2} \\ &= \frac{\sum_{i<j} r_i r_j (v_i - v_j)^2 + \sum_i v_i^2 r_i}{(1 + \sum_s r_s)^2} \end{aligned}$$

In the following we expand $g'''(t)$ as:

$$\begin{aligned} &\frac{\sum_{i<j} r_i r_j (v_i - v_j)^2 [\sum_k (v_i + v_j - 2v_k) r_k] + \sum_i v_i^3 r_i + \sum_i \sum_j v_i^2 r_i r_j (2v_i - 3v_j)}{(1 + \sum_s r_s)^3} \\ = &\frac{\sum_{i<j} r_i r_j (v_i - v_j)^2 [\sum_k (v_i + v_j - 2v_k) r_k] + \sum_i v_i^2 r_i \left(v_i (1 + 2\sum_j r_j) - 3\sum_j v_j r_j\right)}{(1 + \sum_s r_s)^3}, \end{aligned}$$

observe that

$$\frac{1}{1 + \sum_s r_s} \left| \sum_k (v_i + v_j - 2v_k) r_k \right| \leq \sum_k |v_i + v_j - 2v_k| \frac{r_k}{1 + \sum_s r_s} \leq 4\|v\|_2$$

$$\frac{1}{1 + \sum_s r_s} \left| v_i (1 + 2\sum_j r_j) - 3\sum_j v_j r_j \right| \leq 5\|v\|_2$$

Substitute these into definition of *self-concordance* then proof is completed.

Self-concordance properties Self-concordance gives nice characterizations of local curvature of convex functions which plays important role in describing local convexity (Bach, 2014). Some useful results are given upon this condition (see (Bach et al., 2010, Proposition 1)), we list out the three main inequalities as below

For all $w, v \in \mathbb{R}^p, t \in \mathbb{R}$, we have that,

$$F(w + v) \geq F(w) + vF'(w) + \frac{v^\top F''(w)v}{R^2\|v\|_2^2} \cdot \left(e^{-R\|v\|_2} + R\|v\|_2 - 1\right) \tag{25}$$

$$F(w + v) \leq F(w) + vF'(w) + \frac{v^\top F''(w)v}{R^2\|v\|_2^2} \cdot \left(e^{R\|v\|_2} - R\|v\|_2 - 1\right) \tag{26}$$

$$e^{-tR\|v\|_2} F''(w) \preceq F''(w + tv) \preceq e^{tR\|v\|_2} F''(w) \tag{27}$$

Above results give two refined versions of Taylor's expansion along with an upper and lower zero-order Taylor expansion of F''. With these we are ready to give lower bound of KL divergence (see Eq 21), i.e.,

$$\begin{aligned}
&\Phi(\alpha'^\top \hat{h}(x)) - \Phi(\alpha^\top h(x)) - \nabla\Phi(\alpha^\top h(x))^\top (\alpha'^\top \hat{h}(x) - \alpha^\top h(x)) \\
\geq\ & \frac{1}{2} v^\top e^{-tR\|v\|_2} F''(\alpha^\top h(x)) v \\
\geq\ & \frac{1}{2}\lambda_{min}(\Phi''(\alpha^\top h(x)))\|v\|^2 e^{-5\|v\|_2} \\
\geq\ & \frac{1}{2}\lambda_{min}(\Phi''(\alpha^\top h(x)))\|v\|^2 e^{-5(\|\alpha'^\top \hat{h}(x)\| + \|\alpha^\top h(x)\|)} \\
\geq\ & \frac{1}{2}\lambda_{min}(\Phi''(\alpha^\top h(x)))\|v\|^2 \exp(-10\max(\|\alpha'^\top \hat{h}(x)\|, \|\alpha^\top h(x)\|))
\end{aligned}$$

where $v = \alpha'^\top \hat{h}(x) - \alpha^\top h(x)$, t is an parameter exists in $[0, 1]$.

Proof for Lemma 2 is completed. □

Lemma 3. *With Lemma 2 at hand and define $C = \alpha_1\alpha_1^\top$, it is demonstrated that*

$$d_{\mathcal{F}^{\text{down}}}(\hat{h}; h) \leq \frac{1}{\Omega(\tilde{\nu})} d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(\hat{h}; h), \quad \tilde{\nu} = \sigma_r(C) \tag{28}$$

Proof. For function classes $\mathcal{F}^{\text{pre}}$ and $\mathcal{F}^{\text{down}}$ such that $f^{\text{down}} \in \mathcal{F}^{\text{down}}$ and data $(x^{\text{down}}, y^{\text{down}}) \sim \mathbb{P}_{f^{\text{down}} \circ h}$ which is the real underlying distribution, the worst-case representation difference is similar to that in multi-task analysis (Tripuraneni et al., 2020b, Lemma 1):

$$\begin{aligned}
d_{\mathcal{F}^{\text{down}}}(\hat{h}; h) &= \sup_{f^{\text{down}} \in \mathcal{F}^{\text{down}}} \inf_{f' \in \mathcal{F}^{\text{down}}} \mathbb{E}\left\{\ell(f' \circ \hat{h}(x^{\text{down}}), y^{\text{down}}) - \ell(f^{\text{down}} \circ h(x^{\text{down}}), y^{\text{down}})\right\} \\
&\leq \sup_{\|\alpha_s\| \leq c_0} \inf_{\|\alpha'_s\| \leq c_0} \frac{1}{2}\mathbb{E}_{\mathcal{X}^{\text{down}}} \left\|\alpha'^\top \hat{h}(x^{\text{down}}) - \alpha^\top h(x^{\text{down}})\right\|^2, \quad \text{here } s \in [k'-1] \\
&= \sum_{s=1}^{k'-1} \sup_{\|\alpha_s\| \leq c_0} \inf_{\|\alpha'_s\| \leq c_0} \frac{1}{2}\mathbb{E}_{\mathcal{X}^{\text{down}}} \left(\alpha'^\top_s \hat{h}(x^{\text{down}}) - \alpha_s^\top h(x^{\text{down}})\right)^2 \\
&\leq (k'-1)\frac{c_0^2}{2}\sigma_1(\Lambda_{sc}(\hat{h}, h)).
\end{aligned}$$

Here in the last line, the inner infima is considered as the partial minimization of a convex quadratic form (see (Boyd & Vandenberghe, 2004, Example 3.15, Appendix A.5.4)).

Define population covariance if representations $\hat{h}$ and h as

$$\Lambda(\hat{h}, h) = \begin{bmatrix} \mathbb{E}[\hat{h}(x)\hat{h}(x)^\top] & \mathbb{E}[\hat{h}(x)h(x)^\top] \\ \mathbb{E}[h(x)\hat{h}(x)^\top] & \mathbb{E}[h(x)h(x)^\top] \end{bmatrix} = \begin{bmatrix} F_{\hat{h}\hat{h}} & F_{\hat{h}h} \\ F_{h\hat{h}} & F_{hh} \end{bmatrix}$$

$\Lambda_{sc}(\hat{h}, h) = F_{hh} - F_{h\hat{h}}(F_{\hat{h}\hat{h}})^\dagger F_{\hat{h}h}$ is the generalized Schur complement of h with respect to $\hat{h}$.

Controlling the pre-training representation difference with the lower bound is subtler,

$$\begin{aligned}
&d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(\hat{h}; h) \\
=\ & \inf_{f' \in \mathcal{F}^{\text{pre}}} \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}}\{\ell(f' \circ \hat{h}(x^{\text{pre}}), y^{\text{pre}}) - \ell(f^{\text{pre}} \circ h(x^{\text{pre}}), y^{\text{pre}})\} \\
=\ & \inf_{\|\alpha'_s\| \leq c_1} \mathbb{E}_{x^{\text{pre}}, y^{\text{pre}}} \left[\ell(f' \circ \hat{h}(x^{\text{pre}}), y^{\text{pre}}) - \ell(f^{\text{pre}} \circ h(x^{\text{pre}}), y^{\text{pre}})\right], \quad s \in [k-1] \\
\geq\ & c_0\mathbb{E}_{\mathcal{X}^{\text{pre}}} \left[\exp(-10\max(\|\alpha'^\top \hat{h}(x^{\text{pre}})\|, \|\alpha^\top h(x^{\text{pre}})\|)) \cdot \left\|\alpha'^\top \hat{h}(x^{\text{pre}}) - \alpha^\top h(x^{\text{pre}})\right\|^2\right].
\end{aligned}$$

We lower bound each term in the sum over j identically and suppress the j for each of notation. Define event E as

$$\mathbf{1}[E] = \mathbf{1}\left\{\begin{array}{ll} |Z_1^s| = |(\alpha^\top B x^{\mathrm{pre}})_s| \le Cc'\|\alpha_s\|, & s = 1, 2, \cdots, k-1, \\ \\ |Z_2^s| = |({\alpha'}^\top \hat{B} x^{\mathrm{pre}})_s| \le Cc'\|{\alpha'}_s\|, & s = 1, 2, \cdots, k-1. \end{array}\right\},$$

where $Z_1^s \sim \mathrm{subG}(\|\alpha_s\|^2 C^2)$ and $Z_2^s \sim \mathrm{subG}(\|{\alpha'}_s\|^2 C^2)$. Then we use this event to bound the diversity,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X}^{\mathrm{pre}}} \exp(-10 \max(\|{\alpha'}^\top \hat{h}(x^{\mathrm{pre}})\|, \|\alpha^\top h(x^{\mathrm{pre}})\|)) \cdot \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \\
\ge& \mathbb{E}_{\mathcal{X}^{\mathrm{pre}}} \left[\mathbf{1}(E) \exp(-10 \max(\|{\alpha'}^\top \hat{h}(x^{\mathrm{pre}})\|, \|\alpha^\top h(x^{\mathrm{pre}})\|)) \cdot \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right] \\
\ge& e^{-10Cc'c_1 \cdot \sqrt{k-1})} \mathbb{E}_{\mathcal{X}^{\mathrm{pre}}} \left(\mathbf{1}(E) \cdot \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right).
\end{aligned}$$

First write $E_{x^{\mathrm{pre}}} \left(\mathbf{1}(E) \cdot \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right)$ as

$$\mathbb{E}_{x^{\mathrm{pre}}} \left(\left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right) - \mathbb{E}_{x^{\mathrm{pre}}} \left(\mathbf{1}(E^c) \cdot \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right).$$

Observe that for the last term, it holds

$$\begin{aligned}
&\mathbb{E}_{x^{\mathrm{pre}}} \left(\mathbf{1}(E^c) \cdot \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right) \\
=& \sum_{s=1}^{k-1} \mathbb{E}_{x^{\mathrm{pre}}} \left(\mathbf{1}(E^c) \cdot \left[{\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right]_s^2 \right) \\
\le& \sum_{s=1}^{k-1} \sqrt{P[E^c]} \sqrt{\mathbb{E}_{x^{\mathrm{pre}}} \left[{\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right]_s^4}
\end{aligned}$$

Using union bound we have that $P[E^c] \le \frac{2(k-1)}{c'^2}$. Using L4-L2 hyper-contractivity we also conclude

$$\sqrt{\mathbb{E}_{x^{\mathrm{pre}}} \left({\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right)_s^4} \le 10\sigma^2 = 10 \mathbb{E}_{x^{\mathrm{pre}}} \left({\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right)_s^2.$$

Combining these we know

$$\mathbb{E}_{x^{\mathrm{pre}}} \left[\mathbf{1}(E^c) \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right] \le \frac{10\sqrt{2(k-1)}}{c'} \mathbb{E}_{x^{\mathrm{pre}}} \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2.$$

Hence this metric could be claimed to be lower bounded as,

$$\begin{aligned}
&\Omega \left(\inf_{\alpha'} \mathbb{E}_{x^{\mathrm{pre}}} \left\| {\alpha'}^\top \hat{h}(x^{\mathrm{pre}}) - \alpha^\top h(x^{\mathrm{pre}}) \right\|^2 \right) \\
=& \Omega \left(\alpha_1^\top \Lambda_{sc}(\hat{h}; h) \alpha_1 \right) \\
=& \Omega \left(tr(\Lambda_{sc}(\hat{h}; h) C) \right), \quad \text{where } C = \alpha_1 \alpha_1^\top.
\end{aligned}$$

In the second line, we redefine α_1 as parameter α of pre-training for clarity. In this way we conclude that,

$$d_{\mathcal{F}^{\mathrm{pre}}, f^{\mathrm{pre}}}(\hat{h}; h) = \Omega \left(tr(\Lambda_{sc}(\hat{h}, h) C) \right) = \Omega \left(\sigma_1(\Lambda_{sc}(\hat{h}, h)) \sigma_r(C) \right),$$

where C implies expansion of representation $h(x) \in \mathbb{R}^r$, and its condition number $\sigma_r(C)$ indicates how spread out this vector is in $\mathbb{R}^r$:

$$C = \sum_{s=1}^{k-1} (\alpha_1)_s (\alpha_1)_s^\top = \alpha_1 \alpha_1^\top, \quad \alpha_1 \in \mathbb{R}^{r \times (k-1)}$$

Aforementioned calculations show

$$d_{\mathcal{F}^{\mathrm{down}}}(\hat{h}; h) \le \frac{1}{\Omega(\tilde{\nu})} d_{\mathcal{F}^{\mathrm{pre}}, f^{\mathrm{pre}}}(\hat{h}; h), \quad \tilde{\nu} = \sigma_r(C).$$

Proof is completed. □

A.4 Proofs for Linear Subspace Representation

Proof. We begin with bounding each of the complexity terms in the Corrolary 1.

We make use of data-dependent inequalities (Tripuraneni et al., 2020b, Lemma 4) to help upper bound related quantities. Intuitively Definition 5 implies tail-bound properties in a sub-gaussian process.

- $$\begin{aligned}\hat{G}_n(\mathcal{H}) &= \frac{1}{n}\mathbb{E}\left[\sup_{B\in\mathcal{H}}\sum_{k=1}^{r}\sum_{i=1}^{n} g_{ki} b_k^\top x_i^{\text{pre}}\right] \\ &= \mathcal{O}\left(\sqrt{\frac{dr^2}{n}}\right)\end{aligned}$$

- $$\begin{aligned}\hat{G}_n(\mathcal{F}^{\text{pre}}|h\circ x^{\text{pre}}) &= \frac{1}{n}\mathbb{E}\left[\sup_{\alpha_1,\cdots,\alpha_{k-1}}\sum_{s=1}^{k-1}\sum_{i=1}^{n} g_{is}\alpha_s^\top B^\top x_i^{\text{pre}}\right] \\ &= \frac{c_1(k-1)}{n}\mathbb{E}\|\sum_{i=1}^{n} g_{is}B^\top x_i^{\text{pre}}\| \\ &= \frac{c_1(k-1)}{\sqrt{n}}\sqrt{tr(B^\top\Sigma B)}\end{aligned}$$

  then $\bar{G}_n(\mathcal{F}^{\text{pre}}) \le \mathcal{O}\left((k-1)\sqrt{\frac{r}{n}}\right)$.

- Similarly,

  $$\hat{G}_m(\mathcal{F}^{\text{down}}|h\circ x^{\text{down}}) \le \frac{c_1(k'-1)}{\sqrt{m}}\sqrt{\sum_{i=1}^{r}\sigma_i(\hat{B}^\top\Sigma\hat{B})}$$

  then $\bar{G}_m(\mathcal{F}^{\text{down}}) \le \mathcal{O}\left((k'-1)\sqrt{\frac{r}{m}}\right)$.

- boundedness parameter $D_{\mathcal{X}^{\text{pre}}} = \sup_{\alpha,B}\|\alpha^\top B^\top x\| = \mathcal{O}(\sqrt{k-1}D)$

- cross entropy $\ell(\eta;y) = -y^\top\eta + \log\left(1+\sum_{s=1}^{k-1} e^{\eta_s}\right)$, then $\left|\frac{\partial\ell(\eta;y)}{\partial\eta_i}\right| = \left|y_i - \frac{e^{\eta_i}}{1+\sum_{s=1}^{k-1} e^{\eta_s}}\right|$, $|\nabla_\eta\ell(\eta;y)| \le \sqrt{k-1}$, so it is $L^{\text{pre}} = \sqrt{k-1}$–Lipschitz in its first coordinate uniformly over its second for pre-training and $L^{\text{down}} = \sqrt{k'-1}$–Lipschitz for downstream task.

- $|\ell(\eta;y)| \le \mathcal{O}(\|\eta\|)$, where $\|\eta\| = \|x^\top B^{\text{pre}}\alpha\| \lesssim \sqrt{k-1}D$. This means $B^{\text{pre}} = \mathcal{O}(\sqrt{k-1}D)$, $B^{\text{down}} = \mathcal{O}(\sqrt{k'-1}D)$.

In Corollary 1, we define and compute the maximal variance term σ^2 as,

$$\begin{aligned}\sigma^2 &= \frac{1}{m}\sup_{f^{\text{down}}\in\mathcal{F}^{\text{down}}}\sum_{i=1}^{m} Var(\ell'(f^{\text{down}}\circ\hat{h}(x_i^{\text{down}}), y_i^{\text{down}})) \\ &\le \frac{k'-1}{m}\sup_{f^{\text{down}}\in\mathcal{F}^{\text{down}}}\sum_{i=1}^{m} Var(f^{\text{down}}\circ\hat{h}(x_i^{\text{down}})) \\ &= \frac{k'-1}{m}\sup_{\|\alpha_s\|\le\mathcal{O}(1)}\sum_{s=1}^{k'-1}\sum_{i=1}^{m} Var(\alpha_s^\top\hat{B}^\top x_i^{\text{down}}) \\ &= \frac{(k'-1)^2}{m}\sup_{\|\alpha_s\|\le\mathcal{O}(1)}\sum_{i=1}^{m}(\alpha_s\hat{B})^\top\Sigma\hat{B}\alpha_s \\ &= (k'-1)^2\mathcal{O}(\|\hat{B}\Sigma\hat{B}\|_2) \\ &= \mathcal{O}\left((k'-1)^2\right)\end{aligned}$$

With these results in hand, we are now prepared to apply Corollary 1, w.p. at least $1-\delta$

$$R_{\text{down}}(\hat{f}^{\text{down}}, \hat{h}) - R_{\text{down}}(f^{\text{down}}, h)$$
$$\leq d_{\mathcal{F}^{\text{down}}}(\hat{h}; h) + 4\sqrt{\pi}L^{\text{down}} \cdot \mathbb{E}_{\mathcal{X}^{\text{down}}}\hat{G}_m\left(\mathcal{F}^{\text{down}}\middle|\hat{h}\circ x^{\text{down}}\right) + 4\sigma\sqrt{\frac{\log(2/\delta)}{m}} + 50B^{\text{down}}\frac{\log(2/\delta)}{m}$$

where $\hat{G}_m(\mathcal{F}^{\text{down}}|\hat{h}\circ x^{\text{down}})$ is defined in Theorem 2.

Thus $L^{\text{down}} \cdot \mathbb{E}_{\mathcal{X}^{\text{down}}}\hat{G}_m\left(\mathcal{F}^{\text{down}}\middle|\hat{h}\circ x^{\text{down}}\right) \leq L^{\text{down}}\bar{G}_m(\mathcal{F}^{\text{down}}) \leq \mathcal{O}((k'-1)^{\frac{3}{2}}\sqrt{\frac{r}{m}})$, $\sigma \leq \mathcal{O}(k'-1)$, and $B^{\text{down}} \leq \mathcal{O}(\sqrt{k'-1}D)$. Further, we obtain upper bound of worst-case representation difference by diversity parameter and adoption of Theorem 5: w.p. at least $1-\delta$

$$d_{\mathcal{F}^{\text{down}}}(\hat{h}; h)$$
$$\leq \frac{d_{\mathcal{F}^{\text{pre}}, f^{\text{pre}}}(\hat{h}; h)}{\nu}$$
$$\leq \frac{1}{\nu}\left\{4096L\left[\log(n)\cdot[L(\mathcal{F}^{\text{pre}})\cdot G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\text{pre}})] + \frac{\sqrt{k-1}D_{\mathcal{X}^{\text{pre}}}}{n^2}\right] + 4B\sqrt{\frac{\log(2/\delta)}{n}}\right\}$$
$$\lesssim \frac{\sqrt{k}}{\nu}\left\{\log(n)\left(\sqrt{\frac{kdr^2}{n}} + k\sqrt{\frac{r}{n}}\right) + D\left(\frac{k}{n^2} + \sqrt{\frac{\log(1/\delta)}{n}}\right)\right\}$$

The last thing to consider for completing the proof for Theorem 3 is giving accurate characterization of diversity parameter ν, which we leave for the next subsection. □

A.5 Proofs for Deep Neural Network Representation

Proof. In deep neural network, we first review complexity quantities. Adapted from Theorem 8 (Golowich et al., 2018), we have

$$\hat{R}_n(\mathcal{N}) \leq \left(\frac{2}{n}\Pi_{p=1}^{K}M(p)\right)\sqrt{(K+1+\log d)\cdot\max_{j\in[d]}\sum_{i=1}^{n}x_{i,j}^2}$$
$$\leq \frac{2D\sqrt{K+1+\log d}\cdot\Pi_{p=1}^{K}M(p)}{\sqrt{n}}.$$

where $x_{i,j}$ denotes the j-th coordinate of vector x_i.

Then we proceed to bound the Gaussian complexities for our deep neural network and prove Theorem 4. Recall that under the conditions of the result we can use former results to verify the task diversity condition is satisfied with parameters $\Omega(\tilde{\nu})$ with $\tilde{\nu} = \sigma_r(\alpha_1\alpha_1^\top) > 0$. We can see that $\|\mathbb{E}_x[\hat{h}(x)h^*(x)^\top\|_2 \leq \mathbb{E}x\|\hat{h}(x)h^*(x)\| \leq O(M(K)^2)$ using the norm bound from. Hence under this setting we can choose c_1 sufficiently large so that $c_1M(K)^2 \gtrsim \frac{M(K)^2}{c}c_2$. The condition $M(K) \gtrsim 1$ in the theorem statement is simply used to clean up the final bound.

In order to instantiate Theorem 1 we begin by bounding each of the complexity terms in the expression.

- For the feature learning complexity in the training phase, we leverage above results, then

$$\hat{G}_n(\mathcal{H}) = \frac{1}{n}\mathbb{E}[\sup_{\mathcal{W}_K}\sum_{k=1}^{r}\sum_{i=1}^{n}g_{ki}h_k(x_i^{\text{pre}})] \leq \sum_{k=1}^{r}\hat{G}_n(h_k(x_i^{\text{pre}}))$$
$$\leq \log(n)\cdot\sum_{k=1}^{r}\hat{R}_nh_k(x_i^{\text{pre}}) \leq r\log(n)\frac{2D\sqrt{K+1+\log d}\cdot\Pi_{p=1}^{K}M(p)}{\sqrt{n}}.$$

  This also implies the population Gaussian complexity.

- By definition the class $\mathcal{F}$ as linear maps with parameters $\|\alpha_s\|_2 \leq c_1M(K)^2, \forall s\in[k-1]$, we obtain that $L(\mathcal{F}) = c_1\sqrt{k-1}M(K)^2$.

- For the complexity of learning $\mathcal{F}^{\text{pre}}$ in the training phase we obtain,

$$\hat{G}_n(\mathcal{F}^{\text{pre}}|h \circ x^{\text{pre}}) = \frac{1}{n}\mathbb{E}_g[\sup_{\alpha \in \mathcal{F}} \sum_{s=1}^{k-1}\sum_{i=1}^{n} g_{is}\alpha_s^\top h(x_i^{\text{pre}})] \lesssim \frac{(k-1)M(K)^2}{n}\mathbb{E}_g[\| \sum_{i=1}^{n} g_{is}h(x_{1i})\|]$$
$$\lesssim \frac{(k-1)M(K)^2}{n}\sqrt{\sum_{i=1}^{n}\|h(x_i^{\text{pre}})\|^2} \lesssim \frac{(k-1)M(K)^2}{\sqrt{n}} \max_i \|h(x_i^{\text{pre}})\|.$$

For $tanh$ activation function, we simply have

$$\|h(x)\|^2 = \|W_K r_{K-1}\|_2^2 \leq \|W_K\|_{\infty \to 2}^2,$$

where r_{K-1} denotes ourput of the $K-1$th layer,

$$\|h(x)\| \leq \mathcal{O}(M(K)).$$

In conclusion we obtain

$$\bar{G}_n(\mathcal{F}^{\text{pre}}) \leq O\left(\frac{(k-1)M(K)^3}{\sqrt{n}}\right).$$

- Similarly

$$\hat{G}_m(\mathcal{F}^{\text{down}}|h \circ x^{\text{down}}) \leq O\left(\frac{(k'-1)M(K)^3}{\sqrt{m}}\right)$$

Regularity conditions

- Boundedness parameter $D_{\mathcal{X}^{\text{pre}}} = \sup_{\alpha,h} \|\alpha^\top h(x^{\text{pre}})\| = \mathcal{O}(\sqrt{k-1}M(K)^3)$.
- Pre-training loss is $L^{\text{pre}} = \sqrt{k-1}$-Lipschitz and $B^{\text{pre}} = \mathcal{O}(\sqrt{k-1}M(K)^3)$-bounded.
- Downstream loss is $L^{\text{down}} = \sqrt{k'-1}$-Lipschitz and $B^{\text{down}} = \mathcal{O}(\sqrt{k'-1}M(K)^3)$-bounded.

Assembling the previous complexity arguments shows the transfer learning risk is bounded by

$$\lesssim \frac{L^{\text{pre}}}{\tilde{\nu}}\left(\log(n)\left[L(\mathcal{F}^{\text{pre}})r\log(n)\frac{D\sqrt{K}\Pi_{p=1}^{K}M(p)}{\sqrt{n}} + \frac{kM(K)^3}{\sqrt{n}}\right]\right) + \frac{L^{\text{down}}k'M(K)^3}{\sqrt{m}}$$
$$+\left(\frac{1}{\tilde{\nu}}\max\left(\frac{L^{\text{pre}}\sqrt{k}D_{\mathcal{X}^{\text{pre}}}}{n^2}, B^{\text{pre}}\sqrt{\frac{\log(1/\delta)}{n}}\right) + B^{\text{down}}\sqrt{\frac{\log(1/\delta)}{m}}\right)$$

Substitute regularity conditions into it, then the risk is simplified as stated in Theorem 4. □

A.6 Experiment details

Full statistics (including matched and mismatched dev sets for MNLI, accuracy and F1 scores for MRPC and QQP, and (Pearson's correlation + Spearman's correlation)/2 for STS-B. All other tasks uses accuracy as the metric) could be found in Table A.6.

A.6.1 Complete statistics

Here we provide complete results on GLUE dev sets over 5 random seeds.

Table 2: Full statistics on GLUE dev sets.

Model	Statistics	MNLI(m/mm)	MRPC(acc/F1)	SST-2	CoLA	QQP(acc/F1)	QNLI	RTE	STS-B
$\lambda = 0.005$	mean	83.96/84.37	84.90/89.42	92.48	59.99	90.96/87.88	88.11	67.28	89.33
	std	0.26/0.21	2.28/1.34	0.19	0.28	0.05/0.11	0.54	3.43	0.07
$\lambda = 0.05$	mean	83.88/84.14	83.72/88.98	93.00	62.66	90.97/87.96	87.64	60.64	89.57
	std	0.04/0.16	6.69/3.62	0.16	1.07	0.05/0.04	0.44	6.08	0.13
$\lambda = 0.5$	mean	83.96/84.04	87.75/91.09	92.93	60.76	90.85/87.81	88.01	67.93	89.22
	std	0.15/0.24	0.52/0.50	0.24	0.71	0.10/0.14	0.23	1.18	0.23
BERT-base	mean	83.85/84.07	83.48/88.80	92.64	61.46	90.87/87.68	88.10	63.64	89.19
	std	0.13/0.04	6.08/3.19	0.20	0.74	0.07/0.11	0.27	6.64	0.07

Table 3: Performance of reproduced BERT-base model.

	GLUE	MNLI(m/mm)	MRPC(acc/F1)	SST-2	CoLA	QQP(acc/F1)	QNLI	RTE	STS-B
seed	42	83.90/84.04	71.32/82.44	92.66	60.11	90.94/87.72	87.58	66.79	89.25
	0	83.86/84.12	86.27/90.18	92.43	62.05	90.74/87.53	88.19	54.29	89.07
	393	83.78/84.06	86.76/90.63	92.54	61.42	90.93/87.84	88.29	70.36	89.19
	78	84.05/84.02	86.76/90.63	92.55	61.50	90.89/87.60	88.12	57.14	89.18
	3837	83.66/84.11	86.27/90.13	93.00	62.20	90.87/87.73	88.33	69.64	89.26
	mean	83.85/84.07	83.48/88.80	92.64	61.46	90.87/87.68	88.10	63.64	89.19
	std	0.13/0.04	6.08/3.19	0.20	0.74	0.07/0.11	0.27	6.64	0.07

Table 4: Performance of $\lambda = 0.005$ regularized pre-training model.

	GLUE	MNLI(m/mm)	MRPC(acc/F1)	SST-2	CoLA	QQP(acc/F1)	QNLI	RTE	STS-B
seed	42	84.24/84.43	87.25/90.72	92.20	59.99	90.94/87.85	87.09	67.14	89.40
	0	83.91/83.96	86.27/90.47	92.43	60.06	91.03/87.88	88.24	70.71	89.36
	393	83.84/84.51	85.54/89.52	92.55	59.48	90.89/87.68	88.33	71.07	89.22
	78	84.23/84.44	84.80/89.45	92.78	60.13	90.95/87.97	88.71	65.71	89.38
	3837	83.56/84.53	80.64/86.93	92.43	60.30	91.01/88.00	88.17	61.79	89.28
	mean	83.96/84.37	84.90/89.42	92.48	59.99	90.96/87.88	88.11	67.28	89.33
	std	0.26/0.21	2.28/1.34	0.19	0.28	0.05/0.11	0.54	3.43	0.07

Table 5: Performance of $\lambda = 0.05$ regularized pre-training model.

	GLUE	MNLI(m/mm)	MRPC(acc/F1)	SST-2	CoLA	QQP(acc/F1)	QNLI	RTE	STS-B
seed	42	83.88/84.22	86.52/90.27	92.89	61.12	91.06/87.95	86.93	65.00	89.52
	0	83.83/83.92	88.97/92.00	93.00	64.36	90.96/87.93	87.73	54.29	89.65
	393	83.96/83.98	70.59/81.92	93.12	62.22	90.94/88.03	87.47	61.79	89.65
	78	83.86/84.26	87.50/91.06	92.78	63.13	90.94/87.97	88.26	68.93	89.69
	3837	83.86/84.30	85.04/89.66	93.23	62.49	90.93/87.91	87.82	53.21	89.33
	mean	83.88/84.14	83.72/88.98	93.00	62.66	90.97/87.96	87.64	60.64	89.57
	std	0.04/0.16	6.69/3.62	0.16	1.07	0.05/0.04	0.44	6.08	0.13

Table 6: Performance of $\lambda = 0.5$ regularized pre-training model.

	GLUE	MNLI(m/mm)	MRPC(acc/F1)	SST-2	CoLA	QQP(acc/F1)	QNLI	RTE	STS-B
seed	42	83.75/83.87	87.75/90.89	93.00	59.79	90.88/87.75	87.58	65.71	89.00
	0	83.84/84.30	88.24/91.56	92.66	60.94	90.87/87.77	88.14	67.86	89.25
	393	83.98/83.68	87.99/91.46	93.12	60.99	90.98/88.04	88.22	68.21	89.19
	78	84.02/84.29	87.99/91.33	93.23	60.22	90.87/87.86	88.12	68.93	89.65
	3837	84.19/84.05	86.76/90.21	92.66	61.86	90.67/87.61	87.98	68.93	89.03
	mean	83.96/84.04	87.75/91.09	92.93	60.76	90.85/87.81	88.01	67.93	89.22
	std	0.15/0.24	0.52/0.50	0.24	0.71	0.10/0.14	0.23	1.18	0.23

A.6.2 HYPERPARAMETERS

Pre-training Hyperparameters for pre-training are shown in Table 7.

Hyperparam	**phase-1**	**phase-2**
Number of Layers	12	12
Hidden size	768	768
FFN inner hidden size	3072	3072
Attention heads	12	12
Steps	7038	1563
Optimizer	LAMB	LAMB
Learning Rate	9e-3	6e-3
β_1	0.9	0.9
β_2	0.999	0.999
WarmUp	28.43 %	12.80 %
Batch Size	65536	32768

Table 7: Hyperparameters used in pre-training our models. We use the LAMB optimizer (You et al., 2019) for large-batch pretraining of the BERT model, where β_1 and β_2 are its two hyper-parameters.

Finetuning Hyperparameters for downstream tasks are shown in Table 8. We adapt these hyperparameters from Liu et al. (2020), Devlin et al. (2019), and Yang et al. (2019).

	LR	BSZ	# EP	WARMUP	WD	FP16	SEQ
CoLA	1.00E-05	32	20	6%	0.1	O2	128
SST-2	3.00E-05	32	10	6%	0.1	O2	128
MNLI	3.00E-05	32	5	6%	0.1	O2	128
QNLI	3.00E-05	32	10	6%	0.1	O2	128
QQP	3.00E-05	32	5	6%	0.1	O2	128
RTE	3.00E-05	16	5	6%	0.1	O2	128
MRPC	3.00E-05	16	5	6%	0.1	O2	128

Table 8: The hyperparameters used in finetuning our model in downstream tasks. LR: learning rate. BSZ: batch size. #EP: number of epochs. WARMUP: warmup ratio. FP16: automatic mixed precision (AMP) level. SEQ: input sequence length.

A.7 OTHER DETAILS

Computing infrastructure We pretrain our (diversity-regularized) BERT-base models using 32 Nvidia V100 GPUs (32GB RAM each), and the finetuning of the model uses 4 Nvidia V100 GPUs.