On the Generalization Ability of Unsupervised Pretraining

Yuyang Deng
Penn State University

Junyuan Hong
Michigan state university

Jiayu Zhou
Michigan state university

Mehrdad Mahdavi
Penn State University

Abstract

Recent advances in unsupervised learning have shown that unsupervised pre-training, followed by fine-tuning, can improve model generalization. However, a rigorous understanding of how the representation function learned on an unlabeled dataset affects the generalization of the fine-tuned model is lacking. Existing theoretical research does not adequately account for the heterogeneity of the distribution and tasks in pre-training and fine-tuning stage. To bridge this gap, this paper introduces a novel theoretical framework that illuminates the critical factor influencing the transferability of knowledge acquired during unsupervised pre-training to the subsequent fine-tuning phase, ultimately affecting the generalization capabilities of the fine-tuned model on downstream tasks. We apply our theoretical framework to analyze generalization bound of two distinct scenarios: Context Encoder pre-training with deep neural networks and Masked Autoencoder pre-training with deep transformers, followed by fine-tuning on a binary classification task. Finally, inspired by our findings, we propose a novel regularization method during pre-training to further enhances the generalization of fine-tuned model. Overall, our results contribute to a better understanding of unsupervised pre-training and fine-tuning paradigm, and can shed light on the design of more effective pre-training algorithms.

1 Introduction

Unsupervised representation learning has achieved remarkable success in various domains, including computer vision and natural language processing, as evidenced by a rapidly increasing number of empirical studies [Coates and Ng, 2012, Radford et al., 2015, Sun et al., 2019, Dosovitskiy et al., 2020, Feichtenhofer et al., 2022, He et al., 2020, 2022, Devlin et al., 2018, Chen et al., 2020]. In this learning paradigm, the goal is to learn a representation function on a large, possibly unlabeled dataset by optimizing a carefully designed unsupervised learning objective. Then, using the learned representation, a task-specific classifier, such as the head of a neural network, is trained on a small in-house dataset during the fine-tuning stage. This two-stage paradigm addresses the issue of small dataset size in downstream tasks. While unsupervised pre-training for transfer learning has experienced significant empirical growth, a comprehensive understanding of the fundamental factors that influence the generalization performance of fine-tuned models lags considerably behind what has been empirically observed [Neyshabur et al., 2020].

Most existing generalization bounds primarily rely on notions such as distance between the weights of the pre-trained and fine-tuned models [Li and Zhang, 2021, Shachaf et al., 2021] or data-dependent measurements such as Hessian [Ju et al., 2022] through PAC-Bayesian analysis [Arora et al., 2018, Neyshabur et al., 2018] to examine the performance of fine-tuned model. These results inform the design of effective regularization methods [Li and Zhang, 2021, Ju et al., 2022] or incorporating consistent losses [Ju et al., 2022] in fine-tuning stage to improve the generalization of fine-tuned model by mitigating issues such as overfitting caused by fine-tuning a large model on a small training set or instability due to label noise. These generalization bounds, however, do not explicitly incorporate other key factors that may govern the success of fine-tuning such as similarity between the pre-training (on which a model is pre-trained) and target tasks [Shachaf et al., 2021] or task diversity [Tripuraneni et al., 2020], the number of training samples and complexity of model spaces utilized in each stage in a *unified bound*. For example, in real-world learning tasks, the pre-training and fine-tuning tasks may be conducted on completely different domains, and we usually employ some kind of

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

transformation on the pre-training data (i.e., adding noise, rotating or masking), which further exacerbates the data heterogeneity. Consequently, a well-designed generalization theory is expected to take the data heterogeneity into account [Yang et al., 2020]. In modern transfer learning, different tasks can be conducted in the pre-training and fine-tuning stages. For example, in a Masked Autoencoder (MAE) [He et al., 2022], a regression task utilized during pre-training, while a classification task used for fine-tuning. Therefore, a desired theory should allow for flexibility in choosing diverse types of tasks in the pre-training and fine-tuning stages which poses a challenge in formalizing the desired guarantees.

Motivated by the above observations, we aim at formalizing and establishing general generalization bounds on unsupervised pre-training and fine-tuning paradigm that captures aforementioned factors in a unified manner. We introduce the notion of *representation transferrability* to quantify how much knowledge can be transferred from unsupervised representation learning stage to fine-tuned model, in the presence of task heterogeneity. We then establish a bound on the generalization capability of fine-tuned model composed with pre-trained representation model that highlights how representation-induced complexity and distribution mismatch affects the generalization of fine-tuned model. We instantiate our theory to the scenario of Context Encoder [Pathak et al., 2016] with deep neural networks and Masked Autoencoder [Devlin et al., 2018, He et al., 2022] with deep Transformer architectures which highlights the relative merits of learning representations. From a technical perspective, we establish generalization bounds for multi-layer transformers, by deriving the worst case covering number of hypothesis space by expanding upon the machinery that was developed in [Edelman et al., 2022].

Since our theoretical analysis reveals the representation-induced Rademacher complexity as one of the key facotors governing the capacity of the transfer learning, it naturally motivates itself to be incorporated as a regularizer during pre-training. Inspired by this observation, we propose a novel *Rademacher representation regularized* algorithm, dubbed as RadReg, to enhance the generalization capability of the fine-tuned model. We show that by utilizing unlabeled data from the downstream task, we can effectively regularize the pre-trained model to learn representations that entail better generalization after fine-tuning. We propose an efficient algorithm to optimize the new objective and establish its convergence on smooth nonconvex losses.

Contributions. Our main contributions are summarized as follows:

- (Theory) We introduce a formal framework to study the utility of unsupervised representation learning and fine-tuning paradigm (Section 3) and derive the generalization bound for fine-tuned model based on a pre-trained representation function (Section 4). We discover that the generalization capability of model depends on four key factors: Representation transferrability, representation-induced Rademacher complexity, domain heterogeneity, and generalization of the pre-training task.
- (Applications) We apply our theory to derive generalization bound of the pre-training with a context encoder (CE) and a masked autoencoder (MAE) with a transformer followed by a binary classification fine-tuning task (Section 5). We show that, the pre-training tasks defined by regression loss are provably transferrable to downstream binary classification task. In doing so, to our best knowledge, we establish the first generalization analysis of multi-layer transformer models with residual block.
- (Algorithm) Inspired by our generalization bounds, we propose a novel Rademacher Representation Regularized algorithm, RadReg, for improved pre-training and provide convergence guarantees for nonconvex objectives (Section 6). The experimental results show that RadReg can learn better representation than ℓ_2 norm regularized training on downstream tasks with a small dataset (Section 7).

2 Additional Related Works

Theory of Transfer Learning A significant body of work [Tripuraneni et al., 2020, Du et al., 2020] focuses on the theoretical aspects of transfer learning paradigm, trying to answer the question: *why transfer learning can work and what factors affect the learning performance?* Tripuraneni et al. [2020] give the first risk bound capturing *task diversity* among pre-training and fine-tuning stages as the key quantity affecting generalization which is also reflected in our generalization analysis. Xu and Tewari [2021] follow the setup in [Tripuraneni et al., 2020], and show that even though the model architectures used in representation learning and fine-tuning are different, the task diversity remains bounded. Du et al. [2020] study few-shot representation learning, where it considers pre-training a linear representation function by solving the regression problem with squared loss (OLS) on a give large dataset, and then fine-tuning another linear predictor on some target dataset. It shows that the generalization will depend on the number of pre-training data and fine-tuning data. Similar dependencies appear in the generalization bound obtained in our main theorem. Zhang et al. [2023] study the general supervised

pretraining, and highlight the trade-off between the intra and inter class diversity.

Theory of Modern Unsupervised Representation Learning. Recently, due to the rise of contrastive learning [Chen et al., 2020] and masked training [Devlin et al., 2018, He et al., 2022], a line of studies are devoted to understanding the generalization capability or smaple complexity of these learning paradigms [HaoChen et al., 2021, Arora et al., 2019b, Wang and Isola, 2020, Lee et al., 2021, Ge et al., 2023, Gouk et al., 2020, Ju et al., 2022]. Arora et al. [2019b] presents a theoretical framework for studying contrastive learning, and shows that it provably reduces the sample complexity of downstream tasks. HaoChen et al. [2021] considers contrastive learning and establishes the theory without conditional independence of positive data pairs. Wang and Isola [2020] proves that contrastive learning optimizes for alignment and uniformity asymptotically. Zhang et al. [2022] establishes the connection of masked pre-training with contrastive learning over bipartite graphs. Lee et al. [2021] also considers a masking pre-training scenario, but contrary to the present work, it assumes the labels are generated by a function of masked data plus Gaussian noise, and only focuses on the ERM model as a representation function. A recent work [Ge et al., 2023] also examines the unsupervised pre-training framework, but the difference to ours, they consider a maximum likelihood estimation as pre-training method, while we start from general pre-training task and instantiate it in modern machine learning scenario such as Context Encoder and MAE. Gouk et al. [2020] study the end-to-end finetuing scenario, and find that the generalization of finetuned model will depend on the distance that neural network weights traveled away from pretrained model. They hence propose a distance regularization finetuning algorithm and achieve better performance. Ju et al. [2022] also studies the entire model finetuning paradigm, and derive a Hessian based generalization bound via PAC-Bayesian analysis.

3 A Formal Framework

In this section we formalize unsupervised pre-training followed by supervised fine-tuning problem that will enable us to study the relative merits of various unsupervised representation learning approaches and examine their utility on the generalization capability of downstream tasks. In the scenario of unsupervised representation pre-training and fine-tuning on a downstream task, we are given two datasets: one, possibly large, unlabeled pre-training dataset and a small labeled data set. The goal is to learn model $f \circ h$ which is composed of task-specific function $f \in \mathcal{F}$ and representation function $h \in \mathcal{H}$, where $\mathcal{F}$ and $\mathcal{H}$ are model spaces for fine-tuned and representation models, respectively.

Unsupervised pre-training. We assume access to a raw pre-training data $\{\tilde{\mathbf{x}}_i\}_{i=1}^N$ drawn from an unknown, arbitrary distribution $\mathcal{D}$ over an instance domain $\mathcal{X}$ such as images. To learn representations, one first transforms (e.g., masking, adding noise, rotating, or other geometric transformation) unlabeled data into $\tilde{\mathbf{z}}_i = T_1(\tilde{\mathbf{x}}_i) \in \mathcal{X}$ and (self-generated) label $\tilde{\mathbf{y}}_i = T_2(\tilde{\mathbf{x}}_i) \in \mathcal{Z}$ using suitable transformers $T_1 : \mathcal{X} \mapsto \mathcal{X}$ and $T_2 : \mathcal{X} \mapsto \mathcal{Z}$ to generate the pre-training dataset $\widehat{\mathcal{U}} = \{(\tilde{\mathbf{z}}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^N$. For example, in pre-training with masking, our augmented data are masked sentence/image, and self-generated labels are the masked part of data. We denote the transformed distribution over $\mathcal{X} \times \mathcal{Z}$ as $\mathcal{U}$. We note that that marginal distribution $\mathcal{U}_{\mathcal{X}}$ of $\mathcal{U}$ over instance space $\mathcal{X}$ is not necessarily same as $\mathcal{D}$ of raw data due to randomness in data transformation T_1. To learn the representations, we consider a class of decoding and encoding pairs, which is closely inspired by [Hazan and Ma, 2016], and minimize the following empirical risk

$$\min_{g \in \mathcal{G}, h \in \mathcal{H}} \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) := \frac{1}{N} \sum\nolimits_{(\tilde{\mathbf{z}}_i, \tilde{\mathbf{y}}_i) \in \widehat{\mathcal{U}}} \ell(g \circ h(\tilde{\mathbf{z}}_i), \tilde{\mathbf{y}}_i), \tag{1}$$

where $\mathcal{G} \subseteq \{\mathcal{I} \mapsto \mathcal{Z}\}$ and $\mathcal{H} \subseteq \{\mathcal{X} \mapsto \mathcal{I}\}$ are the model spaces for encoder and decoder, respectively, where $\mathcal{I}$ denotes the latent space of representations, and ℓ is the loss function used for pre-training, e.g., $\ell(g \circ h(\tilde{\mathbf{z}}_i), \tilde{\mathbf{y}}_i) = \|g \circ h(\tilde{\mathbf{z}}_i) - \tilde{\mathbf{y}}_i\|_2^2$. Let $\hat{g}$ and $\hat{h}$ denote the decoder and encoder (representation function) obtained by solving (1). We define the following excess risk for pre-training task:

$$\mathcal{E}_{\mathcal{U}}(\hat{g}, \hat{h}) := \mathcal{L}_{\mathcal{U}}(\hat{g} \circ \hat{h}) - \min_{g \in \mathcal{G}, h \in \mathcal{H}} \mathcal{L}_{\mathcal{U}}(g \circ h)$$

where $\mathcal{L}_{\mathcal{U}}(g \circ h) := \mathbb{E}_{(\tilde{\mathbf{z}}, \tilde{\mathbf{y}}) \sim \mathcal{U}}[\ell(f \circ h(\tilde{\mathbf{z}}), \tilde{\mathbf{y}})]$ denotes the generalization ability of pre-training task realized by distribution $\mathcal{U}$. We note that the learned decoder function $\hat{g}$ may be discarded after pre-training. We use $h^*_{\mathcal{U}} = \arg\min_{h \in \mathcal{H}} \min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{U}}(g \circ h) \in \mathcal{H}$ to denote optimal encoder for pre-training task.

Supervised fine-tuning. In fine-tuning stage, we assume access to a labeled downstream dataset $\widehat{\mathcal{T}} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ where feature vector $\mathbf{x}_i$ is sampled based an unknown, arbitrary distribution $\mathcal{T}$ (possibly different from $\mathcal{D}$) on domain $\mathcal{X}$, and its label $\mathbf{y}_i$ is generated based on a labeling function $\mathbf{y}_i = y(\mathbf{x}_i)$. The goal is to utilize the representation function $\hat{h}$ obtained by solving (1) to perform *fine-tuning* on the downstream dataset $\widehat{\mathcal{T}}$ to learn a prediction model $\hat{f}$ from a function class $\mathcal{F}$:

$$\min_{f \in \mathcal{F}} \mathcal{R}_{\widehat{\mathcal{T}}}(f \circ \hat{h}) := \frac{1}{n} \sum\nolimits_{(\mathbf{x}_i, \mathbf{y}_i) \in \widehat{\mathcal{T}}} \phi(f \circ \hat{h}(\mathbf{x}_i), \mathbf{y}_i), \tag{2}$$

where ϕ is the loss function which is not necessarily the same as the pre-training loss.

Our goal is to rigorously analyze the generalization capability of the final model which is the composition of two functions, i.e., $\hat{f} \circ \hat{h}$ where $\hat{f}$ is the solution of (2), by bounding the excess risk

$$\mathcal{E}_{\mathcal{T}}(\hat{f}, \hat{h}) = \mathcal{R}_{\mathcal{T}}(\hat{f} \circ \hat{h}) - \min_{f \in \mathcal{F}, h \in \mathcal{H}} \mathcal{R}_{\mathcal{T}}(f \circ h). \quad (3)$$

Here $\mathcal{R}_{\mathcal{T}}(f \circ h) := \mathbb{E}_{\mathbf{x} \sim \mathcal{T}}[\phi(f \circ h(\mathbf{x}), y(\mathbf{x}))]$ denotes the true risk on downstream task realized by distribution $\mathcal{T}$ over $\mathcal{X}$ and underlying labeling function $y(\cdot)$.

4 On the Utility of Unsupervised Representation Learning

We now turn to establishing the generalization bound of fine-tuned models given a pre-trained representation function, and discuss its implications. Before, we first introduce two key notions. We start by introducing the notion of Rademacher complexity of a hypothesis space when individual models are composed with a fixed representation function (similar measures appear in [Tripuraneni et al., 2020, Xu and Tewari, 2021]).

Definition 1 (Representation-induced Rademacher complexity). *For a hypothesis space $\mathcal{F}$ of set of real (vector)-valued functions defined over input space $\mathcal{X}$ and label space $\mathcal{Y}$, a loss function $\phi : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+$, and a dataset $\widehat{\mathcal{T}} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, the empirical Representation-induced Rademacher complexity of $\mathcal{F}$ with respect to ϕ and $\widehat{\mathcal{T}}$, for a given representation function $\hat{h}$, is defined as*

$$\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi \circ \mathcal{F} \circ \hat{h}) := \mathbb{E}_{\boldsymbol{\varepsilon} \in \{\pm 1\}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum\nolimits_{i=1}^n \varepsilon_i \phi(f \circ \hat{h}(\mathbf{x}_i), \mathbf{y}_i) \right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. *Rademacher random variables with $\mathbb{P}\{\varepsilon_i = 1\} = \mathbb{P}\{\varepsilon_i = -1\} = 1/2$.*

The following definition, relates the generalization of fine-tuned and representation models.

Definition 2 (Represnetation transferability). *Given two representation functions $h, h' \in \mathcal{H}$ and a distribution $\mathcal{U}$ for pre-training data, we say a pre-training task and fine-tuning task satisfies (C_β, β) transferability for some constant $0 < C_\beta < \infty, \beta > 0$ on h, h', if the following statement holds:*

$$\min_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{T}}(f \circ h) - \min_{f' \in \mathcal{F}} \mathcal{R}_{\mathcal{T}}(f' \circ h') \leq C_\beta \left(\min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{U}}(g \circ h) - \min_{g' \in \mathcal{G}} \mathcal{L}_{\mathcal{U}}(g' \circ h') \right)^\beta$$

where $\mathcal{R}_{\mathcal{U}_\mathcal{X}}(f \circ h) := \mathbb{E}_{\mathbf{x} \sim \mathcal{U}_\mathcal{X}}[\phi(f \circ h(\mathbf{x}), y(\mathbf{x}))]$ denotes the risk realized by pre-training marginal data distribution $\mathcal{U}_\mathcal{X}$ and downstream labeling function $y(\cdot)$.

We note that a similar notation is proposed in the analysis of multi-task learning [Hanneke and Kpotufe, 2022, Definition 4], to characterize the transferrability from one task to another task. We emphasize that represnetation transferability is the key to transfer the generalizability of pre-training model to fine-tuned model. Unlike the transfer ratio defined in previous works [Tripuraneni et al., 2020, Ge et al., 2023, Zhang et al., 2023], we have an exponent variable β, which allows the transferrability from losses with different order, e.g., from a quadratic loss to non-quadratic loss. Later on, we show that condition holds essentially under realistic assumptions on suitable data transformations to generate pre-training data and model spaces, such as pre-training with a inpainting autoencoder and a masked autoencoder with a transformer, where both are fine-tuned on a classification task.

The next theorem establishes the generalization bound of the fine-tuned model on a downstream dataset, given a pre-trained representation function $\hat{h}$.

Theorem 1. *Assume $\hat{h}$ and $\hat{g}$ are the pre-trained representation function and its associated decoder function, and real valued non-negative loss ϕ to be G_ϕ Lipschitz and bounded by B_ϕ. Assume pre-training and fine-tuning task admit (C_β, β) representation transferrability on $\hat{h}$ and $h^*_{\mathcal{U}}$. If we solve (2) to get $\hat{f}$, then with probability at least $1 - \nu$, the following statement holds*

$$\mathcal{E}_{\mathcal{T}}(\hat{f}, \hat{h}) \leq C_\beta \mathcal{E}_{\mathcal{U}}(\hat{g}, \hat{h})^\beta + 4 G_\phi \mathfrak{R}_{\widehat{\mathcal{T}}}(\mathcal{F} \circ \hat{h}) + 4 B_\phi \sqrt{\frac{\log(1/\nu)}{n}} + 4 B_\phi \left\| \mathcal{T} - \mathcal{U}_\mathcal{X} \right\|_{\mathrm{TV}} + \min_{f \in \mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}})$$

where $h^*_{\mathcal{U}} = \arg\min_{h \in \mathcal{H}} \min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{U}}(g \circ h)$ is the optimal pre-training representation function, and $\|\mathcal{P} - \mathcal{Q}\|_{\mathrm{TV}} = \sup_{A \in \Omega} |\mathcal{P}(A) - \mathcal{Q}(A)|$ denotes total variation distance between two distributions.

The proof of Theorem 1 is deferred to Appendix B.1. Theorem 1 shows that the generalization of the fine-tuned model depends on four quantities: i) Representation transferrability, ii) Representation-induced Rademacher complexity, iii) domain heterogeneity and iv) generalization of the pre-training task.

Representation transferrability is the key to connect downstream generaliztion with pre-training generalization. It is analogous to task diversity notion in the multi-task learning works [Tripuraneni et al., 2020, Xu and Tewari, 2021], since they all measure how well the knowledge can be transferred across different learning stage. However, our notion is more powerful since we neither assume the pre-training and fine-tuning stage share the same type of task, e.g., both being regression, nor assume a generic nonlinear feature representation is shared across all tasks [Tripuraneni et al., 2020]. As

we will see in the later section, with the help of representation transferrability, we can show that encoder learnt by regression pre-training can be transferred to downstream classification task. The representation-induced Rademacher complexity will play a key role in reflecting how well the learnt representation and $\mathcal{F}$ are coupled. Notice that this complexity is defined over downstream data, and only over class $\mathcal{F}$, which means that in fine-tuning stage we only suffer from a smaller complexity in learning. The price for learning with potential more complex encoder class $\mathcal{H}$ is paid in pre-training task. This observation is consistent with a line of multi-task or transfer learning works [Tripuraneni et al., 2020, Du et al., 2020, Xu and Tewari, 2021, Ge et al., 2023].

The domain heterogeneity term $\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\|_{\mathrm{TV}}$ characterizes the statistical heterogeneity between the pre-training task and the fine-tuning task. Generalization of the pre-training task also appears in the bound which depends on the convergence of the optimization algorithm, and the complexity of the representation function classes $\mathcal{G}$ and $\mathcal{H}$ for decoder and encoder, respectively. The last term $\min_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}})$ is the downstream risk evaluated with optimal representation model $h^*_{\mathcal{U}}$, which characterize the task heterogeneity between pre-training and downstream stages. If the two tasks are well aligned, i.e., they share similar optimal representation, then this quantity is ignorable.

5 The Power of Unsupervised Representation Learning

We now proceed to establish generalization bounds in two distinct settings by refining the generic result presented in the previous section.

5.1 Pre-training with Context Encoder

The setting. We start by applying our theory to the setting where inpainting task is considered as pre-training task and binary classification as downstream task. This learning paradigm is also known as Context Encoder (CE) [Pathak et al., 2016], where in pre-training stage, a deep neural network is trained by reconstructing a random transformation of raw data (e.g, rotating, scaling, adding Gaussian noise or masking) of a given image:

$$\min_{g\in\mathcal{G},h\in\mathcal{H}} \mathcal{L}_{\widehat{\mathcal{U}}}(g\circ h) := \frac{1}{N}\sum_{i=1}^{N} \|g(h(\tilde{\mathbf{z}}_i)) - \mathbf{z}_i\|^2 \quad (4)$$

to learn $\hat{g}$ and $\hat{h}$. Then, we discard the decoder $\hat{g}$, and use the rest layers as an encoder. A linear projection head is added on top of encoder in fine-tuning stage, on the downstream binary classification task with data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ using the learnt encoder $\hat{h}$:

$$\min_{f\in\mathcal{F}} \mathcal{R}_{\widehat{\mathcal{T}}}(f\circ\hat{h}) = \frac{1}{n}\sum_{i=1}^{n} \phi(f(\hat{h}(\mathbf{x}_i)), y_i). \quad (5)$$

The encoder-decoder architecture is defined as follows:

$$\begin{aligned} &\text{encoder:} \quad h(\mathbf{x}) = \sigma\left(\mathbf{W}_L \cdots \sigma\left(\mathbf{W}_1\mathbf{x}\right)\right), \\ &\text{decoder:} \quad g(h(\mathbf{x})) = \mathbf{W}_{L+1}h(\mathbf{x}), \end{aligned}$$

where $\mathbf{W}_1 \in \mathbb{R}^{m\times d}$, $\mathbf{W}_2, \ldots, \mathbf{W}_L \in \mathbb{R}^{m\times m}$, and $\mathbf{W}_{L+1} \in \mathbb{R}^{d\times m}$. In fine-tuning stage, we add a linear head on top of encoder function, i.e., $f(h(\mathbf{x})) = \boldsymbol{\theta}^\top h(\mathbf{x})$. The hypothesis class for encoder is then defined as:

$$\mathcal{H} := \left\{ \mathbf{x} \mapsto \sigma\left(\mathbf{W}_L \cdots \sigma\left(\mathbf{W}_1\mathbf{x}\right)\right) : \|\mathbf{W}_l\| \le W(l), \; \|\mathbf{W}_l\|_{2,1} \le B(l) \right\}$$

where $W(l)$ and $B(l)$ are upper bound on spectral and $(2,1)$ norms of weight matrices, respectively.

The decoder class is defined as:

$$\mathcal{G} := \left\{ \mathbf{x} \mapsto \mathbf{W}_{L+1}\mathbf{x} : \|\mathbf{W}_{L+1}\| \le W(L+1), \; \|\mathbf{W}_{L+1}\|_{2,1} \le B(L+1) \right\}.$$

Generalization bound. The following lemma establishes the representation transferrability of CE pre-training to binary classification task. We need to make the following assumption

Assumption 1 (Realizability)**.** *There exists $g^* \in \mathcal{G}$ and $h^*_{\mathcal{U}} \in \mathcal{H}$ such that $\mathcal{L}_{\mathcal{U}}(g^* \circ h^*_{\mathcal{U}}) = 0$.*

Remark 1. *In Assumption 1 we assume that there exist optimal encoder and decoder that can perfectly realize pre-training task. This is reasonable if we consider overparameterized model, e.g., deep neural network. For example, in masked image reconstruction pre-training, at the most cases, the remaining part of image is enough for deep model to reconstruct the raw image [Pathak et al., 2016, He et al., 2022].*

Lemma 1. *Under Assumption 1, CE pre-training admits an $\left(\Omega\left(1\right), \frac{1}{2}\right)$ representation transferrability to binary classification task.*

The proof of Lemma 1 is deferred to Appendix C.1. This lemma shows that generalization of a pre-training regression task can be effectively transferred to a downstream binary classification task. Here the transfer exponent is $\frac{1}{2}$, which implies that the generalization risk of downstream task will be roughly square root of pre-training generalization. To get the excess risk rate of downstream task, we need to derive the generalization risk of neural network regression. The existing works [Cao and Gu, 2019, 2020, Arora et al., 2019a]

mainly focus on classification task where the loss function is Lipschitz, which is not the case in regression loss. Our technique is to generalize the seminal analysis in [Srebro et al., 2010] for smooth losses and scalar valued hypothesis classes to a vector valued hypothesis class, i.e., neural network class in our case and borrow the standard neural network covering number result from [Bartlett et al., 2017] to conclude the proof.

Theorem 2. *Assume $\hat{h}$ and $\hat{f}$ are the pre-trained representation function and its associated decoder function obtained by solving (4) and (5). Let $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_N]$ and $\mathbf{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_N]$ be pre-training and downstream data, then under Assumption 1 with probability at least $1 - \nu$, the following statement holds:*

$$\mathcal{E}_{\mathcal{T}}(\hat{f}, \hat{h}) \leq \tilde{O}\left(\frac{\sqrt{s_{L+1}\left\|\mathbf{X}\right\|^2}}{n} + \sqrt{\frac{\left\|\tilde{\mathbf{Z}}\right\|^2 s_{L+1}\left(\sum_{l=1}^{L+1}\rho_l\right)^3}{N}}\right) + 4B_\phi\left(\sqrt{\frac{\log(\frac{1}{\nu})}{n}} + \left\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}}\right) + \min_{f\in\mathcal{F}}\mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}}),$$

where $s_l = \prod_{l=1}^{L+1} W^2(l), \rho_l = B(l)/W(l)$.

The proof of Theorem 2 is deferred to Appendix C.3. Here we achieve roughly $O(\frac{\mathcal{C}(\mathcal{F})}{\sqrt{n}} + \frac{\mathcal{C}(\mathcal{G}\circ\mathcal{H})}{\sqrt{N}})$ bound for downstream task where $\mathcal{C}(\cdot)$ denotes the complexity of the set. The cost of learning the complex heavyweight encoder is incurred during the pre-training task, whereas in the fine-tuning stage, we only endure the complexity of learning a lightweight classification head.

5.2 Pre-training with masked autoencoder with tranformer models

The setting. Here we apply our theory to explain the empirical success of masked autoencoder pre-training methods [He et al., 2022]. In masked autoencoder pre-training, taking vision tasks for example, we draw a large set of images $\mathbf{Z}_1, ..., \mathbf{Z}_N \in \mathbb{R}^{K\times d}$, and then randomly mask some patches of each image to get $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_N \in \mathbb{R}^{K\times d}$. Then an encoder-decoder model is trained by recovering the missing patches:

$$\min_{g\in\mathcal{G}, h\in\mathcal{H}} \mathcal{L}_{\widehat{\mathcal{U}}}(g\circ h) := \frac{1}{N}\sum_{i=1}^{N}\left\|g(h(\tilde{\mathbf{Z}}_i)) - \mathbf{Z}_i\right\|_{\mathrm{F}}^2 \quad (6)$$

to get $\hat{g}$ and $\hat{h}$. Finally, we discard the decoder and only fine-tune a new head (e.g., linear projection layer) on the downstream binary classification task with data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ using the encoder:

$$\min_{f\in\mathcal{F}} \mathcal{R}_{\widehat{\mathcal{T}}}(f\circ\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}\phi(f(\hat{h}(\mathbf{X}_i)), y_i). \quad (7)$$

We consider an L-layer transformer as the pre-training encoder model, and a linear projection layer as the pre-train decoder model, and a linear projection layer for binary classification as fine-tune model.

$$\begin{aligned}\textsf{encoder:}\quad & h(\mathbf{X}) = \mathrm{SA}_{\mathbf{W}^L}\left(\mathrm{SA}_{\mathbf{W}^{L-1}}\left(\cdots \mathrm{SA}_{\mathbf{W}^1}(\mathbf{X})\right)\right),\\ \textsf{decoder:}\quad & g(h(\mathbf{X})) = (h(\mathbf{X}))\,\mathbf{W}_D,\end{aligned}$$

where $\mathrm{SA}_{\mathbf{W}}(\cdot)$ is a self-attention module parameterized by $\mathbf{W} = (\mathbf{W}_V, \mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_{\mathrm{FC1}}, \mathbf{W}_{\mathrm{FC2}})$, which is defined as

$$\begin{aligned}\mathrm{SA}_{\mathbf{W}}(\mathbf{X}) &= \alpha_2\sigma\left(\mathbf{Z}\mathbf{W}_{\mathrm{FC1}}\right)\mathbf{W}_{\mathrm{FC2}} + \mathbf{Z},\\ \mathbf{Z} &= \left(\alpha_1\mathbf{A} + \mathbf{X}\right),\\ \mathbf{A} &= \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}\mathbf{W}_K(\mathbf{X}\mathbf{W}_Q)^\top\right)\mathbf{X}\mathbf{W}_V,\end{aligned}$$

where α_1, α_2 are some small constant, as used in practice [Noci et al., 2022]. We assume lth layer's weights' spectral norm is bounded by $W(l)$, and $(2,1)$ norm is bounded by $B(l)$. In downstream task, we aggregate (sum) over all patches from encoder, add a linear projection head $\boldsymbol{\theta}$ on top of $h(\mathbf{X})$ to make a scalar output:

$$\textsf{downstream:}\quad f(h(\mathbf{X})) = (\mathbf{1}^\top h(\mathbf{X}))\boldsymbol{\theta}.$$

Generalization bound. The following lemma establishes the representation transferrability of MAE with a transformer pre-training to binary classification task.

Lemma 2. *MAE pre-training admits an $\left(\Omega(1), \frac{1}{2}\right)$ representation transferrability to binary classification task*

The proof of Lemma 2 is deferred to Appendix D.1. This implies that MAE pre-training with a multi-layer transformer can be transfered to binary classification task, with a constant factor. The exponent is also 1/2. To derive the excess risk bound of downstream task, we need to find the generalization risk of transformer regression, which is characterized by the following lemma.

Lemma 3 (Generalization of MAE pre-training task)**.** *Let $\hat{g}, \hat{h}$ be the solution of (6), and $\tilde{\mathbf{Z}}_{[N]} = [\tilde{\mathbf{Z}}_1; \ldots; \tilde{\mathbf{Z}}_N]$ is the concatenated pre-training data. Then under Assumption 1 with probability at least $1 - \nu$ the following statement holds:*

$$\mathcal{E}_{\mathcal{U}}(\hat{g}, \hat{h}) \leq O\left(s_L^2\left\|\tilde{\mathbf{Z}}_{[N]}\right\|^2\sum_{l=1}^{L+1}\frac{\rho_l}{N} + \frac{\log(\frac{1}{\nu})}{N}\right),$$

where

$$
\begin{aligned}
s_l &:= \prod_{j=1}^{l} \left(\alpha_2 W^2(j) + 1\right) \left(W^2(j)\alpha_1 K + 1\right), \\
\rho_l &:= O\left((\alpha_1\alpha_2 W^2(l) + \alpha_1)^2 B^2(l) \ln(2d^2)\right) \\
&\quad \times \left(K^2 + \frac{\alpha_1 W^4(l) \left(s_{l-1} \max_{i\in[N]} \|\mathbf{X}_i\|\right)^4}{d_K}\right) \\
&\quad + O\left(\alpha_2^2 W^2(l) B^2(l)(1 + \alpha_1^2 K^2 W^2(l)) \ln(2dm)\right).
\end{aligned}
$$

The proof of Lemma 3 is deferred to Appendix D.2. The proof idea is similar to analysis of CE pre-training, where we connect local Rademacher complexity to the covering number of model class, and then use techniques from [Srebro et al., 2010] to establish the generalization of a smooth loss, i.e. MSE. At the heart of our proof is to carefully control the norm of stacked output of transformer on N samples, so that the final covering number does not scale with N, but only depends on spectral norm of concatenated samples. We note that a similar study [Edelman et al., 2022] also establishes the capacity of transformers, but they do not consider residual blocks.

We now proceed to determine the excess risk associated with fine-tuning a transformer model on a binary classification task.

Theorem 3. *Assume $\hat{h}$ and $\hat{f}$ are the pre-trained representation function and its associated decoder function obtained by solving (6) and (7). Let $\tilde{\mathbf{Z}}_{[N]} = [\tilde{\mathbf{Z}}_1; ...\tilde{\mathbf{Z}}_N]$ and $\mathbf{X}_{[N]} = [\mathbf{X}_1; ...\mathbf{X}_N]$ be pre-training and downstream data, then under Assumption 1, with probability at least $1-\nu$, the following statement holds:*

$$
\begin{aligned}
&\mathcal{E}_{\mathcal{T}}(\hat{f}, \hat{h}) \\
&\leq O\left(\frac{\sqrt{s_L \left\|\mathbf{X}_{[N]}\right\|^2}}{n} + \sqrt{\frac{s_L^2 \left\|\tilde{\mathbf{Z}}_{[N]}\right\|^2 \sum_{l=1}^{L} \rho_l}{N}}\right) \\
&+ 4B_\phi \left(\sqrt{\frac{\log(\frac{1}{\nu})}{n}} + \|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\|_{\mathrm{TV}}\right) + \min_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}}),
\end{aligned}
$$

where s_l, ρ_l are constants as defined in Lemma 3.

The proof of Theorem 3 is deferred to Appendix D.3. Our observations are similar to those in the CE scenario: since we train the encoder only on the pre-training dataset, during downstream learning, we mainly contend with the intricacies of a smaller model class, which results in the generalization bound of $O(\frac{\mathcal{C}(\mathcal{F})}{\sqrt{n}} + \frac{\mathcal{C}(\mathcal{G}\circ\mathcal{H})}{\sqrt{N}})$. Meanwhile, it's worth noting that the introduction of masking may potentially reduce the norm of the data, denoted as $\left\|\tilde{\mathbf{Z}}_{[N]}\right\|^2$, thereby diminishing the influence of the second term. On the other hand, it could also amplify the domain discrepancy, i.e., $\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\|_{\mathrm{TV}}$.

6 Effective Learning via Rademacher Representation Regularization

As shown in Theorem 1, a significant quantity that affects the generalization risk of downstream task is $\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi \circ \mathcal{F} \circ \hat{h})$, the Rademacher complexity of $\mathcal{F}$ given learnt representation function $\hat{h}$. Here we devise an algorithm to leverage the unlabeled downstream data in the pre-training stage, to regularize the representation function and further improve the accuracy of fine-tuned model.

Let us first consider binary classification case with binary label $y_i \in \{-1, +1\}$. The idea is that, in the binary classification setting, the Rademacher complexity is *independent* of labels, and hence it can be precisely estimated by only unlabeled downstream dataset. If we assume ϕ is G_ϕ Lipschitz, then according to the contraction property of Rademacher complexity [Ledoux and Talagrand, 2013], we have: $\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi \circ \mathcal{F} \circ \hat{h}) \leq G_\phi \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(\hat{h}(\mathbf{x}_i))\right]$, where we can see that, due to the randomness of Rademacher variables, the upper bound of Rademacher complexity can be estimated *without* knowing the actual labels $\{y_i\}_{i=1}^n$. Hence, we can leverage the unlabeled downstream data in the pre-training stage, to regularize the representation function. This can be cast as the following problem:

$$
\min_{g\in\mathcal{H}, h\in\mathcal{H}} \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) + \lambda \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(h(\mathbf{x}_i))\right],
$$

where λ is the regularization coefficient. The idea can also be generalized to multi-class classification, where we use a vector contraction lemma to estimate the upper bound of this complexity which we discuss in Section 6.1. To estimate the expectation, we sample B configurations of Rademacher variables $\{\boldsymbol{\sigma}^j = [\sigma_1^j, ..., \sigma_n^j]\}_{j=1}^B$. If we assume f and h are parameterized by $\mathbf{v} \in \mathcal{V}$ and $\mathbf{w} \in \mathcal{W}$, respectively, we have

$$
\min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_{\widehat{\mathcal{U}}}(\mathbf{w}) + \frac{\lambda}{B}\sum_{j=1}^{B}\left[\max_{\mathbf{v}_j\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} \mathfrak{R}_j(\mathbf{v}_j, \mathbf{w}; \mathbf{x}_i)\right] \quad (8)
$$

where $\mathfrak{R}_j(\mathbf{v}, \mathbf{w}; \mathbf{x}_i) := \sigma_i^j f_{\mathbf{v}}(h_{\mathbf{w}}(\mathbf{x}_i))$.

Optimization method. To solve the aforementioned optimization problems, we adapt the celebrated SGDA algorithm [Lin et al., 2019] (Algorithm 1). At the beginning of each iteration, we first sample a batch of n' pre-training data $\{\mathbf{z}_i^t\}_{i=1}^{n'}$, and then do mini-batch

Algorithm 1: RadReg: Rademacher Regularized Pre-training

Input: Number of iterations T; regularization parameter λ

Sample B configurations of Rademacher variables $\{\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^B\}$,

Initialize $\mathbf{v}_j^0 = \mathbf{0},\ \forall j \in [B]$

for $t = 0, \ldots, T-1$ **do**

  Sample a batch of data from pre-training dataset $\{\tilde{\mathbf{z}}_1^t, \ldots, \tilde{\mathbf{z}}_{n'}^t\}$

  Sample a batch of data from downstreaming dataset $\{\tilde{\mathbf{x}}_1^t, ..., \tilde{\mathbf{x}}_{n'}^t\}$

  for $j = 1, ..., B$ **do**

    $\mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \gamma \frac{1}{n'} \sum_{i=1}^{n'} \nabla_{\mathbf{v}} \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t; \tilde{\mathbf{x}}_i^t)$.
    # Dual variable update

  end

  $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \frac{1}{n'} \sum_{i=1}^{n'} \nabla \mathcal{L}_{\widehat{\mathcal{U}}}(\mathbf{w}^t; \tilde{\mathbf{z}}_i^t)$
  $- \eta\lambda \frac{1}{B} \sum_{j=1}^{B} \left[\frac{1}{n'} \sum_{i=1}^{n'} \nabla_{\mathbf{w}} \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t; \tilde{\mathbf{x}}_i^t) \right]$
  # Representation model update

end

Output: $\hat{\mathbf{w}}$ uniformly sampled from $\{\mathbf{w}^t\}_{t=1}^T$.

stochastic gradient descent:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \frac{1}{n'} \sum\nolimits_{i=1}^{n'} \nabla \mathcal{L}_{\widehat{\mathcal{U}}}(\mathbf{w}^t; \tilde{\mathbf{z}}_i^t) - \eta\lambda \frac{1}{B} \sum\nolimits_{j=1}^{B} \left[\frac{1}{n'} \sum\nolimits_{i=1}^{n'} \nabla_{\mathbf{w}} \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t; \tilde{\mathbf{x}}_i^t) \right].$$

To solving the inner max problem, we sample another batch of n' downstream (unlabeled) data, and then we do one step mini-batch stochastic gradient ascent:$\mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \gamma \frac{1}{n'} \sum_{i=1}^{n'} \nabla_{\mathbf{v}} \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t; \tilde{\mathbf{x}}_i^t)$.

Convergence analysis of RadReg. To establish the convergence of RadReg on (8), we consider the following primal function:$\Psi(\mathbf{w}) := \mathcal{L}_{\widehat{\mathcal{U}}}(\mathbf{w}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \left[\max_{\mathbf{v}_j \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} \mathfrak{R}_j(\mathbf{v}_j, \mathbf{w}; \mathbf{x}_i) \right]$. Then we follow [Lin et al., 2020] and consider the following Moreau envelope function:

Definition 3 (Moreau Envelope)**.** *A function $\Psi_\rho(\mathbf{w})$ is the ρ-Moreau envelope of a function Ψ if $\Psi_\rho(\mathbf{w}) := \min_{\mathbf{w}' \in \mathcal{W}} \{\Psi(\mathbf{w}') + \frac{1}{2\rho} \|\mathbf{w}' - \mathbf{w}\|^2\}$.*

Theorem 4 (Convergence of RadReg with Linear Top Layer, Informal)**.** *RadReg (Algorithm 1 converge to ϵ-stationary point of $\Psi_{1/4L}(\mathbf{w})$ with gradient complexity bounded by $O\left(B/\epsilon^8\right)$.*

The formal version of Theorem 4 as well as the proof is deferred to Appendix E. We can see that the proposed optimization algorithm can find an ϵ-stationary point with at most $O\left(B/\epsilon^8\right)$ stochastic gradient evaluations. Given that complexity increases with respect to B, it becomes crucial to have an appropriately sized sample of Rademacher variable.

6.1 Multi-class

The proposed regularization idea can also be generalized to multi-class classification. If the model $f(\cdot)$ is a vector-valued function, i.e., in multi-class classification, $f(h(\mathbf{x})) : \mathcal{X} \mapsto \mathbb{R}^o$, we can apply the following *vector-valued contraction lemma* of Rademacher complexity:

Lemma 4. *[Maurer, 2016] Let $\phi(\cdot, \cdot) : \mathbb{R}^o \mapsto \mathbb{R}$ be G_ϕ-Lipschitz in the first argument, and $f(\cdot) : \mathbb{R}^{d'} \mapsto \mathbb{R}^o$ be vector-valued function. Then, the following facts hold true for Rademacher complexity over $\mathcal{F}$ and any $h : \mathcal{X} \mapsto \mathbb{R}^{d'}$:*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi(f(h(\mathbf{x}_i)), y_i) \right] \leq \sqrt{2} G_\phi \mathbb{E}_{\boldsymbol{\varepsilon}_i} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i f(h(\mathbf{x}_i)) \right],$$

where $\boldsymbol{\varepsilon}_i \in \{-1, +1\}^o$ is Rademacher vector.

Now the empirical minimization problem becomes:

$$\min_{\mathbf{w} \in \mathcal{W}} \mathcal{L}_{\widehat{\mathcal{U}}}(\mathbf{w}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \left[\max_{\mathbf{V} \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\sigma}_i^j)^\top f(h(\mathbf{x}_i)) \right].$$

Specifically, if the top layer is a linear projection layer, i.e., $f(h(\mathbf{x})) = \mathbf{V} h(\mathbf{x})$, the objective is equivalent to:

$$\min_{\mathbf{w} \in \mathcal{W}} \mathcal{L}_{\widehat{\mathcal{U}}}(\mathbf{w}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \left[\max_{\mathbf{V} \in \mathcal{V}} \operatorname{tr} \left(\mathbf{V} \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{x}_i) (\boldsymbol{\sigma}_i^j)^\top \right) \right],$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. Here the inner problem is convex and easy to solve with simple (stochastic) gradient ascent.

7 Experiments

In this section, we empirically evaluate the proposed regularization method in improving the generalization of unsupervised pre-training to downstream tasks. We utilize the Masked AutoEncoder (MAE) [He et al., 2022] as the base unsupervised pre-training method. We conduct experiments using 50,000 images from CIFAR10 dataset [Krizhevsky et al., 2009] for pre-training and 4,096 few-shot STL [Coates et al., 2011] samples for finetuning. Since our regularization requires unlabeled data from the downstream task, but L2 and non-regularization methods cannot leverage those data, for a fair comparison, we incorporate the fine-tuning data into a separate unsupervised loss with the same

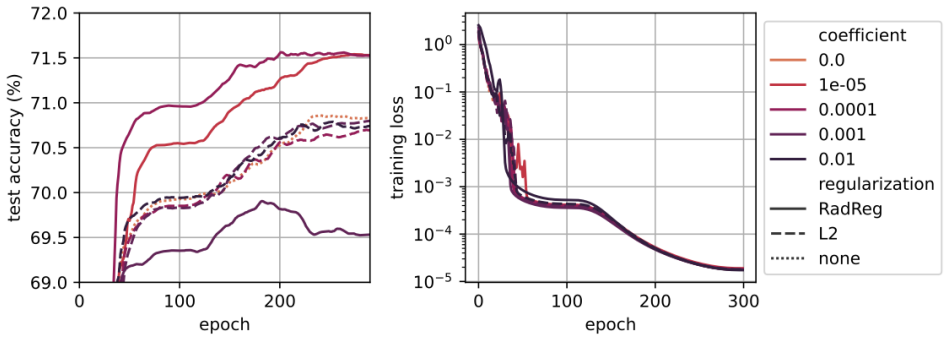

Figure 1: Testing and training accuracy by epochs, averaged by three repetitions.

Reg.	λ	Final Acc	Best Acc	Train Acc
None	-	70.8 (0.2)	70.9 (0.3)	100 (0.)
L_2	10^{-4}	70.7 (0.3)	70.7 (0.3)	100 (0.)
	10^{-3}	70.8 (0.6)	70.9 (0.5)	100 (0.)
	10^{-2}	70.7 (0.6)	70.9 (0.5)	100 (0.)
RadReg	10^{-5}	**71.5** (0.2)	**71.8** (0.3)	100 (0.)
	10^{-4}	71.5 (0.4)	71.6 (0.7)	100 (0.)
	10^{-3}	69.6 (0.6)	69.6 (0.5)	100 (0.)

(a) End-to-end fine-tuning

Reg.	λ	Final Acc	Best Acc	Train Acc
None	-	55.3 (0.1)	55.5 (0.0)	65.7 (0.2)
L_2	10^{-5}	55.3 (0.1)	55.5 (0.0)	57.5 (0.1)
	10^{-4}	55.9 (0.1)	56.0 (0.1)	58.2 (0.1)
	10^{-3}	54.4 (0.0)	54.5 (0.0)	58.2 (0.1)
RadReg	10^{-5}	**56.8** (0.0)	**56.9** (0.1)	65.7 (0.2)
	10^{-4}	26.8 (0.0)	27.0 (0.1)	40.6 (0.1)

(b) Linear fine-tuning

Table 1: Evaluation of MAE. Average fine-tuning accuracy is reported with its standard deviations in brackets.

formulation as the MAE loss:

$$\min_{g,h} \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) + \alpha \cdot \mathcal{L}_{\widehat{\mathcal{D}}}(g \circ h) + \lambda \mathfrak{R}_{\widehat{\mathcal{T}}}(\mathcal{F} \circ h) \text{ (RadReg)},$$

$$\min_{g,h} \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) + \alpha \cdot \mathcal{L}_{\widehat{\mathcal{D}}}(g \circ h) + \lambda \|\mathbf{W}\|^2 \quad (L_2),$$

$$\min_{g,h} \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) + \alpha \cdot \mathcal{L}_{\widehat{\mathcal{D}}}(g \circ h) \quad \text{(Non-regularized)},$$

where we assume h is parameterized by $\mathbf{W}$ and α is fixed as 0.01. Our proposed regularization will further leverage the data to control the complexity of learned representations. The details of experiments are included in Appendix F.

In Table 1, we compare our method to non-regularized MAE training and the one with L_2 regularization. We repeat the fine-tuning three times by randomly selecting 4096 samples from the preset STL10 training set, and report the mean and standard deviations. We vary the coefficient for our and L_2 regularization and compare the best test accuracy on fine-tuning. We observe that our method can effectively improve the downstream performance as early as in the pre-training stage without using any labels. Compared to L_2 regularization, our method can achieve higher test accuracy.

In Figure 1, we show the learning curves by different regularization strategies. Due to the large capacity of the pre-trained ViT encoder, all methods can sufficiently fit the training set approaching 100% training accuracy, but the testing accuracy reaches the ceiling. Our method can improve the best test accuracy by limiting the representation complexity as early as the pre-training stage. Our method also improves the convergence rate at fine-tuning, when our method reaches the 71% test accuracy at epoch 80 but the best baseline reaches the same accuracy after 200 epochs.

8 Conclusion

This paper establishes a generic learning bound in unsupervised representation pre-training and fine-tuning paradigm. We discover that the generalization depends on representation transferrbaility, representation-induced Rademacher complexity, task heterogeneity and generalization of pre-training task. We apply our theory to analyze the generalization of CE and MAE pre-training. Motivated by our theory, we propose Rademacher representation regularization, with a provable convergence guarantee. The experiments validate the superiority of our algorithm. As a future direction, it would be interesting to expand our analysis to end-to-end model fine-tuning, where task specific head and encoder are jointly updated in fine-tuning stage.

Acknowledgement

The work of YD and MM was partially supported by NSF CAREER Award #2239374 and NSF CNS Award #1956276. JZ was supported by NSF #IIS-2212174 and IIS-1749940 and NIA #1RF1AG072449.

References

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019a.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019b.

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

Olivier Bousquet. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. 01 2002.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*, pages 561–580. Springer, 2012.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. *arXiv preprint arXiv:2303.01566*, 2023.

Henry Gouk, Timothy Hospedales, et al. Distance-based regularisation of deep networks for fine-tuning. In *International Conference on Learning Representations*, 2020.

Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Elad Hazan and Tengyu Ma. A non-generative framework and convex relaxations for unsupervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Haotian Ju, Dongyue Li, and Hongyang R Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. In *International Conference on Machine Learning*, pages 10431–10461. PMLR, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Dongyue Li and Hongyang Zhang. Improved regularization and robustness for fine-tuning in neural networks. *Advances in Neural Information Processing Systems*, 34:27249–27262, 2021.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

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

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33: 512–523, 2020.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *arXiv preprint arXiv:2206.03126*, 2022.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex concave min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Axel Ruhe. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10:343–354, 1970. URL https://api.semanticscholar.org/CorpusID:122004897.

Gal Shachaf, Alon Brutzkus, and Amir Globerson. A theoretical analysis of fine-tuning with linear teachers. *Advances in Neural Information Processing Systems*, 34:15382–15394, 2021.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *17th European Conference Computer Vision–ECCV 2022:*, pages 68–85. Springer, 2022.

Ziping Xu and Ambuj Tewari. Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34:4792–4804, 2021.

Fan Yang, Hongyang R Zhang, Sen Wu, Weijie J Su, and Christopher Ré. Analysis of information transfer from heterogeneous sources via precise high-dimensional asymptotics. *arXiv preprint arXiv:2010.11750*, 2020.

Jieyu Zhang, Bohan Wang, Zhengyu Hu, Pang Wei Koh, and Alexander Ratner. On the trade-off of intra-/inter-class diversity for supervised pre-training. *arXiv preprint arXiv:2305.12224*, 2023.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *arXiv preprint arXiv:2210.08344*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Organization The appendix is organized as follows. In Appendix A we will introduce some helper inequalities that we will be utilized in our proofs and prove the main generalization theorem in Appendix B. In Appendices C and D we will provide the proofs in Section 5.1 (generalization of pre-training with a context encoder) and Section 5.2 (generalization of pre-training with masked autoencoder with a transformer), respectively. In Appendix E, we provide the proof of convergence of the proposed algorithm in Section 6. At last, in Appendix F we will provide the details of setup for our experiments and report additional results.

A Basic Inequalities

In this section, we provide some general technical results that will be used in our proofs.

Proposition 1 (Total variation distance and L_1 distance). *[Levin and Peres, 2017, Proposition 4.2] Given two probability measures $\mathcal{P}$ and $\mathcal{Q}$ defined over instance space $\mathcal{X}$, the following inequality holds:*

$$\|\mathcal{P} - \mathcal{Q}\|_{\mathrm{TV}} = \frac{1}{2} \sum_{\mathbf{x} \in \mathcal{X}} |\mathcal{P}(\mathbf{x}) - \mathcal{Q}(\mathbf{x})|.$$

Proposition 2 (Ruhe's trace inequality). *[Ruhe, 1970] If $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite Hermitian matrices with eigenvalues,*

$$a_1 \geq ... \geq a_n \geq 0, \;\; b_1 \geq ... \geq b_n \geq 0, \tag{9}$$

repsectively, then

$$\sum_{i=1}^{n} a_i b_{n-i+1} \leq \mathsf{tr}\left(\mathbf{AB}\right) \leq \sum_{i=1}^{n} a_i b_i.$$

B Proof of Main Generalization Theorem

In this section we provide the proof of main result on generalization of fine-tuned model composed with an unsupervised pre-trained model stated in Theorem 1. For readability purposes, we re-state the theorem here:

Theorem 5 (Theorem 1 restated). *Assume $\hat{h}$ and $\hat{g}$ are the pre-trained representation function and its associated decoder function, and real valued non-negative loss ϕ to be G_ϕ Lipschitz and bounded by B_ϕ. Assume pre-training and fine-tuning task admit a (C_β, β) representation transferrability on $\hat{h}$ and $h^*_{\mathcal{U}}$. If we solve (2) to get $\hat{f}$, then with probability at least $1 - \nu$, the following statement holds*

$$\mathcal{E}_{\mathcal{T}}(\hat{f}, \hat{h}) \leq C_\beta \left(\mathcal{E}_{\mathcal{U}}(\hat{g}, \hat{h})\right)^\beta + 4G_\phi \mathfrak{R}_{\widehat{\mathcal{T}}}(\mathcal{F} \circ \hat{h}) + 4B_\phi \sqrt{\frac{\log(1/\nu)}{n}} + 4B_\phi \left\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}} + \min_{f \in \mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}}),$$

where $h^*_{\mathcal{U}} = \arg\min_{h \in \mathcal{H}} \min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{U}}(g \circ h)$ is the optimal pre-training representation function, and $\|\mathcal{P} - \mathcal{Q}\|_{\mathrm{TV}} = \sup_{A \in \Omega} |\mathcal{P}(A) - \mathcal{Q}(A)|$ denotes total variation distance between two distributions.

B.1 Proof of Theorem 1

Proof. For the ease of presentation we define

$$f^*_{\mathcal{T}}(h) = \arg\min_{f \in \mathcal{F}} \mathcal{R}_{\mathcal{T}}(f \circ h) := \mathbb{E}_{\mathbf{x} \sim \mathcal{T}}[\phi(f \circ h(\mathbf{x}), y(\mathbf{x}))].$$

That is, the optimal fine-tuned risk minimizer in function class $\mathcal{F}$ w.r.t. distribution $\mathcal{T}$ over domain, given a representation function h, which denotes the optimal risk minimizer for downstream task with labeling function $y(\cdot)$, for a given representation function. Also, recall $h^*_{\mathcal{U}} = \arg\min_{h \in \mathcal{H}} \min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{U}}(g \circ h)$ denotes the optimal pre-training representation function.

By standard risk decomposition we have:

$$
\begin{aligned}
\mathcal{E}_{\mathcal{T}}(\hat{f},\hat{h}) &= \mathcal{R}_{\mathcal{T}}(\hat{f}\circ\hat{h}) - \min_{f\in\mathcal{F},h\in\mathcal{H}} \mathcal{R}_{\mathcal{T}}(f\circ h) \\
&= \mathcal{R}_{\mathcal{T}}(\hat{f}\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h}) + \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h}) - \min_{f\in\mathcal{F},h\in\mathcal{H}} \mathcal{R}_{\mathcal{T}}(f\circ h) \\
&= \underbrace{\mathcal{R}_{\mathcal{T}}(\hat{f}\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h})}_{\mathsf{I}} \\
&\quad + \underbrace{\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ h^*_{\mathcal{U}})}_{\mathsf{II}} \\
&\quad + \underbrace{\left(\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ\hat{h})\right)}_{\mathsf{III}} \\
&\quad - \underbrace{\left(\min_{f\in\mathcal{F},h\in\mathcal{H}} \mathcal{R}_{\mathcal{T}}(f\circ h) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ h^*_{\mathcal{U}})\right)}_{\mathsf{IV}}
\end{aligned}
$$

We now turn to bounding each term in RHS of above inequality.

Bounding I. The term I can be bounded by following standard results in uniform convergence and noting the fact that $\hat{f}$ is empirical risk minimizer of downstream task by fixing the pre-training representation function $\hat{h}$:

$$
\begin{aligned}
\mathsf{I} &= \mathcal{R}_{\mathcal{T}}(\hat{f}\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h}) \\
&= \mathcal{R}_{\mathcal{T}}(\hat{f}\circ\hat{h}) - \mathcal{R}_{\widehat{\mathcal{T}}}(\hat{f}\circ\hat{h}) + \underbrace{\mathcal{R}_{\widehat{\mathcal{T}}}(\hat{f}\circ\hat{h}) - \mathcal{R}_{\widehat{\mathcal{T}}}(f^*_{\mathcal{T}}(\hat{h})\circ\hat{h})}_{\leq 0} + \mathcal{R}_{\widehat{\mathcal{T}}}(f^*_{\mathcal{T}}(\hat{h})\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h}) \\
&\leq 4\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi\circ\mathcal{F}\circ\hat{h}) + 4B_\phi\sqrt{\frac{\log(1/\nu)}{n}}.
\end{aligned}
$$

Bounding III. To bound III, we define $f^*_{\mathcal{U}}(h) = \arg\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ\hat{h})$, where $\mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ h) := \mathbb{E}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{X}}}[\phi(f\circ h(\mathbf{x}), y(\mathbf{x}))]$ denotes the risk realized by pre-training marginal data distribution $\mathcal{U}_{\mathcal{X}}$ and downstream labeling function $y(\cdot)$ (Definition 2). We have:

$$
\begin{aligned}
\mathsf{III} &= \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ\hat{h}) \leq \mathcal{R}_{\mathcal{T}}(f^*_{\mathcal{U}}(\hat{h})\circ\hat{h}) - \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f^*_{\mathcal{U}}(\hat{h})\circ\hat{h}) \\
&= \mathbb{E}_{\mathbf{x}\sim\mathcal{T}}[\phi(f^*_{\mathcal{U}}(\hat{h})\circ\hat{h}(\mathbf{x}),\mathbf{y})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{U}}[\phi(f^*_{\mathcal{U}}(\hat{h})\circ\hat{h}(\mathbf{x}),\mathbf{y})] \\
&= \sum_{\mathbf{x}\in\mathcal{X}} |\mathcal{T}(\mathbf{x}) - \mathcal{U}_{\mathcal{X}}(\mathbf{x})|\cdot\phi(f^*_{\mathcal{U}}(\hat{h})\circ\hat{h}(\mathbf{x}),\mathbf{y}) \\
&\leq B_\phi \sum_{\mathbf{x}\in\mathcal{X}} |\mathcal{T}(\mathbf{x}) - \mathcal{U}_{\mathcal{X}}(\mathbf{x})| \\
&\leq 2B_\phi \left\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}}.
\end{aligned}
$$

where the last step follows from Proposition 1.

Bounding IV. For IV, recalling that $f^*_{\mathcal{T}}(h) = \arg\min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ h)$, and we have

$$
\begin{aligned}
\mathsf{IV} &= \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ h^*_{\mathcal{U}}) - \min_{f\in\mathcal{F},h\in\mathcal{H}} \mathcal{R}_{\mathcal{T}}(f\circ h) \\
&= \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ h^*_{\mathcal{U}}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ h^*_{\mathcal{U}}) + \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ h^*_{\mathcal{U}}) - \min_{f\in\mathcal{F},h\in\mathcal{H}} \mathcal{R}_{\mathcal{T}}(f\circ h) \\
&\leq \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f^*_{\mathcal{T}}(h^*_{\mathcal{U}})\circ h^*_{\mathcal{U}}) - \mathcal{R}_{\mathcal{T}}(f^*_{\mathcal{T}}(h^*_{\mathcal{U}})\circ h^*_{\mathcal{U}}) + \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{T}}(f\circ h^*_{\mathcal{U}}) - \min_{f\in\mathcal{F},h\in\mathcal{H}} \mathcal{R}_{\mathcal{T}}(f\circ h) \\
&\leq 2B_\phi \left\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}} + \min_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}})
\end{aligned}
$$

where at last step we use the same reasoning we used in bounding III, and the definition of $\mathcal{E}_{\mathcal{T}}(\cdot)$.

Bounding II. It remains to bound II. Under the representation transferability assumption, we know

$$\begin{aligned}
\text{II} &= \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ\hat{h}) - \min_{f\in\mathcal{F}} \mathcal{R}_{\mathcal{U}_{\mathcal{X}}}(f\circ h^*_{\mathcal{U}}) \\
&\le C_\beta \left(\min_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{U}}(g\circ\hat{h}) - \min_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{U}}(g\circ h^*_{\mathcal{U}})\right)^\beta \\
&\le C_\beta \left(\mathcal{L}_{\mathcal{U}}(\hat{g}\circ\hat{h}) - \min_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{U}}(g\circ h^*_{\mathcal{U}})\right)^\beta \\
&= C_\beta \left(\mathcal{L}_{\mathcal{U}}(\hat{g}\circ\hat{h}) - \min_{g\in\mathcal{G},h\in\mathcal{H}} \mathcal{L}_{\mathcal{U}}(g\circ h)\right)^\beta \\
&= C_\beta \left(\mathcal{E}_{\mathcal{U}}(\hat{g},\hat{h})\right)^\beta.
\end{aligned}$$

where the last step follows from the definition of $\mathcal{E}_{\mathcal{U}}(\cdot)$.

Putting pieces I-IV together yields:

$$\begin{aligned}
\mathcal{E}_{\mathcal{T}}(\hat{f},\hat{h}) \le C_\beta \left(\mathcal{E}_{\mathcal{U}}(\hat{g},\hat{h})\right)^\beta + 4G_\phi \mathfrak{R}_{\widehat{\mathcal{T}}}(\mathcal{F}\circ\hat{h}) + 4B_\phi\sqrt{\frac{\log(1/\nu)}{n}} \\
+ 4B_\phi \left\|\mathcal{T}-\mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}} + \min_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}}),
\end{aligned}$$

thus leading to the desired generalization bound stated in Theorem 1.

□

As mentioned earlier, to instantiate Theorem 1 to a particular application, we need to establish bounds on representation transferrability, generalization of pre-training task, and representation-induced Rademacher complexity as we demonstrate on two specific pre-training tasks. We note that similar notions to representation transferrability were proposed in [Tripuraneni et al., 2020, Ge et al., 2023, Du et al., 2020, Zhang et al., 2023], but they do not have exponent in definition, so cannot capture the transferrability when pre-training and downstream task losses are not homogeneous. The term $\min_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}})$ characterizes how well the optimal pre-training task encoder is when applied on downstream task. It will depend on specific pre-training and downstream distribution. Since we do not make distributional assumption, analyzing this term is beyond the scope of this paper.

C Proof of Generalization for Pre-training with Context Encoder

In this section we prove the results on generalization of pre-training with Context Encoder (CE) and fine-tuning on binary classification as downstream task provided in Subsection 5.1. Recall during pre-training, we draw a set of unlabeled data,e.g., images $\{\mathbf{z}_1,...,\mathbf{z}_N\}$, and corrupt these data to make $\{\tilde{\mathbf{z}}_1,...,\tilde{\mathbf{z}}_N\}$, then a deep neural network is trained by reconstructing the corrupted pixel of a given image. The encoder-decoder architecture is defined as follows:

$$\begin{aligned}
&\text{encoder:} \quad h(\mathbf{x}) = \sigma\left(\mathbf{W}_L \cdots \sigma\left(\mathbf{W}_1\mathbf{x}\right)\right), \\
&\text{decoder:} \quad g(h(\mathbf{x})) = \mathbf{W}_{L+1} h(\mathbf{x}).
\end{aligned}$$

where $\mathbf{W}_1 \in \mathbb{R}^{m\times d}$, $\mathbf{W}_2,...,\mathbf{W}_L \in \mathbb{R}^{m\times m}$, and $\mathbf{W}_{L+1} \in \mathbb{R}^{d\times m}$ (for simplicity we assume the hiddent layers share the same dimension m). We assume each layer's weight is with bounded norm: $\|\mathbf{W}_l\| \le W(l)$, $\|\mathbf{W}_l\|_{2,1} \le B(l), \forall l \in [L+1]$. The hypothesis class for encoder is then defined as:

$$\mathcal{H} := \left\{\mathbf{x} \mapsto \sigma\left(\mathbf{W}_L \cdots \sigma\left(\mathbf{W}_1\mathbf{x}\right)\right) : \|\mathbf{W}_l\| \le W(l), \|\mathbf{W}_l\|_{2,1} \le B(l), \forall l \in [L]\right\}$$

and decoder class is defined as:

$$\mathcal{G} := \left\{\mathbf{x} \mapsto \mathbf{W}_{L+1}\mathbf{x} : \|\mathbf{W}_{L+1}\| \le W(L+1), \|\mathbf{W}_{L+1}\|_{2,1} \le B(L+1)\right\}.$$

In pre-training stage we optimize the following empirical unsupervised losses:

$$\min_{g\in\mathcal{G},h\in\mathcal{H}} \mathcal{L}_{\widehat{\mathcal{U}}}(g\circ h) := \frac{1}{2}\sum_{i=1}^{N} \|g(h(\tilde{\mathbf{z}}_i)) - \mathbf{z}_i\|^2, \tag{10}$$

where $\tilde{\mathbf{z}}_i = T_1(\mathbf{z}_i)$, and $T_1 : \mathcal{X} \mapsto \mathcal{X}$ is some random transformation, e.g, rotating, scaling, adding Gaussian noise or masking pixels.

After pre-training, we discard the top layer of the network, and use the rest layers as an encoder. A linear projection head is added on top of encoder in downstream training:

$$\text{downstream model:} \quad f(\hat{h}(\mathbf{x})) = \boldsymbol{\theta}^\top \hat{h}(\mathbf{x}),$$

with $\|\boldsymbol{\theta}\| \le R$, and we assume that only the linear head is trainable during fine-tune stage. We optimize a binary classification task with Lipschitz loss function as fine-tuning task:

$$\min_{\|\boldsymbol{\theta}\|\le R} \mathcal{R}_{\widehat{\mathcal{T}}}(\boldsymbol{\theta}\circ\hat{h}) = \frac{1}{n}\sum\nolimits_{i=1}^{n} \phi(\boldsymbol{\theta}^\top h(\mathbf{x}_i), y_i),$$

to get $\hat{f}$, where $y_i \in \{-1, +1\}$ is binary labeling function for downstream task.

Roadmap. We will provide proof of Theorem 2 in the following subsections. The roadmap is as follows: in Appendix C.1 we first show that the CE pre-training admits bounded representation transferrability to downstream task (the proof of Lemma 1), and then in Appendix C.2 we prove the generalization of CE pre-training task (Lemma 5), and finally in Appendix C.3 we conclude the proof for Theorem 2 by showing that the representation-induced Rademacher complexity is bounded.

C.1 Proof of Transferability

In this subsection we provide the proof of Lemma 1. For notational convenience we define the following quantities:

$$\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) = \min_{\|\boldsymbol{\theta}\|\le R} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}}[\phi(\boldsymbol{\theta}^\top\hat{h}(\tilde{\mathbf{z}}))] - \min_{\|\tilde{\boldsymbol{\theta}}\|\le R} \mathbb{E}_{\tilde{\mathbf{z}}\sim\mathcal{U}_{\mathcal{X}}}[\phi(\tilde{\boldsymbol{\theta}}^\top h^*_{\mathcal{U}}(\tilde{\mathbf{z}}))],$$

$$\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) = \min_{\mathbf{W}_{L+1}} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left\|\mathbf{W}_{L+1}\hat{h}(\tilde{\mathbf{z}}) - \mathbf{z}\right\|^2 - \min_{\widetilde{\mathbf{W}}_{L+1}} \mathbb{E}_{\tilde{\mathbf{z}}\sim\mathcal{U}_{\mathcal{X}}} \left\|\widetilde{\mathbf{W}}_{L+1}h^*_{\mathcal{U}}(\tilde{\mathbf{z}}) - \mathbf{z}\right\|^2$$

To prove Lemma 1, we are going to show $\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) \le C_\beta \left(\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}})\right)^\beta$ holds for some C_β, β.

Upper bounding $\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}})$: We examine $\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}})$ first. We define the optimal head for classification task on distribution $\mathcal{U}_{\mathcal{X}}$ under represetation $h^*_{\mathcal{U}}$ as $\tilde{\boldsymbol{\theta}}^* = \arg\min_{\|\tilde{\boldsymbol{\theta}}\|\le R} \mathbb{E}_{\tilde{\mathbf{z}}\sim\mathcal{U}_{\mathcal{X}}}[\phi(\tilde{\boldsymbol{\theta}}^\top h^*_{\mathcal{U}}(\tilde{\mathbf{z}}))]$.

$$\begin{aligned}
\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*) &= \min_{\|\boldsymbol{\theta}\|\le R} \mathbb{E}_{\mathbf{x}\sim\mathcal{U}}[\phi(\boldsymbol{\theta}^\top\hat{h}(\tilde{\mathbf{z}}))] - \mathbb{E}_{\mathbf{x}\sim\mathcal{U}}[\phi(\tilde{\boldsymbol{\theta}}^{*\top} h^*_{\mathcal{U}}(\tilde{\mathbf{z}}))] \\
&\le \min_{\|\boldsymbol{\theta}\|\le R} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left|\boldsymbol{\theta}^\top\hat{h}(\tilde{\mathbf{z}}) - \tilde{\boldsymbol{\theta}}^{*\top} h^*_{\mathcal{U}}(\tilde{\mathbf{z}})\right| \\
&\le \min_{\|\boldsymbol{\theta}\|\le R} \sqrt{\mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left(\boldsymbol{\theta}^\top\hat{h}(\tilde{\mathbf{z}}) - \tilde{\boldsymbol{\theta}}^{*\top} h^*_{\mathcal{U}}(\tilde{\mathbf{z}})\right)^2} \\
&= \min_{\|\boldsymbol{\theta}\|\le R} \sqrt{\boldsymbol{\theta}^\top \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}})\hat{h}^\top(\tilde{\mathbf{z}})\right]\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}})h^{*\top}_{\mathcal{U}}(\tilde{\mathbf{z}})\right]\tilde{\boldsymbol{\theta}}^* + \tilde{\boldsymbol{\theta}}^{*\top}\mathbb{E}\left[h^*_{\mathcal{U}}(\tilde{\mathbf{z}})h^{*\top}_{\mathcal{U}}(\tilde{\mathbf{z}})\right]\tilde{\boldsymbol{\theta}}^*}
\end{aligned}$$

Since $\sqrt{f(x)}$ and $f(x)$ attain the minimum at the same point, we examine the minimum of $\boldsymbol{\theta}^\top \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}})\hat{h}^\top(\tilde{\mathbf{z}})\right]\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}})h^{*\top}_{\mathcal{U}}(\tilde{\mathbf{z}})\right]\tilde{\boldsymbol{\theta}}^* + \tilde{\boldsymbol{\theta}}^{*\top}\mathbb{E}\left[h^*_{\mathcal{U}}(\tilde{\mathbf{z}}){h_{\mathcal{U}}}^{*\top}(\tilde{\mathbf{z}})\right]\tilde{\boldsymbol{\theta}}^*$ over $\boldsymbol{\theta}$. Under unconstrained setting, the minimum of above statement is $\tilde{\boldsymbol{\theta}}^{*\top}\Lambda\tilde{\boldsymbol{\theta}}^*$

when $\boldsymbol{\theta} = \left(\mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}})\hat{h}^\top(\tilde{\mathbf{z}})\right]\right)^\dagger \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}}){h^*_{\mathcal{U}}}^\top(\tilde{\mathbf{z}})\right]\tilde{\boldsymbol{\theta}}^*$, and

$$\Lambda = \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}})\hat{h}^\top(\tilde{\mathbf{z}})\right] - \mathbb{E}\left[h^*_{\mathcal{U}}(\tilde{\mathbf{z}})\hat{h}^\top(\tilde{\mathbf{z}})\right]\left(\mathbb{E}\left[h^*_{\mathcal{U}}(\tilde{\mathbf{z}}){h^*_{\mathcal{U}}}^\top(\tilde{\mathbf{z}})\right]\right)^\dagger \mathbb{E}\left[\hat{h}(\tilde{\mathbf{z}}){h^*_{\mathcal{U}}}^\top(\tilde{\mathbf{z}})\right].$$

Hence we have

$$\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) \le \sqrt{\tilde{\boldsymbol{\theta}}^{*\top}\Lambda\tilde{\boldsymbol{\theta}}^*} = \sqrt{\mathsf{tr}(\Lambda\tilde{\boldsymbol{\theta}}^{*\top}\tilde{\boldsymbol{\theta}}^*)} \le \sqrt{d\sigma_{\max}(\Lambda)\sigma_{\max}(\tilde{\boldsymbol{\theta}}^{*\top}\tilde{\boldsymbol{\theta}}^*)}, \tag{11}$$

where we applied Ruhe's Trace Inequalities at last step (Proposition 2): $\mathsf{tr}(\mathbf{AB}) \le \sum_{i=1}^d \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) \le d\sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{B})$.

Finally, we choose large enough R so that we can attain the optimum.

Lower bounding $\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*)$ Now we switch to lower bounding $\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*)$. We have:

$$\begin{aligned}
\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) &= \min_{\mathbf{W}_{L+1}:\|\mathbf{W}\|\le W(L+1)} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left\|\mathbf{W}_{L+1}\hat{h}(\tilde{\mathbf{z}}) - \mathbf{z}\right\|^2 - \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left\|\mathbf{W}^*_{L+1}h^*_{\mathcal{U}}(\tilde{\mathbf{z}}) - \mathbf{z}\right\|^2 \\
&= \min_{\mathbf{W}_{L+1}:\|\mathbf{W}\|\le W(L+1)} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left\|\mathbf{W}_{L+1}\hat{h}(\tilde{\mathbf{z}}) - \mathbf{W}^*_{L+1}h^*_{\mathcal{U}}(\tilde{\mathbf{z}})\right\|^2
\end{aligned}$$

where the last step is due to our realizability Assumption 1, the optimal encoder-decoder exists in the hypothesis class which can perfectly recover masked data.

Hence

$$\begin{aligned}
\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) &= \min_{\mathbf{W}_{L+1}\in\mathbb{R}^{d\times m}} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left\|\mathbf{W}_{L+1}\hat{h}(\tilde{\mathbf{z}}) - \mathbf{W}^*_{L+1}h^*(\tilde{\mathbf{z}})\right\|^2 \\
&= \min_{\mathbf{w}_r\in\mathbb{R}^m, r\in[d]} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \sum_{r=1}^d \left\|\mathbf{w}_r^\top\hat{h}(\tilde{\mathbf{z}}) - {\mathbf{w}^*_r}^\top h^*(\tilde{\mathbf{z}})\right\|^2 \\
&\ge \sum_{r=1}^d \min_{\mathbf{w}_r\in\mathbb{R}^m} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left\|\mathbf{w}_r^\top\hat{h}(\tilde{\mathbf{z}}) - {\mathbf{w}^*_r}^\top h^*(\tilde{\mathbf{z}})\right\|^2 \\
&\ge \sum_{r=1}^d \min_{\mathbf{w}_r\in\mathbb{R}^m} \mathbb{E}_{(\tilde{\mathbf{z}},\mathbf{z})\sim\mathcal{U}} \left(\mathbf{w}_r^\top\hat{h}(\tilde{\mathbf{z}})\hat{h}(\tilde{\mathbf{z}})^\top\mathbf{w}_r - 2\mathbf{w}_r^\top\hat{h}(\tilde{\mathbf{z}})h^*_{\mathcal{U}}(\tilde{\mathbf{z}})^\top\mathbf{w}^*_r + {\mathbf{w}^*_r}^\top h^*_{\mathcal{U}}(\tilde{\mathbf{z}})h^*_{\mathcal{U}}(\tilde{\mathbf{z}})^\top\mathbf{w}^*_r\right).
\end{aligned}$$

According to similar reasoning in the proof of upper bound, with Λ defined in the same way as (11), we have

$$\begin{aligned}
\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}}) &\ge \sum_{r=1}^d {\mathbf{w}^*_r}^\top\Lambda\mathbf{w}^*_r \\
&= \mathrm{tr}\left(\Lambda\sum_{r=1}^m \mathbf{w}^*_r{\mathbf{w}^*_r}^\top\right) \\
&\ge \sigma_{\max}(\Lambda)\sigma_{\min}\left(\sum_{r=1}^d \mathbf{w}^*_r{\mathbf{w}^*_r}^\top\right)
\end{aligned}$$

where at last step we apply Ruhe's trace inequality (Proposition 2)): $\mathsf{tr}\left(\mathbf{AB}\right) \ge \sigma_{\max}(\mathbf{A})\sigma_{\min}(\mathbf{B})$. Therefore, we can conclude that

$$\frac{\Delta^{ft}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}})}{\left(\Delta^{pt}_{\mathcal{U}}(\hat{h}, h^*_{\mathcal{U}})\right)^{1/2}} \le O\left(\frac{\sqrt{d\sigma_{\max}(\tilde{\boldsymbol{\theta}}^*\tilde{\boldsymbol{\theta}}^{*\top})}}{\sqrt{\sigma_{\min}\left(\sum_{r=1}^d \mathbf{w}^*_r{\mathbf{w}^*_r}^\top\right)}}\right),$$

which indicates that Context Encoder pretraining admits an $\left(\Omega\left(\frac{\sqrt{d\sigma_{\max}(\tilde{\boldsymbol{\theta}}^*\tilde{\boldsymbol{\theta}}^{*\top})}}{\sqrt{\sigma_{\min}\left(\sum_{r=1}^d \mathbf{w}_r^*\mathbf{w}_r^{*\top}\right)}}\right), \frac{1}{2}\right)$ representation transferrability to binary classification task. In the main paper Lemma 1 we omit the constant dependency for ease of exposition.

C.2 Proof of generalization of CE pretraining task

In this section we are going to derive generalization bound of the CE pre-training. The generalization is given in the following lemma:

Lemma 5 (Generalization of pre-training task). *Let $\hat{g}, \hat{h}$ be the solution of (4), and $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_N]$ is the concatenated pre-training data. Then with probability at least 0.99 the following statement holds:*

$$\mathcal{E}_{\mathcal{U}}(\hat{g}, \hat{h}) \leq O\left(\frac{\left(\left\|\tilde{\mathbf{Z}}\right\|^2 \ln(2m^2)\right)\left(\prod_{l=1}^{L+1} W^2(l)\right)\left(\sum_{l=1}^{L+1}(\frac{B(l)}{W(l)})^{\frac{2}{3}}\right)^3}{N}\right).$$

To prove Lemma 5, we first introduce the following worst case covering number quantity:

Definition 4 (L_2 covering number). *Given a hypothesis class $\mathcal{H}$ and a set of data $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, let $h(\mathbf{X}) = [h(\mathbf{x}_1); ...; h(\mathbf{x}_N)]$ denote the concatenated output of N points.The the covering number $\mathcal{N}(\mathcal{H}(\mathcal{S}), \epsilon, \|\cdot\|)$ is the least cardinality of set $\mathcal{C}$, such that for every $h \in \mathcal{H}$, there exists a $h_\epsilon \in \mathcal{C}$, and ensures that*

$$\|h(\mathbf{X}) - h_\epsilon(\mathbf{X})\| \leq \epsilon.$$

Definition 5 (L_∞ covering number). *Given a hypothesis class $\mathcal{H}$ and a set of data $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, the worst case covering number $\mathcal{N}_\infty(\mathcal{H}(\mathcal{S}), \epsilon, \|\cdot\|)$ is the least cardinality of set $\mathcal{C}$, such that for every $h \in \mathcal{H}$, there exists a $h_\epsilon \in \mathcal{C}$, and ensures that*

$$\max_{i\in[N]} \|h(\mathbf{x}_i) - h_\epsilon(\mathbf{x}_i)\| \leq \epsilon.$$

The following result will relate the Rademacher complexity of the local loss class induced by a hypothesis class $\mathcal{H}$, to the L_∞ covering number of $\mathcal{H}$.

Theorem 6 ([Srebro et al., 2010, Theorem 1]). *Given a non-negative H-smooth loss ℓ bounded by b and a set of data pairs $\widehat{\mathcal{S}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, Define a local loss class $\mathcal{L}(r) = \left\{(\mathbf{x}, \mathbf{y}) \mapsto \ell(h(\mathbf{x}), \mathbf{y}) : h \in \mathcal{H}, \mathcal{L}_{\widehat{\mathcal{S}}}(h) \leq r\right\}$ for some $0 \leq r < \infty$. Then , for all $f \in \mathcal{F}$ simultaneously*

$$\mathfrak{R}_{\widehat{\mathcal{S}}}\left(\mathcal{L}(r)\right) \leq \inf_\alpha \left(\frac{\alpha}{\sqrt{N}} + \int_\alpha^{\sqrt{br}} \sqrt{\frac{\ln \mathcal{N}_\infty(\mathcal{H}, \frac{\epsilon}{\sqrt{12Hr}}, \|\cdot\|)}{N}} d\epsilon\right)$$

where the empirical Rademacher complexity of loss class is defined as

$$\mathfrak{R}_{\widehat{\mathcal{S}}}\left(\mathcal{L}(r)\right) = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{h\in\mathcal{H}, \mathcal{L}_{\widehat{\mathcal{S}}}(h)\leq r}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell(h(\mathbf{x}_i), \mathbf{y}_i)\right|\right]. \qquad (\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \text{unif}\{\pm 1\})$$

The above theorem relates the complexity of loss class to the worst case spectral covering number of function class, in our case, vector valued neural networks. Hence, it remains to find worst case (L_∞) covering number of our encoder class

$$\mathcal{G} \circ \mathcal{H} := \{\mathbf{x} \mapsto \mathbf{W}_{L+1}\sigma\left(\mathbf{W}_L \cdots \sigma(\mathbf{W}_1\mathbf{x})\right) : \|\mathbf{W}_l\| \leq W(l), \|\mathbf{W}_l\|_{2,1} \leq B(l)\ \forall l \in [L+1]\}. \qquad (12)$$

Lemma 6 (Implication of [Bartlett et al., 2017, Theorem 3.3]). *Given a set of data pairs $\widehat{\mathcal{S}} = \{\tilde{\mathbf{z}}_i\}_{i=1}^N$, and hypothesis class defined in (12), then the following statement holds:*

$$\ln \mathcal{N}_\infty(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|) \leq \ln \mathcal{N}(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|) \leq \left(\frac{\left\|\tilde{\mathbf{Z}}\right\|^2 \ln(2m^2)}{\epsilon^2}\right)\left(\prod_{l=1}^{L+1} W^2(l)\right)\left(\sum_{l=1}^{L+1}(\frac{B(l)}{W(l)})^{\frac{2}{3}}\right)^3.$$

where $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1; ...; \tilde{\mathbf{z}}_N]$.

Proof. We define $g \circ h(\mathbf{X}) = [g(h(\mathbf{x}_1)); ...; g(h(\mathbf{x}_N))] \in \mathbb{R}^{N \times d}$. Notice the fact that 2-norm of a row of a matrix, is always less than the spectral norm of the matrix:

$$\max_{i \in [N]} \|g \circ h(\mathbf{x}_i) - g' \circ h'(\mathbf{x}_i)\| \leq \max_{\|\mathbf{a}\| \leq 1} \left\|(g \circ h(\mathbf{X}) - g' \circ h'(\mathbf{X}))^\top \mathbf{a}\right\| = \|g \circ h(\mathbf{X}) - g' \circ h'(\mathbf{X})\|,$$

hence we can have the following fact for covering numbers:

$$\ln \mathcal{N}_\infty(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|) \leq \ln \mathcal{N}(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|). \tag{13}$$

At last plugging the bound for $\ln \mathcal{N}(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|)$ from [Bartlett et al., 2017] concludes the proof. □

Equipped with above results, we are ready to show the local Rademacher complexity of loss class induced by encoder-decoder function class $\mathcal{G} \circ \mathcal{H}$:

Lemma 7. *Given a hypothesis class $\mathcal{H}$, if the logarithm of its L_∞ covering number $\ln \mathcal{N}_\infty(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|)$ is bounded by $\frac{c}{\epsilon^2}$, then the following bound for local Rademcaher complexity holds true:*

$$\mathfrak{R}_{\widehat{\mathcal{S}}}(\mathcal{L}(r)) \leq 10\sqrt{\frac{cHr}{N}} + 10\sqrt{\frac{cHr}{N}}\left(\ln\sqrt{br} - \ln\left(\frac{5}{2}\sqrt{\frac{cHr}{N}}\right)\right).$$

Proof. According to Theorem 6 we have

$$\begin{aligned}
\mathfrak{R}_{\widehat{\mathcal{S}}}(\mathcal{L}(r)) &\leq 4\alpha + 10\int_\alpha^{\sqrt{br}} \sqrt{\frac{\ln \mathcal{N}_\infty(\mathcal{H}, \frac{\epsilon}{\sqrt{12Hr}}, N)}{N}} d\epsilon \\
&\leq 4\alpha + 10\int_\alpha^{\sqrt{br}} \sqrt{\frac{cHr}{N\epsilon^2}} d\epsilon \\
&\leq 4\alpha + 10\sqrt{\frac{cHr}{N}}(\ln\sqrt{br} - \ln(\alpha)).
\end{aligned}$$

Choosing $\alpha = \frac{5}{2}\sqrt{\frac{B^2 cHr}{N}} = \frac{5}{2\sqrt{N}}\sqrt{cHr}$ will minimize above bound, and yields:

$$\mathfrak{R}_{\widehat{\mathcal{S}}}(\mathcal{L}(r)) \leq 10\sqrt{\frac{cHr}{N}} + 10\sqrt{\frac{cHr}{N}}\left(\ln\sqrt{br} - \ln\left(\frac{5}{2}\sqrt{\frac{Hr \cdot c}{N}}\right)\right).$$

□

The following theorem connects local Rademacher complexity to population risk.

Theorem 7. *[Bousquet, 2002, Theorem 6.1] Given a loss class $\mathcal{L}(r)$, let $\phi(r)$ be the function such that*

$$\mathfrak{R}_{\widehat{\mathcal{S}}}(\mathcal{L}(r)) \leq \phi(r).$$

then with probability at least $1 - \exp(-\nu)$,

$$\mathcal{L}_{\mathcal{S}}(h) \leq \mathcal{L}_{\widehat{\mathcal{S}}}(h) + 45r^* + \sqrt{\mathcal{L}_{\mathcal{S}}(h)}\left(\sqrt{8r_n^*} + \sqrt{\frac{4b(\log(1/\nu) + 6\log\log N)}{N}}\right) + 20\frac{b(\nu + 6\log\log N)}{N}$$

where r^* is the largest solution such that $\phi(r) = r$.

C.2.1 Proof of Lemma 5

Proof. First we evoke Lemma 7 with $c = 12\left\|\tilde{\mathbf{Z}}\right\|^2 \ln(2m^2)\left(\prod_{l=1}^{L+1} W^2(l)\right)\left(\sum_{l=1}^{L+1}(\frac{B(l)}{W(l)})^{\frac{2}{3}}\right)^3$

$$\begin{aligned}
\mathfrak{R}_{\widehat{\mathcal{S}}}(\mathcal{L}(r)) &\leq 10\sqrt{\frac{Hr \cdot c}{N}} + 10\sqrt{\frac{cHr}{N}}\left(\ln\sqrt{br} - \ln\left(\frac{5}{2}\sqrt{\frac{Hr \cdot c}{N}}\right)\right) \\
&= 10\sqrt{\frac{Hr \cdot c}{N}} + 10\sqrt{\frac{cHr}{N}}\ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)
\end{aligned}$$

We set $\phi(r) = 10\sqrt{\frac{Hr\cdot c}{N}} \cdot \max\left\{1, \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)\right\}$. Solving the follwoing equation to get r^*

$$\phi(r) = 10\sqrt{\frac{Hr\cdot c}{N}} \cdot \max\left\{1, \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)\right\} = r,$$

$$\iff r^* = 100\frac{H\cdot c}{N} \cdot \max\left\{1, \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)\right\}^2$$

Now, according to Theorem 7, and the fact that

$$A \le B + C\sqrt{A} \Longrightarrow A \le B + C^2 + \sqrt{B}C,$$

we have

$$\mathcal{L}_{\mathcal{U}}(g\circ h) \le \mathcal{L}_{\widehat{\mathcal{U}}}(g\circ h) + 45r^* + \left(\sqrt{8r^*} + \sqrt{\frac{4b(\log(1/\nu) + 6\log\log N)}{N}}\right)^2$$
$$+ 20\frac{b(\nu + 6\log\log N)}{N} + \sqrt{\mathcal{L}_{\widehat{\mathcal{U}}}(g\circ h) + 45r^* + 20\frac{b(\nu + 6\log\log N)}{N}}\left(\sqrt{8r^*} + \sqrt{\frac{4b(\log(1/\nu) + 6\log\log N)}{N}}\right).$$

Plugging r^*, and empirical risk minimizers $\hat{g}, \hat{h}$ will conclude the proof. □

C.3 Proof of Theorem 2

Proof. Recall that in Theorem 1, the generalization bound is given by

$$\mathcal{E}_{\mathcal{T}}(\hat{f}, \hat{h}) \le C_\beta\left(\mathcal{E}_{\mathcal{U}}(\hat{g}, \hat{h}) + \mu\right)^\beta + 4G_\phi\mathfrak{R}_{\widehat{\mathcal{T}}}(\mathcal{F}\circ\hat{h}) + 4B_\phi\sqrt{\frac{\log(1/\nu)}{n}} + 4B_\phi\left\|\mathcal{T} - \mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}} + \min_{f\in\mathcal{F}}\mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}}).$$

Since in the previous subsection we prove the bounded transferrability and generalization of pre-training task, it remains to show the upper bound of representation-induced Rademacher complexity. To this end, we have

$$\begin{aligned}
\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi\circ\mathcal{F}\circ\hat{h}) &= \mathbb{E}_{\boldsymbol{\varepsilon}\in\{\pm1\}^n}\left[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|\le R}\frac{1}{n}\sum\nolimits_{i=1}^n \varepsilon_i\phi(\boldsymbol{\theta}^\top\hat{h}(\mathbf{x}_i), y_i)\right] \\
&\le RG_\phi\mathbb{E}_{\boldsymbol{\varepsilon}\in\{\pm1\}^n}\left[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|\le R}\frac{1}{n}\sum\nolimits_{i=1}^n \varepsilon_i\boldsymbol{\theta}^\top\hat{h}(\mathbf{x}_i)\right] \\
&= \frac{RG_\phi}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum\nolimits_{i=1}^n \varepsilon_i\hat{h}(\mathbf{x}_i)\right\| \\
&\le \frac{RG_\phi}{n}\sqrt{\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum\nolimits_{i=1}^n \varepsilon_i\hat{h}(\mathbf{x}_i)\right\|^2} \\
&= \frac{RG_\phi}{n}\sqrt{\sum\nolimits_{i=1}^n\left\|\hat{h}(\mathbf{x}_i)\right\|^2}
\end{aligned}$$

where at first inequality we apply Ledoux-Talagrand's inequality to peel off Lipschitz loss $\phi(\cdot)$, and at last inequality we use the fact that ε_i are i.i.d. with zero mean, so that the cross terms disappear. For each $\left\|\hat{h}(\mathbf{x}_i)\right\|^2$, we have:

$$\left\|\hat{h}(\mathbf{x}_i)\right\|^2 \le \prod_{l=1}^{L+1} W^2(l)\left\|\mathbf{x}_i\right\|^2,$$

hence we arrive at

$$\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi\circ\mathcal{F}\circ\hat{h}) \le \frac{RG_\phi\sqrt{\prod_{l=1}^{L+1} W^2(l)\sum_{i=1}^n\left\|\mathbf{x}_i\right\|^2}}{n}.$$

Plugging Lemmas 1 and 5 back into Theorem 1 as well as above bound will complete the proof of Theorem 2. □

D Proof of Pre-training with Masked Autoencoder with Tranformer Models

We turn to proving the generalization of pretraining with masked autoencoder (MAE) with tranformer models (Section 5.2).

Recall, in MAE pre-training for vision tasks as an example, we draw a large set of images $\mathbf{Z}_1, ..., \mathbf{Z}_N \in \mathbb{R}^{K\times d}$, and then randomly mask some patches of each image to get $\tilde{\mathbf{Z}}_1, ..., \tilde{\mathbf{Z}}_N \in \mathbb{R}^{K\times d}$. Then an encoder-decoder model is trained by recovering the missing patches (e.g., by utilizing MSE loss $\ell(\hat{\mathbf{Z}}, \mathbf{Z}) = \left\|\hat{\mathbf{Z}} - \mathbf{Z}\right\|_{\mathrm{F}}^2$ as pre-training loss).

We will consider L-layer transformer as the pre-train encoder model, a single self-attention layer transformer as the pre-train decoder model, and a linear projection layer for binary classification as fine-tune model.

Encoder Architecture In a L-layer transformer, given a input $\mathbf{X}$, the lth layer's output is define as:

$$\mathbf{X}^l = \begin{cases} \mathbf{X}, & l = 0 \\ \mathrm{SA}_{\mathbf{W}^l}(\mathbf{X}^{l-1}), & l = [L], \end{cases}$$

where $\mathrm{SA}_{\mathbf{W}^l}(\cdot)$ is the l-layer self attention module given a collection of weight matrices $\mathbf{W}^l = \left(\mathbf{W}_V^l, \mathbf{W}_K^l, \mathbf{W}_Q^l, \mathbf{W}_{\mathrm{FC1}}^l, \mathbf{W}_{\mathrm{FC2}}^l\right) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d_K} \times \mathbb{R}^{d\times d_K} \times \mathbb{R}^{d\times m} \times \mathbb{R}^{m\times d}$ defined as:

$$\begin{aligned} \mathrm{SA}_{\mathbf{W}^l}(\mathbf{X}^{l-1}) &= \alpha_2 \sigma\left(\mathbf{Z}^l \mathbf{W}_{\mathrm{FC1}}^l\right) \mathbf{W}_{\mathrm{FC2}}^l + \mathbf{Z}^l, \\ \mathbf{Z}^l &= \left(\alpha_1 \mathbf{A}^l + \mathbf{X}^{l-1}\right), \\ \mathbf{A}^l &= \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}} \mathbf{X}\mathbf{W}_K^l (\mathbf{X}\mathbf{W}_Q^l)^\top\right) \mathbf{X}\mathbf{W}_V^l, \end{aligned}$$

where α_1, α_2 are some small constant, as used in practice [Noci et al., 2022]. We use the Lth layer's output as the final output of encoder, i.e., $h(\mathbf{X}) = \mathbf{X}^L$.

$$\text{encoder:} \quad h(\mathbf{X}) = \mathbf{X}^L.$$

The hypothesis class of encoder is defined as:

$$\mathcal{H} = \left\{ \begin{array}{l} \mathbf{X} \mapsto \mathrm{SA}_{\mathbf{W}^L}\left(\mathrm{SA}_{\mathbf{W}^{L-1}} ... \mathrm{SA}_{\mathbf{W}^1}(\mathbf{X})\right) : \\ \qquad \left\|\mathbf{W}_{\mathrm{FC1}}^l\right\|, \left\|\mathbf{W}_{\mathrm{FC2}}^l\right\|, \left\|\mathbf{W}_K^l\right\|, \left\|\mathbf{W}_Q^l\right\|, \left\|\mathbf{W}_V^l\right\| \le W(l), \\ \qquad \left\|\mathbf{W}_{\mathrm{FC1}}^l\right\|_{2,1}, \left\|\mathbf{W}_{\mathrm{FC2}}^l\right\|_{2,1}, \left\|\mathbf{W}_K^l\right\|_{2,1}, \left\|\mathbf{W}_Q^l\right\|_{2,1}, \left\|\mathbf{W}_V^l\right\|_{2,1} \le B(l), \forall l \in [L] \end{array} \right\}. \tag{14}$$

Decoder Architecture When encoder finished processing masked sequence, we will send the encoder output $h(\tilde{\mathbf{Z}})$ to decoder. The decoder is a simple linear projection layer:

$$\text{decoder:} \quad g(h(\tilde{\mathbf{Z}})) = h(\tilde{\mathbf{Z}})\mathbf{W}^D,$$

To learn the representation model, we solve the following[1]:

$$\min_{g\in\mathcal{G}, h\in\mathcal{H}} \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) := \frac{1}{2} \sum_{i=1}^{N} \left\|g(h(\tilde{\mathbf{Z}}_i)) - \mathbf{Z}_i\right\|_{\mathrm{F}}^2, \tag{15}$$

to get representation $\hat{h}$.

Then, in the fine-tuning stage for a binary classification tasks with labels $y_i \in \{-1, +1\}$, we consider a linear model parameterized by $\boldsymbol{\theta}$

$$\text{downstream model:} \quad f(h(\mathbf{X})) = \mathbf{1}^\top \hat{h}(\mathbf{X}_i)\boldsymbol{\theta},$$

[1] In some implementation of MAE pre-training, the MSE loss is not computed on full patches, but only the masked patches. It can be adapted by changing our objective to $\frac{1}{2}\sum_{i=1}^{N} \left\|\mathbf{A} \odot (g(h(\tilde{\mathbf{Z}}_i)) - \mathbf{Z}_i)\right\|_{\mathrm{F}}^2$ where $\mathbf{A} \in \mathbb{R}^{K\times d}$ is the indicator matrix with j row to be $\mathbf{1}$ if jth patch is masked, otherwise $\mathbf{0}$. This adaptation will not affect our analysis significantly.

with classification loss $\phi(\cdot,\cdot)$ and optimize:

$$\min_{\|\boldsymbol{\theta}\|_2\leq R}\mathcal{R}_{\widehat{\mathcal{T}}}(\boldsymbol{\theta}\circ\hat{h}(\mathbf{X}))=\frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{1}^\top\hat{h}(\mathbf{X}_i)\boldsymbol{\theta},y_i),$$

to get $\hat{f}$ (or $\hat{\boldsymbol{\theta}}$ in this setting), the aggregated patch over all patches is used for linear projection in classification task.

Roadmap. We will provide proof of Theorem 3 in the following subsections. The roadmap is that in Appendix D.1 we first show the MAE pre-training admits bounded representation transferrability to downstream task (Lemma 2), and then in Appendix D.2 we prove the generalization of MAE pre-training task (Lemma 3). The heart of the proof in this part is to derive worst case covering number of transformer class. Finally in Appendix D.3 we conclude the proof for Theorem 3 by showing that the representation-induced Rademacher complexity is bounded.

D.1 Proof of Task Transferability of MAE

Similar to proof of DAE transferability, we define the following quantity:

$$\Delta_{\mathcal{U}}^{ft}(\hat{h},h_{\mathcal{U}}^*)=\min_{\|\boldsymbol{\theta}\|\leq R}\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}[\phi(\boldsymbol{\theta}^\top(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top)]-\min_{\|\tilde{\boldsymbol{\theta}}\|\leq R}\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}[\phi(\tilde{\boldsymbol{\theta}}^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top)],$$

$$\Delta_{\mathcal{U}}^{pt}(\hat{h},h_{\mathcal{U}}^*)=\min_{\mathbf{W}^D\in\mathbb{R}}\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}\left\|\hat{h}(\tilde{\mathbf{Z}})\mathbf{W}^D-\mathbf{Z}\right\|_{\mathrm{F}}^2-\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}\left\|h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\mathbf{W}^{D*}-\mathbf{Z}\right\|_{\mathrm{F}}^2$$

where $\mathbf{1}=[1,1,1,...]\in\mathbb{R}^K$.

Upper bounding $\Delta_{\mathcal{U}}^{ft}(\hat{h},h_{\mathcal{U}}^*)$ We examine $\Delta_{\mathcal{U}}^{ft}(\hat{h},h_{\mathcal{U}}^*)$ first. Similar to DAE proof, We define the optimal head for classification task on distribution $\mathcal{U}_{\mathcal{X}}$ under representation $h_{\mathcal{U}}^*$ as $\tilde{\boldsymbol{\theta}}^*=\arg\min_{\|\tilde{\boldsymbol{\theta}}\|\leq R}\mathbb{E}_{\tilde{\mathbf{Z}}\sim\mathcal{U}_{\mathcal{X}}}[\phi(\tilde{\boldsymbol{\theta}}^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})))]$.

$$\begin{aligned}
&\Delta_{\mathcal{U}}^{ft}(\hat{h},h_{\mathcal{U}}^*)=\min_{\|\boldsymbol{\theta}\|\leq R}\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}[\phi(\boldsymbol{\theta}^\top(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top)]-\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}[\phi(\tilde{\boldsymbol{\theta}}^{\star\top}(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top)]\\
&\leq\min_{\|\boldsymbol{\theta}\|\leq R}G_\phi\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}|(\boldsymbol{\theta}^\top(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top)-(\tilde{\boldsymbol{\theta}}^{\star\top}(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top)|\\
&\leq\min_{\|\boldsymbol{\theta}\|\leq R}G_\phi\sqrt{\mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}}\left(\boldsymbol{\theta}^\top(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top-\tilde{\boldsymbol{\theta}}^{\star\top}(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top\right)^2}\\
&=\min_{\|\boldsymbol{\theta}\|\leq R}\\
&\quad G_\phi\sqrt{\boldsymbol{\theta}^\top\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))\right]\boldsymbol{\theta}-2\boldsymbol{\theta}^\top\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))\right]\tilde{\boldsymbol{\theta}}^\top+\tilde{\boldsymbol{\theta}}^{\star\top}\mathbb{E}\left[(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top\hat{h}_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))\right]\tilde{\boldsymbol{\theta}}}
\end{aligned}$$

Since $\sqrt{f(x)}$ and $f(x)$ attain the minimum at the same point, we examine the minimum of above statement over $\boldsymbol{\theta}$ without square root. Under unconstrained setting, the minimum of above statement is $\tilde{\boldsymbol{\theta}}^{\star\top}\Lambda\left(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}),\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\tilde{\boldsymbol{\theta}}^*$ when $\boldsymbol{\theta}^*=\left(\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\mathbf{X}))^\top(\mathbf{1}^\top\hat{h}(\mathbf{X}))\right]\right)^\dagger\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\mathbf{X}))^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\mathbf{X}))\right]\tilde{\boldsymbol{\theta}}^*$. Hence we have

$$\begin{aligned}
\Delta_{\mathcal{U}}^{ft}(\hat{h},h_{\mathcal{U}}^*)&\leq\sqrt{\tilde{\boldsymbol{\theta}}^{\star\top}\Lambda\left(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}),\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\tilde{\boldsymbol{\theta}}^*}=\sqrt{\mathsf{tr}(\Lambda\left(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}),\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\tilde{\boldsymbol{\theta}}^*\tilde{\boldsymbol{\theta}}^{\star\top})}\\
&\leq\sqrt{d\sigma_{\max}(\Lambda\left(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}),\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right))\sigma_{\max}(\tilde{\boldsymbol{\theta}}^*\tilde{\boldsymbol{\theta}}^{\star\top})},
\end{aligned}$$

where

$$\begin{aligned}
&\Lambda\left(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}),\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\\
&=\mathbb{E}\left[(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))\right]-\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))\right]\left(\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))\right]\right)^\dagger\mathbb{E}\left[(\mathbf{1}^\top\hat{h}(\tilde{\mathbf{Z}}))^\top(\mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))\right].
\end{aligned}$$

At last, by choosing a properly large R, we can guarantee the optimum can be attained.

Lower bounding $\Delta_{\mathcal{U}}^{pt}(\hat{h}, h_{\mathcal{U}}^*)$ Similar to CE proof, we have:

$$
\begin{aligned}
&\Delta_{\mathcal{U}}^{pt}(\hat{h}, h_{\mathcal{U}}^*) \\
&= \min_{\mathbf{W}^D \in \mathbb{R}^{d\times d}} \mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}} \left\|\hat{h}(\widetilde{\mathbf{Z}})\mathbf{W}^D - \mathbf{Z}\right\|_{\mathrm{F}}^2 - \mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}} \left\|h_{\mathcal{U}}^*(\widetilde{\mathbf{Z}})\mathbf{W}^{D^*} - \mathbf{Z}\right\|_{\mathrm{F}}^2 \\
&= \min_{\mathbf{W}^D \in \mathbb{R}^{d\times d}} \mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}} \left\|\hat{h}(\widetilde{\mathbf{Z}})\mathbf{W}^D - h_{\mathcal{U}}^*(\widetilde{\mathbf{Z}})\mathbf{W}^{D^*}\right\|_{\mathrm{F}}^2 \\
&\geq \sum_{i=1}^{d} \min_{\mathbf{w}_r \in \mathbb{R}^d} \mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}} \left\|\hat{h}(\widetilde{\mathbf{Z}})\mathbf{w}_r - h_{\mathcal{U}}^*(\widetilde{\mathbf{Z}})\mathbf{w}_r^*\right\|^2 \\
&= \sum_{i=1}^{d} \min_{\mathbf{w}_r \in \mathbb{R}^d} \mathbb{E}_{(\tilde{\mathbf{Z}},\mathbf{Z})\sim\mathcal{U}} \left(\mathbf{w}_r^\top \hat{h}(\widetilde{\mathbf{Z}})^\top \hat{h}(\widetilde{\mathbf{Z}})\mathbf{w}_r - 2\mathbf{w}_r^\top \hat{h}(\widetilde{\mathbf{Z}})^\top h_{\mathcal{U}}^*(\widetilde{\mathbf{Z}})\mathbf{w}_r^* + \mathbf{w}_r^{*\top} h_{\mathcal{U}}^*(\widetilde{\mathbf{Z}})^\top h_{\mathcal{U}}^*(\widetilde{\mathbf{Z}})\mathbf{w}_r^*\right)
\end{aligned}
$$

where the second step is due to our realizability Assumption 1, $\mathbf{w}_r \in \mathbb{R}^d$ represents rth colum of $\mathbf{W}^D$ and so is $\mathbf{w}_r^* \in \mathbb{R}^d$ represents rth colum of $\mathbf{W}^{D^*}$.

Similar to Context Encoder proof, we define Schur complement as:

$$
\Lambda\left(\hat{h}(\tilde{\mathbf{Z}}), h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right) = \mathbb{E}\left[(h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right] - \mathbb{E}\left[(h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top (\hat{h}(\tilde{\mathbf{Z}}))\right] \left(\mathbb{E}\left[(h_{\mathcal{U}}^*(\tilde{\mathbf{Z}}))^\top h_{\mathcal{U}}^{*\top}(\tilde{\mathbf{Z}})\right]\right)^\dagger \mathbb{E}\left[(\hat{h}(\tilde{\mathbf{Z}}))^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right].
$$

By computing the closed form solution of quadratic form we arrived at:

$$
\begin{aligned}
\Delta_{\mathcal{U}}^{pt}(\hat{h}, h_{\mathcal{U}}^*) &\geq \sum_{r=1}^{d} \mathbf{w}_r^{*\top} \Lambda\left(\hat{h}(\tilde{\mathbf{Z}}), h^*(\tilde{\mathbf{Z}})\right) \mathbf{w}_r^* \\
&\geq \operatorname{tr}\left(\Lambda\left(\hat{h}(\tilde{\mathbf{Z}}), h^*(\tilde{\mathbf{Z}})\right) \sum_{j=1}^{d} \mathbf{w}_j^* \mathbf{w}_j^{*\top}\right) \\
&\geq \sigma_{\max}\left(\Lambda\left(\hat{h}(\tilde{\mathbf{Z}}), h^*(\tilde{\mathbf{Z}})\right)\right) \sigma_{\min}\left(\sum_{j=1}^{d} \mathbf{w}_j^* \mathbf{w}_j^{*\top}\right).
\end{aligned}
$$

Recall that

$$
\Delta_{\mathcal{U}}^{ft}(\hat{h}, h_{\mathcal{U}}^*) \leq \sqrt{d\sigma_{\max}\left(\Lambda\left(\mathbf{1}^\top \hat{h}(\tilde{\mathbf{Z}}), \mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\right) \sigma_{\max}(\tilde{\boldsymbol{\theta}}^* \tilde{\boldsymbol{\theta}}^{\star\top})}.
$$

Hence, we can conclude that

$$
\frac{\Delta_{\mathcal{U}}^{ft}(\hat{h}, h_{\mathcal{U}}^*)}{\left(\Delta_{\mathcal{U}}^{pt}(\hat{h}, h_{\mathcal{U}}^*)\right)^{1/2}} \leq O\left(\frac{\sqrt{\sigma_{\max}\left(\Lambda\left(\mathbf{1}^\top \hat{h}(\tilde{\mathbf{Z}}), \mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\right)}}{\sqrt{\sigma_{\max}\left(\Lambda\left(\hat{h}(\tilde{\mathbf{Z}}), h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\right)}} \sqrt{\frac{d\sigma_{\max}(\tilde{\boldsymbol{\theta}}^* \tilde{\boldsymbol{\theta}}^{\star\top})}{\sigma_{\min}\left(\sum_{j=1}^{d} \mathbf{w}_j^* \mathbf{w}_j^{*\top}\right)}}\right),
$$

which indicates that MAE pre-training admits an

$$
\left(\Omega\left(\frac{\sqrt{\sigma_{\max}\left(\Lambda\left(\mathbf{1}^\top \hat{h}(\tilde{\mathbf{Z}}), \mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\right)}}{\sqrt{\sigma_{\max}\left(\Lambda\left(\hat{h}(\tilde{\mathbf{Z}}), h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})\right)\right)}} \sqrt{\frac{d\sigma_{\max}(\tilde{\boldsymbol{\theta}}^* \tilde{\boldsymbol{\theta}}^{\star\top})}{\sigma_{\min}\left(\sum_{j=1}^{d} \mathbf{w}_j^* \mathbf{w}_j^{*\top}\right)}}\right), \frac{1}{2}\right)
$$

representation transferrability to binary classification task. Notice that the transfer constant C_β mainly depends on the Schur complement of $\hat{h}(\tilde{\mathbf{Z}}), h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})$, and $\mathbf{1}^\top \hat{h}(\tilde{\mathbf{Z}}), \mathbf{1}^\top h_{\mathcal{U}}^*(\tilde{\mathbf{Z}})$. In the main paper Lemma 2 we omit this constant dependency.

D.2 Proof of Generalization of MAE Pre-training Task

In this section we are going to derive generalization bound of the masking pre-training with Transformer. In pursuit of optimal generalization bound of pretraining task, i.e., regression with deep transformer, we again need to employ the framework we introduced in CE analysis (Appendix C.2). Hence, we need to upper bound the worst case L_2 covering number of deep transformer class. The following result establishes the worst case spectral covering number of L-layer self-attention transformer defined in 14.

Lemma 8 (Covering number of transformer class)**.** *Let $\mathbf{X}_{[N]} = [\mathbf{X}_1; ...\mathbf{X}_N] \in \mathbb{R}^{NK\times d}$ denotes the concatenated data matrix. Then the worst case covering number of L-layer transformer class $\mathcal{H}$ defined in 14 is bounded as follows:*

$$\ln \mathcal{N}_\infty(\mathcal{H}(\mathcal{S}), \epsilon, \|\cdot\|) \leq O\left(s_L^2 \left\|\mathbf{X}_{[N]}\right\|^2 \sum_{l=1}^{L} \frac{\rho_l}{\epsilon^2}\right),$$

where

$$\begin{aligned}
s_l &:= \prod_{j=1}^{l} \left(\alpha_2 W^2(j) + 1\right)\left(W^2(j)\alpha_1 K + 1\right), \\
\rho_l &:= O\left(\alpha_1^2(\alpha_2 W^2(l)+1)^2 B^2(l)\ln(2d^2)\left(K^2 + \frac{\alpha_1 W^2(l)\left(s_{l-1}\|\mathbf{X}_*\|\right)^2}{d_K}\right)\right) \\
&\quad + O\left(\alpha_2^2 W^2(l) B^2(l)(W^2(l) + \alpha_1^2 K^2 W^2(l))\ln(2dm)\right), \\
\|\mathbf{X}_*\| &:= \max_{i\in[N]} \|\mathbf{X}_i\|.
\end{aligned}$$

Roughly speaking, ρ_l is the price for covering the parameter of lth self-attention layer, and extending the cover to the whole model yields the sum over l. Notice that to ensure a L_∞ cover, it suffices to ensure that $\sum_{i=1}^N \|h(\mathbf{X}_i) - h_\epsilon(\mathbf{X}_i)\|^2 \leq \epsilon^2$. However, if we trivially cover each individual loss $\|h(\mathbf{X}_i) - h_\epsilon(\mathbf{X}_i)\|^2$ with ϵ^2/N radius, the final covering number will be N times larger, which make the later generalization bound vacuous, i.e., greater than 1. To avoid this N factor, we directly consider the cover over concatenated data matrix $\mathbf{X}_{[N]}$, and consider the covering $\left\|\hat{h}(\mathbf{X}_{[N]}) - h(\mathbf{X}_{[N]})\right\|^2 \leq \epsilon^2$. Using the fact that matrix covering bound is independent of dimension of $\mathbf{X}_{[N]}$, but only depends the spectral norm of $\mathbf{X}_{[N]}$, the final covering number will only have logrithmic dependency on N.

To prove Lemma 8, first we introduce the following matrix covering number bound from [Bartlett et al., 2017].

Lemma 9. *[Bartlett et al., 2017, Lemma 3.2] Let conjugate exponents (p, q) and (r, s) be given with $p \leq 2$, as well as positive reals (a, b, ϵ) and positive integer m. Let matrix $\mathbf{X} \in \mathbb{R}^{NK\times d}$ be given with $\|\mathbf{X}\|_p \leq b$. Then*

$$\ln \mathcal{N}\left(\left\{\mathbf{XW} : \mathbf{W} \in \mathbb{R}^{d\times m}, \|\mathbf{W}\|_{q,s} \leq a\right\}, \epsilon, \|\cdot\|_2\right) \leq \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \right\rceil \ln(2dm).$$

Lemma 10 (Covering number of attention matrix)**.** *Given a set of data $\mathcal{S} = \{\mathbf{X}_1, ..., \mathbf{X}_N\}$ and the attention matrix class:*

$$\mathcal{H}_S(\mathcal{S}) = \left\{ \mathbf{S} = \begin{bmatrix} \mathbf{S}_1, \mathbf{0}, ..., \mathbf{0}, \\ \ddots \\ \mathbf{0}, ..., \mathbf{0}, \mathbf{S}_N \end{bmatrix} : \mathbf{S}_i = \text{softmax}\left(\frac{1}{\sqrt{d_K}} \mathbf{X}_i \mathbf{W}_K (\mathbf{X}_i \mathbf{W}_Q)^\top\right) : \|\mathbf{W}_K\|, \|\mathbf{W}_Q\| \leq W, \right.$$
$$\left. \|\mathbf{W}_K\|_{2,1}, \|\mathbf{W}_Q\|_{2,1} \leq B, \right\}$$

the following covering number bound holds true:

$$\ln \mathcal{N}(\mathcal{H}_S(\mathcal{S}), \epsilon, \|\cdot\|) \leq O\left(\frac{KW^2B^2\|\mathbf{X}_*\|^4}{d_K\epsilon^2}\ln(2d^2)\right).$$

Proof. we define set $\mathcal{K} = \left\{\mathbf{X}\mathbf{W}_K : \|\mathbf{W}_K\| \le W, \|\mathbf{W}_K\|_{2,1} \le B\right\}$, $\mathcal{Q} = \left\{\mathbf{X}\mathbf{W}_Q : \|\mathbf{W}_Q\| \le W, \|\mathbf{W}_Q\|_{2,1} \le B\right\}$. We define ϵ_K cover of $\mathcal{K}$ as $\mathcal{C}_K$, and ϵ_Q cover of $\mathcal{Q}$ as $\mathcal{C}_Q$. We construct the following set:

$$\mathcal{C}_S = \left\{ \mathbf{S} = \begin{bmatrix} \mathbf{S}_1, \mathbf{0}, ..., \mathbf{0}, \\ \ddots \\ \mathbf{0}, ..., \mathbf{0}, \mathbf{S}_N \end{bmatrix} : \mathbf{S}_i = \text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}_i\mathbf{W}_K(\mathbf{X}_i\mathbf{W}_Q)^\top\right) : \mathbf{W}_K \in \mathcal{C}_K, \mathbf{W}_Q \in \mathcal{C}_Q \right\}$$

Next we will show that $\mathcal{C}_S$ is a cover of $\mathcal{H}_S(\mathcal{S})$ with some radius. For any $\mathbf{S}_{[N]} \in \mathcal{H}_S$, we can find $\hat{\mathbf{S}}_{[N]} \in \mathcal{C}_S$ such that:

$$\begin{aligned}
\left\|\mathbf{S}_{[N]} - \hat{\mathbf{S}}_{[N]}\right\| &\le \max_{i\in[N]} \left\|\mathbf{S}_i - \hat{\mathbf{S}}_i\right\| \\
&= \max_{i\in[N]} \left\|\text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}_i\mathbf{W}_K(\mathbf{X}_i\mathbf{W}_Q)^\top\right) - \text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}_i\hat{\mathbf{W}}_K(\mathbf{X}_i\hat{\mathbf{W}}_Q)^\top\right)\right\| \\
&\le \frac{\sqrt{K}}{\sqrt{d_K}} \max_{i\in[N]} \left\|\mathbf{X}_i\mathbf{W}_K(\mathbf{X}_i\mathbf{W}_Q)^\top - \mathbf{X}_i\hat{\mathbf{W}}_K(\mathbf{X}_i\hat{\mathbf{W}}_Q)^\top\right\| \\
&\le \frac{\sqrt{K}}{\sqrt{d_K}} \max_{i\in[N]} \left\|(\mathbf{X}_i\mathbf{W}_K - \mathbf{X}_i\hat{\mathbf{W}}_K)(\mathbf{X}_i\mathbf{W}_Q)^\top\right\| \\
&\quad + \frac{\sqrt{K}}{\sqrt{d_K}} \max_{i\in[N]} \left\|\mathbf{X}_i\hat{\mathbf{W}}_K(\mathbf{X}_i\mathbf{W}_Q)^\top - \mathbf{X}_i\hat{\mathbf{W}}_K(\mathbf{X}_i\hat{\mathbf{W}}_Q)^\top\right\| \\
&\le \frac{W\sqrt{K}}{\sqrt{d_K}} (\epsilon_K + \epsilon_Q) \max_{i\in[N]} \|\mathbf{X}_i\|,
\end{aligned}$$

where the first inequality is due to the property of block diagonal matrices. We define $\|\mathbf{X}_*\| = \max_{i\in[N]} \|\mathbf{X}_i\|$. To ensure above bound is less than ϵ, we choose $\epsilon_K = \epsilon_Q = \frac{\sqrt{d_K}}{2W\sqrt{K}\|\mathbf{X}_*\|}\epsilon$. According to Lemma 9, we know:

$$\ln|\mathcal{C}_S| \le \ln|\mathcal{C}_K| + \ln|\mathcal{C}_Q| \le O\left(\frac{KW^2B^2\|\mathbf{X}_*\|^4}{d_K\epsilon^2}\ln(2d^2)\right).$$

□

Proposition 3 (Covering number of single self-attention layer)**.** *Consider the following function class of self-attention module:*

$$\mathcal{H}_{SA} := \left\{ \begin{array}{r} \mathbf{X} \mapsto \sigma\left(\mathbf{Z}\mathbf{W}_{\text{FC1}}\right)\mathbf{W}_{\text{FC2}} + \mathbf{Z} : \mathbf{Z} = (\mathbf{A} + \mathbf{X}), \mathbf{A} = \text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}\mathbf{W}_K(\mathbf{X}\mathbf{W}_Q)^\top\right)\mathbf{X}\mathbf{W}_V \\ \|\mathbf{W}_{\text{FC1}}\|, \|\mathbf{W}_{\text{FC2}}\|, \|\mathbf{W}_K\|, \|\mathbf{W}_Q\|, \|\mathbf{W}_V\| \le W, \\ \|\mathbf{W}_{\text{FC1}}\|_{2,1}, \|\mathbf{W}_{\text{FC2}}\|_{2,1}, \|\mathbf{W}_K\|_{2,1}, \|\mathbf{W}_Q\|_{2,1}, \|\mathbf{W}_V\|_{2,1} \le B, \end{array} \right\}$$

then the following bound holds for its covering number:

$$\begin{aligned}
\ln\mathcal{N}(\mathcal{H}_{SA}(\mathcal{S}), \epsilon, \|\cdot\|) &\le O\left(\frac{(\alpha_1\alpha_2W^2 + \alpha_1)^2B^2\left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon^2}\ln(2d^2)\right)\left(K^2 + \frac{\alpha_1W^2\|\mathbf{X}_*\|^2}{d_K}\right) \\
&\quad + O\left(\frac{\alpha_2^2W^2B^2(W^2\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2K^2W^2\left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon^2}\ln(2dm)\right).
\end{aligned}$$

Proof. Recall that $\mathbf{X}_{[N]} \in \mathbb{R}^{NK\times d}$ is the concatenated data matrix, and we shall use $h(\mathbf{X}_{[N]}) \in \mathbb{R}^{NK\times d}$ to denote the concatenated encoder output, i.e., $h(\mathbf{X}_{[N]}) = [h(\mathbf{X}_1); ..., h(\mathbf{X}_N)]$. Our goal is to find the cardinality of a cover such that for any $h \in \mathcal{H}$ we can find a $h_\epsilon \in \mathcal{C}_{\text{SA}}$ such that

$$\left\|h(\mathbf{X}_{[N]}) - h_\epsilon(\mathbf{X}_{[N]})\right\| \le \epsilon.$$

I: Covering number of input layer by value matrix Let $\mathcal{C}_V$ to be ϵ_V cover of set $\mathcal{H}_V(\mathcal{S}) = \left\{\mathbf{X}_{[N]}\mathbf{W}_V : \|\mathbf{W}_V\| \leq W, \|\mathbf{W}_V\|_{2,1} \leq B\right\}$, then evoking Lemma 9 we have:

$$\ln \mathcal{N}(\mathcal{H}_V, \epsilon_V, \|\cdot\|) \leq O\left(\frac{B^2 \left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon_V^2} \ln(2dm)\right).$$

II: Covering number of Attention layer Next, consider the set of attention matrix

$$\mathcal{H}_S(\mathcal{S}) = \left\{ \mathbf{S} = \begin{bmatrix} \mathbf{S}_1, \mathbf{0}, ..., \mathbf{0}, \\ \ddots \\ \mathbf{0}, ..., \mathbf{0}, \mathbf{S}_N \end{bmatrix} : \mathbf{S}_i = \text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}_i\mathbf{W}_K(\mathbf{X}_i\mathbf{W}_Q)^\top\right) : \|\mathbf{W}_K\|, \|\mathbf{W}_Q\| \leq W, \right.$$
$$\left. \|\mathbf{W}_K\|_{2,1}, \|\mathbf{W}_Q\|_{2,1} \leq B, \right\}$$

From Lemma 10 we know its covering number can be bounded as:

$$\ln \mathcal{N}\left(\mathcal{H}_S(\mathcal{S}), \epsilon, \|\cdot\|\right) \leq \ln \mathcal{N}\left(\mathcal{H}_{\tilde{S}}, \epsilon_S, \|\cdot\|\right) \leq O\left(\frac{KW^2B^2 \left\|\mathbf{X}_*\right\|^4}{d_K\epsilon_S^2} \ln(d^2)\right).$$

Now we can proceed to bounding the covering number of following set:

$$\mathcal{H}_A(\mathcal{S}) = \left\{ \alpha_1 \text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{X}_{[N]}\mathbf{W}_K(\mathbf{X}_{[N]}\mathbf{W}_Q)^\top\right)\mathbf{X}_{[N]}\mathbf{W}_V : \|\mathbf{W}_K\|, \|\mathbf{W}_Q\|, \|\mathbf{W}_V\| \leq W, \right.$$
$$\left. \|\mathbf{W}_K\|_{2,1}, \|\mathbf{W}_Q\|_{2,1}, \|\mathbf{W}_V\|_{2,1} \leq B, \right\}$$

For every element $\hat{\mathbf{V}}_{[N]} \in \mathcal{C}_V$, we construct the set $\alpha_1\mathcal{H}_S(\mathcal{S}) \circ \hat{\mathbf{V}}_{[N]} := \left\{\alpha_1\mathbf{S}_{[N]}\hat{\mathbf{V}}_{[N]} : \mathbf{S}_{[N]} \in \mathcal{H}_S(\mathcal{S})\right\}$. Then we define ϵ_A-covering of $\mathcal{H}_S \circ \hat{\mathbf{V}}$ as $\mathcal{C}(\mathcal{H}_S \circ \hat{\mathbf{V}}, \epsilon_A, \|\cdot\|)$. To construct $\mathcal{H}_S \circ \hat{\mathbf{V}}$ as $\mathcal{C}(\mathcal{H}_S \circ \hat{\mathbf{V}}, \epsilon_A, \|\cdot\|)$, we consider $\mathcal{C}_S$. For any $\mathbf{S}_{[N]}\hat{\mathbf{V}}_{[N]} \in \mathcal{H}_S(\mathcal{S}) \circ \hat{\mathbf{V}}_{[N]}$, we can find $\hat{\mathbf{S}}_{[N]} \in \mathcal{C}_S$, such that

$$\begin{aligned}\left\|\alpha_1\mathbf{S}_{[N]}\hat{\mathbf{V}}_{[N]} - \alpha_1\hat{\mathbf{S}}_{[N]}\hat{\mathbf{V}}_{[N]}\right\| &\leq \alpha_1 \left\|\mathbf{S}_{[N]} - \hat{\mathbf{S}}_{[N]}\right\| \left\|\hat{\mathbf{V}}_{[N]}\right\| \\ &\leq \alpha_1\epsilon_S \left\|\mathbf{X}_{[N]}\right\| W.\end{aligned}$$

Setting $\epsilon_S = \frac{\epsilon_A}{\alpha_1\left\|\mathbf{X}_{[N]}\right\|W}$ we can conclude that $\mathcal{C}(\mathcal{H}_S \circ \hat{\mathbf{V}}, \epsilon_A, \|\cdot\|)$ actually ϵ_A covers $\mathcal{H}_S \circ \hat{\mathbf{V}}$ and the following fact holds for the covering number

$$\ln\left|\mathcal{C}(\mathcal{H}_S(\mathcal{S}) \circ \hat{\mathbf{V}}_{[N]}, \epsilon_A, \|\cdot\|)\right| \leq \sup_{\hat{\mathbf{V}}_{[N]} \in \mathcal{C}_V} \ln \mathcal{N}(\mathcal{H}_S(\mathcal{S}) \circ \hat{\mathbf{V}}_{[N]}, \epsilon_A, \|\cdot\|) \leq O\left(\frac{\alpha_1^2KB^2W^4 \left\|\mathbf{X}_*\right\|^4 \left\|\mathbf{X}_{[N]}\right\|^2}{d_K\epsilon_A^2} \ln(2d^2)\right).$$

Then we construct a cover $\mathcal{C}_A$ for $\mathcal{H}_A$ by:

$$\mathcal{C}_A = \bigcup_{\hat{\mathbf{V}}_{[N]} \in \mathcal{C}_V} \mathcal{C}(\alpha_1\mathcal{H}_S(\mathcal{S}) \circ \hat{\mathbf{V}}_{[N]})$$

It is not hard to verify the cardinality of this cover:

$$\begin{aligned}\ln|\mathcal{C}_A| &\leq \ln|\mathcal{C}_V| + \sup_{\hat{\mathbf{V}}_{[N]} \in \mathcal{C}_V} \ln\left|\mathcal{C}(\alpha_1\mathcal{H}_S(\mathcal{S}) \circ \hat{\mathbf{V}}_{[N]}, \epsilon_A, \|\cdot\|)\right| \\ &\leq O\left(\frac{B^2 \left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon_V^2} \ln(2dm)\right) + O\left(\frac{\alpha_1^2KB^2W^4 \left\|\mathbf{X}_*\right\|^4 \left\|\mathbf{X}_{[N]}\right\|^2}{d_K\epsilon_A^2} \ln(2d^2)\right).\end{aligned}$$

III: Covering number of fully-connected layer 1 By similar reasoning, we can show that the covering number of

$$\mathcal{H}_{\text{FC1}}(\mathcal{S}) = \left\{\mathbf{Z}_{[N]}\mathbf{W}_{\text{FC1}} : \mathbf{Z}_{[N]} = \alpha_1\mathbf{A}_{[N]} + \mathbf{X}_{[N]}, \mathbf{A}_{[N]} \in \mathcal{H}_A(\mathcal{S}), \|\mathbf{W}_{\text{FC1}}\| \le W, \|\mathbf{W}_{\text{FC1}}\|_{2,1} \le B\right\}$$

For every element $\hat{\mathbf{A}}_{[N]} \in \mathcal{C}_A$, we define set

$$\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}} = \left\{(\alpha_1\hat{\mathbf{A}}_{[N]} + \mathbf{X}_{[N]})\mathbf{W}_{\text{FC1}}, \|\mathbf{W}_{\text{FC1}}\| \le W, \|\mathbf{W}_{\text{FC1}}\|_{2,1} \le B\right\}$$

We denote ϵ_{FC1}-cover of $\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}$ as $\mathcal{C}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{FC_1}, \epsilon_{\text{FC1}}, \|\cdot\|)$, and the covering number of $\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}$ is bounded by:

$$\begin{aligned}
\ln|\mathcal{C}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}, \epsilon_{\text{FC1}}, \|\cdot\|)| &\le \sup_{\hat{\mathbf{A}}_{[N]} \in \mathcal{C}_A} \ln\mathcal{N}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}, \epsilon_{\text{FC1}}, \|\cdot\|) \\
&= \sup_{\hat{\mathbf{S}}_{[N]} \in \mathcal{C}_S} O\left(\frac{B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2W^2\left\|\mathbf{X}_{[N]}\right\|^2\left\|\hat{\mathbf{S}}_{[N]}\right\|^2)}{\epsilon_{\text{FC1}}^2}\ln(dm)\right) \\
&\le O\left(\frac{B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2K^2W^2\left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon_{\text{FC1}}^2}\ln(dm)\right)
\end{aligned}$$

Now, we construct the ϵ_{FC1}-cover of $\mathcal{H}_{\text{FC1}}(\mathcal{S})$ as

$$\mathcal{C}_{\text{FC1}} = \bigcup_{\hat{\mathbf{A}}_{[N]} \in \mathcal{C}_A} \mathcal{C}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}, \epsilon_{\text{FC1}}, \|\cdot\|)$$

And the covering number is bounded:

$$\begin{aligned}
\ln|\mathcal{C}_{\text{FC1}}| &\le \ln|\mathcal{C}_A| + \sup_{\hat{\mathbf{A}}_{[N]} \in \mathcal{C}_A} \ln|\mathcal{C}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}, \epsilon_{\text{FC1}}, \|\cdot\|)| \\
&\le O\left(\frac{B^2\left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon_V^2}\ln(2dm)\right) + O\left(\frac{\alpha_1^2KB^2W^4\left\|\mathbf{X}_*\right\|^4\left\|\mathbf{X}_{[N]}\right\|^2}{d_K\epsilon_A^2}\ln(2d^2)\right) \\
&\quad + O\left(\frac{B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2K^2W^2\left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon_{\text{FC1}}^2}\ln(dm)\right).
\end{aligned}$$

IV: Covering number of fully-connected layer 2 The analysis this part is almost identical to **III**. We try to find the covering number of the set $\mathcal{H}_{\text{SA}}$. For every element $\hat{\mathbf{F}}_{[N]} \in \mathcal{C}_{\text{FC1}}$ and $\hat{\mathbf{A}}_{[N]} \in \mathcal{C}_A$, define the set

$$\alpha_2\hat{\mathbf{F}}_{[N]} \circ \mathcal{W}_{\text{FC2}} + \hat{\mathbf{Z}}_{[N]} = \left\{\alpha_2\sigma(\hat{\mathbf{F}}_{[N]})\mathbf{W}_{\text{FC2}} + \hat{\mathbf{Z}}_{[N]} : \hat{\mathbf{Z}}_{[N]} = \alpha_1\hat{\mathbf{A}}_{[N]} + \mathbf{X}_{[N]}, \|\mathbf{W}_{\text{FC2}}\| \le W, \|\mathbf{W}_{\text{FC2}}\|_{2,1} \le B\right\}$$

.

We denote ϵ_{FC2}-cover of $\alpha_2\hat{\mathbf{F}}_{[N]} \circ \mathcal{W}_{\text{FC2}} + \hat{\mathbf{Z}}$ as $\mathcal{C}(\alpha_2\hat{\mathbf{F}}_{[N]} \circ \mathcal{W}_{\text{FC2}} + \hat{\mathbf{Z}}, \epsilon_{\text{FC2}}, \|\cdot\|)$, and the cardinality of this set is bounded by:

$$\begin{aligned}
\ln|\mathcal{C}(\alpha_2\hat{\mathbf{F}}_{[N]} \circ \mathcal{W}_{\text{FC2}} + \hat{\mathbf{Z}}_{[N]}, \epsilon_{\text{FC2}}, \|\cdot\|)| &\le \sup_{\hat{\mathbf{F}}_{[N]} \in \mathcal{C}_{\text{FC1}}, \hat{\mathbf{A}} \in \mathcal{C}_A} \ln\mathcal{N}(\hat{\mathbf{F}}_{[N]} \circ \mathcal{W}_{\text{FC2}} + \hat{\mathbf{Z}}_{[N]}, \epsilon_{\text{FC2}}, \|\cdot\|) \\
&= O\left(\frac{\alpha_2^2B^2\left(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2K^2W^2\left\|\mathbf{X}_{[N]}\right\|^2\right)W^2}{\epsilon_{\text{FC2}}^2}\ln(2dm)\right)
\end{aligned}$$

Now, we construct the ϵ_{FC2}-cover of $\mathcal{H}_{\text{FC2}}$ as

$$\mathcal{C}_{\text{FC2}} = \bigcup_{\hat{\mathbf{F}}_{[N]} \in \mathcal{H}_{FC1}, \hat{\mathbf{A}} \in \mathcal{H}_A} \mathcal{C}(\hat{\mathbf{F}}_{[N]} \circ \mathcal{W}_{\text{FC2}} + \hat{\mathbf{Z}}_{[N]}, \epsilon_{\text{FC2}}, \|\cdot\|)$$

And the covering number is bounded:

$$\begin{aligned}
\ln|\mathcal{C}_{\text{FC2}}| &\leq \ln|\mathcal{C}_A| + \ln|\mathcal{C}_{FC1}| + \max_{\hat{\mathbf{A}} \in \mathcal{C}_A} \ln|\mathcal{C}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\text{FC1}}, \tilde{\epsilon}, \|\cdot\|)| \\
&\leq O\left(\frac{B^2 \left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon_V^2} \ln(2dm)\right) + O\left(\frac{\alpha_1^2 K B^2 W^4 \left\|\mathbf{X}_*\right\|^4 \left\|\mathbf{X}_{[N]}\right\|^2}{d_K \epsilon_A^2} \ln(2d^2)\right) \\
&\quad + O\left(\frac{B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2 K^2 W^2 \left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon_{\text{FC1}}^2} \ln(dm)\right) \\
&\quad + O\left(\frac{\alpha_2^2 B^2 \left(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2 K^2 W^2 \left\|\mathbf{X}_{[N]}\right\|^2\right) W^2}{\epsilon_{\text{FC2}}^2} \ln(2dm)\right).
\end{aligned}$$

V: Verification of $\mathcal{C}_{\text{FC2}}$ being an ϵ cover of $\mathcal{H}_{\text{SA}}$ It remains to verify $\mathcal{C}_{FC2}$ is an ϵ cover of $\mathcal{H}_{SA}$. Given any $\mathbf{H}_{[N]} \in \mathcal{H}_{SA}$, we can find a $\hat{\mathbf{H}}_{[N]} \in \mathcal{C}_{\text{FC2}}$ such that

$$\begin{aligned}
\left\|\mathbf{H}_{[N]} - \hat{\mathbf{H}}_{[N]}\right\| &= \left\|\alpha_2 \sigma(\mathbf{Z}_{[N]} \mathbf{W}_{\text{FC1}}) \mathbf{W}_{\text{FC2}} + \mathbf{Z}_{[N]} - \alpha_2 \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}}) \hat{\mathbf{W}}_{\text{FC2}} - \hat{\mathbf{Z}}_{[N]}\right\| \\
&\leq \alpha_2 \left\|\sigma(\mathbf{Z}_{[N]} \mathbf{W}_{\text{FC1}}) \mathbf{W}_{\text{FC2}} - \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}}) \mathbf{W}_{\text{FC2}}\right\| \\
&\quad + \left\|\alpha_2 \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}}) \mathbf{W}_{\text{FC2}} + \mathbf{Z}_{[N]} - \alpha_2 \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}}) \hat{\mathbf{W}}_{\text{FC2}} - \hat{\mathbf{Z}}_{[N]}\right\| \\
&\leq \alpha_2 W \left\|\sigma(\mathbf{Z}_{[N]} \mathbf{W}_{\text{FC1}}) - \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}})\right\| + \epsilon_{\text{FC2}} + \left\|\mathbf{Z}_{[N]} - \hat{\mathbf{Z}}_{[N]}\right\|.
\end{aligned}$$

We bound $\left\|\sigma(\mathbf{Z}_{[N]} \mathbf{W}_{\text{FC1}}) - \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}})\right\|$ first as follows:

$$\begin{aligned}
\left\|\sigma(\mathbf{Z}_{[N]} \mathbf{W}_{\text{FC1}}) - \sigma(\hat{\mathbf{Z}}_{[N]} \hat{\mathbf{W}}_{\text{FC1}})\right\| &\leq \left\|(\alpha_1 \mathbf{A}_{[N]} + \mathbf{X}_{[N]}) \mathbf{W}_{\text{FC1}} - (\alpha_1 \hat{\mathbf{A}}_{[N]} + \mathbf{X}_{[N]}) \hat{\mathbf{W}}_{\text{FC1}}\right\| \\
&\leq \left\|(\alpha_1 \mathbf{A}_{[N]} + \mathbf{X}_{[N]}) \mathbf{W}_{\text{FC1}} - (\alpha_1 \hat{\mathbf{A}}_{[N]} + \mathbf{X}_{[N]}) \mathbf{W}_{\text{FC1}}\right\| \\
&\quad + \left\|(\alpha_1 \hat{\mathbf{A}}_{[N]} + \mathbf{X}_{[N]}) \mathbf{W}_{\text{FC1}} - (\alpha_1 \hat{\mathbf{A}}_{[N]} + \mathbf{X}_{[N]}) \hat{\mathbf{W}}_{\text{FC1}}\right\| \\
&\leq \alpha_1 W \left\|\mathbf{A}_{[N]} - \hat{\mathbf{A}}_{[N]}\right\| + \epsilon_{\text{FC1}}.
\end{aligned}$$

For $\left\|\mathbf{A}_{[N]} - \hat{\mathbf{A}}_{[N]}\right\|$, we have

$$\begin{aligned}
\left\|\mathbf{A}_{[N]} - \hat{\mathbf{A}}_{[N]}\right\| &= \left\|\mathbf{S}_{[N]} \mathbf{X}_{[N]} \mathbf{W}_V - \hat{\mathbf{S}}_{[N]} \mathbf{X}_{[N]} \hat{\mathbf{W}}_V\right\| \\
&\leq \left\|\mathbf{S}_{[N]} \mathbf{X}_{[N]} \mathbf{W}_V - \mathbf{S}_{[N]} \mathbf{X}_{[N]} \hat{\mathbf{W}}_V\right\| + \left\|\mathbf{S}_{[N]} \mathbf{X}_{[N]} \hat{\mathbf{W}}_V - \hat{\mathbf{S}}_{[N]} \mathbf{X}_{[N]} \hat{\mathbf{W}}_V\right\| \\
&\leq K \left\|\mathbf{X}_{[N]} \mathbf{W}_V - \mathbf{X}_{[N]} \hat{\mathbf{W}}_V\right\| + \epsilon_A \\
&\leq K \epsilon_V + \epsilon_A.
\end{aligned}$$

Putting pieces together yields:

$$\left\|\sigma(\mathbf{Z}\mathbf{W}_{\text{FC1}}) - \sigma(\hat{\mathbf{Z}} \hat{\mathbf{W}}_{\text{FC1}})\right\| \leq \alpha_1 W (K \epsilon_V + \epsilon_A) + \epsilon_{\text{FC1}}.$$

Now we switch to bounding $\left\|\mathbf{Z}_{[N]} - \hat{\mathbf{Z}}_{[N]}\right\|$:

$$\left\|\mathbf{Z}_{[N]} - \hat{\mathbf{Z}}_{[N]}\right\| = \alpha_1 \left\|\mathbf{A}_{[N]} - \hat{\mathbf{A}}_{[N]}\right\| \leq \alpha_1(K\epsilon_V + \epsilon_A)$$

Hence we know:

$$\begin{aligned}\left\|\mathbf{H}_{[N]} - \hat{\mathbf{H}}_{[N]}\right\| &\leq \alpha_2 W\left(\alpha_1 W(K\epsilon_V + \epsilon_A) + \epsilon_{\mathrm{FC}_1}\right) + \alpha_1(K\epsilon_V + \epsilon_A) + \epsilon_{\mathrm{FC2}} \\ &= (\alpha_1\alpha_2 W^2 K + \alpha_1 K)\epsilon_V + (\alpha_1\alpha_2 W^2 + \alpha_1)\epsilon_A + \alpha_2 W\epsilon_{\mathrm{FC1}} + \epsilon_{\mathrm{FC2}}\end{aligned}$$

To make sure RHS is less than ϵ, we set

$$\epsilon_V = \frac{\epsilon}{4(\alpha_1\alpha_2 W^2 K + \alpha_1 K)}, \epsilon_A = \frac{\epsilon}{4(\alpha_1\alpha_2 W^2 + \alpha_1)}, \epsilon_{\mathrm{FC1}} = \frac{\epsilon}{4\alpha_2 W}, \epsilon_{\mathrm{FC2}} = \frac{\epsilon}{4}.$$

Recall that

$$\begin{aligned}\ln|\mathcal{C}_{\mathrm{FC2}}| &\leq \ln|\mathcal{C}_A| + \ln|\mathcal{C}_{FC1}| + \max_{\hat{\mathbf{A}}\in\mathcal{C}_A} \ln|\mathcal{C}(\hat{\mathbf{A}}_{[N]} \circ \mathcal{W}_{\mathrm{FC1}}, \tilde{\epsilon}, \|\cdot\|)| \\ &\leq O\left(\frac{B^2\left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon_V^2}\ln(2dm)\right) + O\left(\frac{\alpha_1^2 K B^2 W^4\left\|\mathbf{X}_*\right\|^4\left\|\mathbf{X}_{[N]}\right\|^2}{d_K\epsilon_A^2}\ln(2d^2)\right) \\ &\quad + O\left(\frac{B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2 K^2 W^2\left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon_{\mathrm{FC1}}^2}\ln(dm)\right) \\ &\quad + O\left(\frac{\alpha_2^2 B^2\left(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2 K^2 W^2\left\|\mathbf{X}_{[N]}\right\|^2\right)W^2}{\epsilon_{\mathrm{FC2}}^2}\ln(2dm)\right).\end{aligned}$$

Hence we can upper bound the covering number of $\mathcal{H}_{SA}$ as follows:

$$\begin{aligned}\ln\mathcal{N}(\mathcal{H}_{SA}, \epsilon, \|\cdot\|) &\leq O\left(\frac{(\alpha_1\alpha_2 W^2 + \alpha_1)^2 K^2 B^2\left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon^2}\ln(2dm)\right) \\ &\quad + O\left(\frac{(\alpha_1\alpha_2 W^2 + \alpha_1)^2\alpha_1^2 K B^2 W^4\left\|\mathbf{X}_*\right\|^4\left\|\mathbf{X}_{[N]}\right\|^2}{d_K\epsilon^2}\ln(2d^2)\right) \\ &\quad + O\left(\frac{\alpha_2^2 W^2 B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2 K^2 W^2\left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon^2}\ln(2dm)\right) \\ &= O\left(\frac{(\alpha_1\alpha_2 W^2 + \alpha_1)^2 B^2\left\|\mathbf{X}_{[N]}\right\|^2}{\epsilon^2}\ln(2d^2)\right)\left(K^2 + \frac{\alpha_1 W^4\left\|\mathbf{X}_*\right\|^4}{d_K}\right) \\ &\quad + O\left(\frac{\alpha_2^2 W^2 B^2(\left\|\mathbf{X}_{[N]}\right\|^2 + \alpha_1^2 K^2 W^2\left\|\mathbf{X}_{[N]}\right\|^2)}{\epsilon^2}\ln(2dm)\right).\end{aligned}$$

□

Proposition 4 (Contraction mapping of self-attention layer)**.** *For a single attention layer parameterized by* $\mathbf{W}$*, with* $\|\mathbf{W}\| \leq W$*, the following statement holds:*

$$\left\|\mathrm{SA}_{\mathbf{W}}(\mathbf{X}) - \mathrm{SA}_{\mathbf{W}}(\hat{\mathbf{X}})\right\| \leq (\alpha_2 W^2 + 1)(\alpha_1 K W + 1)\left\|\mathbf{X} - \hat{\mathbf{X}}\right\|.$$

Proof. The proof follows by definition and simple algebraic manipulation:

$$\begin{aligned}\left\|\mathrm{SA}_{\mathbf{W}}(\mathbf{X})-\mathrm{SA}_{\mathbf{W}}(\hat{\mathbf{X}})\right\| &\leq \left\|\alpha_2\sigma\left(\mathbf{Z}\mathbf{W}_{\mathrm{FC1}}\right)\mathbf{W}_{\mathrm{FC2}}+\mathbf{Z}-\beta\sigma\left(\hat{\mathbf{Z}}\mathbf{W}_{\mathrm{FC1}}\right)\mathbf{W}_{\mathrm{FC2}}-\hat{\mathbf{Z}}\right\| \\ &\leq \alpha_2 W^2\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\|+\left\|\mathbf{Z}-\hat{\mathbf{Z}}\right\| \\ &\leq (\alpha_2W^2+1)\left\|\mathbf{A}+\mathbf{X}-\hat{\mathbf{A}}-\hat{\mathbf{X}}\right\| \\ &\leq (\alpha_2W^2+1)\left(\left\|\alpha_1\mathbf{S}\mathbf{X}\mathbf{W}_V-\alpha_1\mathbf{S}\hat{\mathbf{X}}\mathbf{W}_V\right\|+\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|\right) \\ &\leq (\alpha_2W^2+1)\left(\alpha_1KW+1\right)\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|.\end{aligned}$$

□

D.2.1 Proof of Lemma 8

Proof. We first examine the norm of each self-attention layer's output:

$$\begin{aligned}\left\|\mathbf{X}_i^l\right\| &\leq \alpha_2W^2(l)\left\|\mathbf{Z}_i^l\right\|+\left\|\mathbf{Z}_i^l\right\| \\ &\leq \left(\alpha_2W^2(l)+1\right)\left(\alpha\left\|\mathbf{A}_i^l\right\|+\left\|\mathbf{X}_i^{l-1}\right\|\right) \\ &\leq \left(\alpha_2W^2(l)+1\right)\left(W^2(l)\alpha K+1\right)\left\|\mathbf{X}^{l-1}\right\| \\ &\leq \prod_{j=1}^{l}\left(\alpha_2W^2(j)+1\right)\left(W^2(j)\alpha_1K+1\right)\left\|\mathbf{X}_i\right\| \end{aligned} \tag{16}$$

and grouped output

$$\begin{aligned}\left\|\mathbf{X}_{[N]}^l\right\| &\leq \alpha_2W^2(l)\left\|\mathbf{Z}_{[N]}^l\right\|+\left\|\mathbf{Z}_{[N]}^l\right\| \\ &\leq \left(\alpha_2W^2(l)+1\right)\left(\alpha\left\|\mathbf{A}_{[N]}^l\right\|+\left\|\mathbf{X}_{[N]}^{l-1}\right\|\right) \\ &\leq \left(\alpha_2W^2(l)+1\right)\left(W^2(l)\alpha K+1\right)\left\|\mathbf{X}_{[N]}^{l-1}\right\| \\ &\leq \prod_{j=1}^{l}\left(\alpha_2W^2(j)+1\right)\left(W^2(j)\alpha_1K+1\right)\left\|\mathbf{X}_{[N]}\right\|.\end{aligned}$$

For the ease of presentation, we define the class of lth layer output:

$$\mathcal{H}_l=\left\{\begin{array}{c}\mathrm{SA}^l\left(\mathrm{SA}^{l-1}...\mathrm{SA}^1(\mathbf{X})\right):\left\|\mathbf{W}_{\mathrm{FC1}}^j\right\|,\left\|\mathbf{W}_{\mathrm{FC2}}^j\right\|,\left\|\mathbf{W}_K^j\right\|,\left\|\mathbf{W}_Q^l\right\|,\left\|\mathbf{W}_V^j\right\|\leq W(j), \\ \left\|\mathbf{W}_{\mathrm{FC1}}^j\right\|_{2,1},\left\|\mathbf{W}_{\mathrm{FC2}}^j\right\|_{2,1},\left\|\mathbf{W}_K^j\right\|_{2,1},\left\|\mathbf{W}_Q^j\right\|_{2,1},\left\|\mathbf{W}_V^j\right\|_{2,1}\leq B(j),\forall j\in[l]\end{array}\right\},$$

and it will be useful to define set of weight matrices at lth layer:

$$\mathcal{W}_l=\left\{\mathbf{W}:\|\mathbf{W}\|\leq W(l),\|\mathbf{W}\|_{2,1}\leq B(l).\right\}.$$

We shall construct the cover with certain radius for each $\mathcal{H}_l, l\in[L]$.

For base case $l=1$: we create ϵ_1 cover of $\mathrm{SA}^1(\mathbf{X}_{[N]})$

$$\mathcal{C}_1=\mathcal{C}(\mathrm{SA}^1(\mathbf{X}_{[N]}),\epsilon_1,\|\cdot\|).$$

For $1<l+1\leq L$, for each element $\hat{\mathbf{X}}^l\in\mathcal{C}_l$, we construct the ϵ_{l+1}-cover of the following set:

$$\mathrm{SA}^{l+1}(\hat{\mathbf{X}}^l):=\left\{\mathrm{SA}_{\mathbf{W}^{l+1}}(\hat{\mathbf{X}}^l),\mathbf{W}^{l+1}\in\mathcal{W}_{l+1}\right\}$$

and we denote the cover as $\mathcal{C}\left(\mathrm{SA}^{l+1}(\hat{\mathbf{X}}^l), \epsilon_{l+1}, \|\cdot\|\right)$. We first examine the cardinality of this cover as follows:

$$
\begin{aligned}
&\ln\left|\mathcal{C}\left(\mathrm{SA}^{l+1}(\mathbf{X}^l_{[N]}), \epsilon_{l+1}, \|\cdot\|\right)\right| \\
&\leq \max_{\mathbf{X}^l\in\mathcal{C}_l} O\left(\frac{(\alpha_1\alpha_2W^2+\alpha_1)^2B^2}{\epsilon^2}\ln(2d^2)\right)\left(K^2+\frac{\alpha_1W^4\left(\max_{i\in[N]}\left\|\mathbf{X}^l_i\right\|\right)^4}{d_K}\right)\left\|\mathbf{X}^l_{[N]}\right\|^2 \\
&\quad + O\left(\frac{\alpha_2^2W^2(l+1)B^2(l+1)(1+\alpha_1^2K^2W^2(l+1))}{\epsilon^2}\ln(2dm)\right)\left\|\mathbf{X}^l_{[N]}\right\|^2 \\
&\leq O\left(\frac{(\alpha_1\alpha_2W^2(l+1)+\alpha_1)^2B^2(l+1)}{\epsilon^2}\ln(2d^2)\right)\left(K^2+\frac{\alpha_1W^4(l+1)\left(s_l\|\mathbf{X}_*\|\right)^4}{d_K}\right)s_l^2\left\|\mathbf{X}_{[N]}\right\|^2 \\
&\quad + O\left(\frac{\alpha_2^2W^2(l+1)B^2(l+1)(1+\alpha_1^2K^2W^2(l+1))}{\epsilon^2}\ln(2dm)\right)s_l^2\left\|\mathbf{X}_{[N]}\right\|^2 \\
&:= \ln N_{l+1}
\end{aligned}
$$

where $s_l := \prod_{j=1}^{l}\left(\alpha_2W^2(j)+1\right)\left(W^2(j)\alpha_1K+1\right)$.

We then construct cover for $\mathcal{H}_{l+1}$ as:

$$
\mathcal{C}_{l+1} = \bigcup_{\mathbf{X}^l\in\mathcal{C}_l}\mathcal{C}\left(\mathrm{SA}^{l+1}(\mathbf{X}^l), \epsilon_{l+1}, \|\cdot\|\right).
$$

It is not hard to check the cardinality of $\mathcal{C}_{l+1}$

$$
|\mathcal{C}_{l+1}| = \left|\bigcup_{\mathbf{X}^l\in\mathcal{C}_l}\mathcal{C}\left(\mathrm{SA}^{l+1}(\mathbf{X}^l), \epsilon_{l+1}, \|\cdot\|\right)\right| \leq |C_l|N_{l+1} \leq \prod_{l'=1}^{l+1}N_{l'}
$$

Let

$$
\begin{aligned}
\rho_l := &\, O\left((\alpha_1\alpha_2W^2(l)+\alpha_1)^2B^2(l)\ln(2d^2)\right)\left(K^2+\frac{\alpha_1W^4(l)\left(s_{l-1}\|\mathbf{X}_*\|\right)^4}{d_K}\right) \\
&+ O\left(\alpha_2^2W^2(l)B^2(l)(1+\alpha_1^2K^2W^2(l))\ln(2dm)\right),
\end{aligned}
$$

we have

$$
\ln|\mathcal{C}_{l+1}| \leq \sum_{l'=1}^{l+1}\ln N_{l'} \leq \sum_{l'=1}^{l+1}\frac{\rho_{l'}}{\epsilon_{l'}^2}s_{l'}^2\left\|\mathbf{X}_{[N]}\right\|^2.
$$

Now it remains to verify $\mathcal{C}_L$ is a cover of $\mathcal{H}_L$. For any $\mathbf{X}^L\in\mathcal{H}_L$ we can find a $\hat{\mathbf{X}}^L\in\mathcal{C}_L$ such that

$$
\begin{aligned}
\left\|\mathbf{X}^L-\hat{\mathbf{X}}^L\right\| &= \left\|\mathrm{SA}_{\mathbf{W}^L}(\mathbf{X}^{L-1})-\mathrm{SA}_{\hat{\mathbf{W}}^L}(\hat{\mathbf{X}}^{L-1})\right\| \\
&\leq \left\|\mathrm{SA}_{\mathbf{W}^L}(\mathbf{X}^{L-1})-\mathrm{SA}_{\mathbf{W}^L}(\hat{\mathbf{X}}^{L-1})\right\| + \left\|\mathrm{SA}_{\mathbf{W}^L}(\hat{\mathbf{X}}^{L-1})-\mathrm{SA}_{\hat{\mathbf{W}}^L}(\hat{\mathbf{X}}^{L-1})\right\| \\
&\leq \left(\alpha_2W^2(L)+1\right)\left(\alpha_1KW(L)+1\right)\left\|\mathbf{X}^{L-1}-\hat{\mathbf{X}}^{L-1}\right\| + \epsilon_L \\
&\leq \sum_{l=0}^{L}\prod_{j=l+1}^{L}\left(\alpha_2W^2(j)+1\right)\left(\alpha_1KW(j)+1\right)\epsilon_l
\end{aligned}
$$

We choose $\epsilon_j = \left(L\prod_{j=l+1}^{L}(\alpha_2W^2(j)+1)\left(\alpha_1KW(j)+1\right)\right)^{-1}\epsilon$, and let $s_{l+1\mapsto L} := \prod_{j=l+1}^{L}(\alpha_2W^2(j)+1)\left(\alpha_1KW(j)+1\right)$. Hence we have:

$$
\frac{\rho_l}{\epsilon_l^2} = \frac{\rho_l s_{l+1\mapsto L}^2}{\epsilon^2}\left\|\mathbf{X}_{[N]}\right\|^2
$$

and conclude the covering number of $\mathcal{H}_L$ as follows:

$$\begin{aligned}\ln \mathcal{N}(\mathcal{H}_L, \epsilon, \|\cdot\|) = \ln |\mathcal{C}_L| &\leq \sum_{l=1}^{L} \frac{\rho_l}{\epsilon_l^2} \left(s_l \left\|\mathbf{X}_{[N]}\right\|\right)^2 \\ &= \ln |\mathcal{C}_L| \leq O\left(s_L^2 \left\|\mathbf{X}_{[N]}\right\|^2 \sum_{l=1}^{L} \frac{\rho_l}{\epsilon^2}\right).\end{aligned}$$

Finally according to covering number fact (13),

$$\ln \mathcal{N}_\infty(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|) \leq \ln \mathcal{N}(\mathcal{G} \circ \mathcal{H}(S), \epsilon, \|\cdot\|), \tag{17}$$

we can conclude the proof. □

Now, equipped with covering number bound for the transformer, we are ready to show the generalization of MAE pre-training task.

D.2.2 Proof of Lemma 3

Proof. Similar to the proof in CE section, we evoke Lemma 7 with $c = O\left(s_{L+1}^2 \left\|\tilde{\mathbf{Z}}_{[N]}\right\|^2 \sum_{l=1}^{L+1} \rho_l\right)$, where s_l, ρ_l are defined in Lemma 8.

$$\begin{aligned}\mathfrak{R}_{\widehat{\mathcal{S}}}\left(\mathcal{L}(r)\right) &\leq 10\sqrt{\frac{Hr \cdot c}{N}} + 10\sqrt{\frac{cHr}{N}} \left(\ln \sqrt{br} - \ln\left(\frac{5}{2}\sqrt{\frac{Hr \cdot c}{N}}\right)\right) \\ &= 10\sqrt{\frac{Hr \cdot c}{N}} + 10\sqrt{\frac{cHr}{N}} \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right).\end{aligned}$$

We set $\phi(r) = 10\sqrt{\frac{Hr \cdot c}{N}} \cdot \max\left\{1, \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)\right\}$. Solving the follwoing equation to get r^*

$$\begin{aligned}&\phi(r) = 10\sqrt{\frac{Hr \cdot c}{N}} \cdot \max\left\{1, \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)\right\} = r, \\ \iff &r^* = 100\frac{H \cdot c}{N} \cdot \max\left\{1, \ln\left(\frac{2}{5}\sqrt{\frac{bN}{Hc}}\right)\right\}^2\end{aligned}$$

Now, according to Theorem 7, and the fact that

$$A \leq B + C\sqrt{A} \Longrightarrow A \leq B + C^2 + \sqrt{B}C,$$

we have

$$\begin{aligned}\mathcal{L}_{\mathcal{U}}(g \circ h) \leq \mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) + 45r^* &+ \left(\sqrt{8r^*} + \sqrt{\frac{4b(\log(1/\nu) + 6\log\log N)}{N}}\right)^2 \\ &+ 20\frac{b(\nu + 6\log\log N)}{N} \\ &+ \sqrt{\mathcal{L}_{\widehat{\mathcal{U}}}(g \circ h) + 45r^* + 20\frac{b(\nu + 6\log\log N)}{N}} \left(\sqrt{8r^*} + \sqrt{\frac{4b(\log(1/\nu) + 6\log\log N)}{N}}\right)\end{aligned}$$

Plugging r^* and empirical risk minimizers $\hat{g}, \hat{h}$ will conclude the proof. □

D.3 Proof of Theorem 3

Proof. Again, recall in Theorem 1, the generalization bound of downstream task is given by

$$\mathcal{E}_{\mathcal{T}}(\hat{f},\hat{h}) \le C_\beta \left(\mathcal{E}_{\mathcal{U}}(\hat{g},\hat{h})\right)^\beta + 4G_\phi \mathfrak{R}_{\widehat{\mathcal{T}}}(\mathcal{F}\circ\hat{h}) + 4B_\phi\sqrt{\frac{\log(1/\nu)}{n}} + 4B_\phi \left\|\mathcal{T}-\mathcal{U}_{\mathcal{X}}\right\|_{\mathrm{TV}} + \min_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{T}}(f, h^*_{\mathcal{U}}).$$

Since in the previous subsection we prove the bounded transferrability and generalization of pre-training task, it remains to show the upper bound of representation-induced Rademacher complexity.

$$\begin{aligned}
\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi\circ\mathcal{F}\circ\hat{h}) &= \mathbb{E}_{\boldsymbol{\varepsilon}\in\{\pm1\}^n}\left[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|\le R}\frac{1}{n}\sum\nolimits_{i=1}^n \varepsilon_i\phi(\boldsymbol{\theta}^\top(\mathbf{1}^\top\hat{h}(\mathbf{X}_i))^\top, y_i)\right]\\
&\le G_\phi\mathbb{E}_{\boldsymbol{\varepsilon}\in\{\pm1\}^n}\left[\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|\le R}\frac{1}{n}\sum\nolimits_{i=1}^n \varepsilon_i\boldsymbol{\theta}^\top(\mathbf{1}^\top\hat{h}(\mathbf{X}_i))^\top\right]\\
&= \frac{RG_\phi}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum\nolimits_{i=1}^n \varepsilon_i(\mathbf{1}^\top\hat{h}(\mathbf{X}_i))^\top\right\|\\
&\le \frac{RG_\phi}{n}\sqrt{\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum\nolimits_{i=1}^n \varepsilon_i(\mathbf{1}^\top\hat{h}(\mathbf{X}_i))^\top\right\|^2}\\
&\le \frac{RG_\phi}{n}\sqrt{\sum\nolimits_{i=1}^n\left\|\mathbf{1}^\top\hat{h}(\mathbf{X}_i)\right\|^2}\\
&\le \frac{RG_\phi}{n}\sqrt{\sum\nolimits_{i=1}^n K\left\|\hat{h}(\mathbf{X}_i)\right\|^2}
\end{aligned}$$

where at first inequality we apply Ledoux-Talagrand's inequality to peel of Lipschitz loss $\phi(\cdot)$, and at last inequality we use the fact that ε_i are i.i.d. with zero mean, so that the cross terms disappear. For each $\left\|\hat{h}(\mathbf{X}_i)\right\|^2$, evoking (16) we have:

$$\left\|\hat{h}(\mathbf{X}_i)\right\| \le \prod_{j=1}^{l}\left(\alpha_2 W^2(j)+1\right)\left(W^2(j)\alpha_1 K+1\right)\|\mathbf{X}_i\|,$$

hence we arrive at

$$\mathfrak{R}_{\widehat{\mathcal{T}}}(\phi\circ\mathcal{F}\circ\hat{h}) \le \frac{RG_\phi\sqrt{\prod_{j=1}^{l}\left(\alpha_2 W^2(j)+1\right)^2\left(W^2(j)\alpha_1 K+1\right)^2\sum_{i=1}^n\|\mathbf{X}_i\|^2}}{n}.$$

Plugging Lemmas 2 and 3 as well as above bound will complete the proof. □

E Proof of Convergence RadReg Algorithm

In this section we provide the missing proofs from Section 6. Then we provide the proof of convergence.

E.1 Convergence result of RadReg

In this section we provide formal version of convergence results for RadReg. First let us introduce the following Moreau envelope concept.

Definition 6 (Moreau Envelope)**.** *A function $\Psi_\rho(\mathbf{w})$ is the ρ-Moreau envelope of a function Ψ if $\Psi_\rho(\mathbf{w}) := \min_{\mathbf{w}'\in\mathcal{W}}\{\Psi(\mathbf{w}')+\frac{1}{2\rho}\|\mathbf{w}'-\mathbf{w}\|^2\}$.*

We have the following property of the Moreau Envelope of a nonsmooth function:

Lemma 11. *[Davis and Drusvyatskiy, 2019] Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}'\in\mathcal{W}}\Psi(\mathbf{w}')+\frac{1}{2\rho}\|\mathbf{w}'-\mathbf{w}\|^2$, then we have the following facts: $\|\hat{\mathbf{w}}-\mathbf{w}\| \le \rho\|\nabla\Phi_\rho(\mathbf{w})\|$, $\min_{\mathbf{g}\in\partial\Psi(\hat{\mathbf{w}})}\|\mathbf{g}\| \le \|\nabla\Phi_\rho(\mathbf{w})\|$.*

Lemma 11 shows that, if we find a $\mathbf{w}$ such that $\|\nabla\Psi_\rho(\mathbf{w})\|$ is small, then we can demonstrate that $\mathbf{w}$ is near some point $\hat{\boldsymbol{x}}$ which is a near-stationary point of Ψ. We will use $1/4L$-Moreau envelope of Ψ, following the setting in [Lin et al., 2020, Rafique et al., 2018], and state the convergence rate in terms of $\|\nabla\Psi_{1/4L}(\mathbf{w})\|$. We also define quantity $\hat{\Delta}_{\Psi_{1/4L}} = \Psi_{1/4L}(\mathbf{w}_0) - \min_{\mathbf{w}\in\mathcal{W}} \Psi_{1/4L}(\mathbf{w})$ that will be used in stating the convergence rates.

Assumption 2 (Bounded Variance). *Let $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{x}}$ be uniformly sampled from $\hat{\mathcal{U}}$ and $\hat{\mathcal{D}}$. Then, the variance of stochastic gradients is bounded:*

$$\mathbb{E}\left\|\nabla\mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w};\tilde{\mathbf{z}}) - \nabla\mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w})\right\|^2 \le \delta^2,$$
$$\mathbb{E}\left\|\nabla\mathfrak{R}_j(\mathbf{v},\mathbf{w};\tilde{\mathbf{x}}) - \frac{1}{n}\sum\nolimits_{i=1}^{n}\nabla\mathfrak{R}_j(\mathbf{v},\mathbf{w};\mathbf{x}_i)\right\|^2 \le \delta^2.$$

Assumption 3 (Smooth and Bounded Linear Head). *$\mathcal{L}_{\hat{\mathcal{U}}}$ and $\mathfrak{R}_j(\mathbf{v},\mathbf{w}';\mathbf{x})$ are L smooth w.r.t. $\mathbf{w}$, $\forall j\in[B]$, and $\mathbf{x}\in\mathcal{X}$:*

$$\left\|\nabla\mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w}) - \nabla\mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w}')\right\| \le L\left\|\mathbf{w}-\mathbf{w}'\right\|,$$
$$\left\|\nabla_{\mathbf{w}}\mathfrak{R}_j(\mathbf{v},\mathbf{w};\mathbf{x}) - \nabla_{\mathbf{w}}\mathfrak{R}_j(\mathbf{v},\mathbf{w}';\mathbf{x})\right\| \le L\left\|\mathbf{w}-\mathbf{w}'\right\|.$$

Also, we assume $\mathfrak{R}_j(\mathbf{v},\mathbf{w};\mathbf{x})$ is linear in $\mathbf{v}$, and $\max_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|\le D$.

Assumption 4 (Lipschitzness). *$\mathcal{L}_{\hat{\mathcal{U}}}$ and $\mathfrak{R}_j(\mathbf{v},\mathbf{w}';\mathbf{x})$ are G Lipschitz w.r.t. $\mathbf{w}$, $\forall\mathbf{v}\in\mathcal{V}, j\in[B]$, and $\mathbf{x}\in\mathcal{X}$, i.e.,*

$$\left\|\mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w}) - \mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w}')\right\| \le G\left\|\mathbf{w}-\mathbf{w}'\right\|,$$
$$\left\|\mathfrak{R}_j(\mathbf{v},\mathbf{w};\mathbf{x}) - \mathfrak{R}_j(\mathbf{v},\mathbf{w}';\mathbf{x})\right\| \le G\left\|\mathbf{w}-\mathbf{w}'\right\|.$$

We are now ready to state the formal version of Theorem 4 as follows.

Theorem 8 (Convergence of RadReg with Linear Top Layer). *Under Assumptions 2 and 3, if we use RadReg (Algorithm 1 with one step update) to optimize (8), by choosing $\eta = \Theta\left(\frac{\epsilon^6}{L^3D^2G}\right)$ and $\gamma = \Theta\left(\frac{\epsilon^2}{L\delta^2}\right)$ it holds that:*

$$\frac{1}{T+1}\sum\nolimits_{t=0}^{T}\mathbb{E}\left\|\Psi_{1/4L}(\mathbf{w}^t)\right\|^2 \le \epsilon^2,$$

with the gradient complexity bounded by:

$$O\left(\frac{BL^3(G^2+\frac{\delta^2}{n'})D^2\frac{\delta^2}{n'}\Delta_{\Psi_{1/4L}}}{\epsilon^8}\right).$$

We can see that the proposed optimization algorithm can find an ϵ stationary point with at most $O\left(\frac{B}{\epsilon^8}\right)$ stochastic gradient evaluations. Since the complexity grows in terms of B, a proper sampling size of Rademacher variable is crucial.

In the rest of this section, we prove the convergence rate of RadReg. The proof idea mainly follows the framework developed in Lin et al. [2019]. But before we state a few intermediate results that the main proof relies on.

Lemma 12. *Under the conditions of Theorem 8, the following one iteration recursion relation holds true:*

$$\eta\mathbb{E}\left\|\nabla\Psi(\mathbf{w}^{t-1})\right\|^2 = \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^{t-1})] - \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^t)]$$
$$+ 4\eta L\frac{1}{B}\sum_{j=1}^{B}\mathbb{E}\left[\mathfrak{R}_j(\mathbf{v}_j^*(\mathbf{w}^{t-1}),\mathbf{w}^{t-1}) - \mathfrak{R}_j(\mathbf{v}_j^{t-1},\mathbf{w}^{t-1})\right] + \eta^2L(G^2+\frac{\delta^2}{n'}),$$

where $\mathbf{v}_j^*(\mathbf{w}) := \arg\max_{\mathbf{v}\in\mathcal{V}}\mathfrak{R}_j(\mathbf{v},\mathbf{w}^{t-1})$.

Proof. Recall the definition of Ψ and $\mathfrak{R}_j$:

$$\Psi(\mathbf{w}) := \mathcal{L}_{\hat{\mathcal{U}}}(\mathbf{w}) + \lambda\frac{1}{B}\sum_{j=1}^{B}\left[\max_{\mathbf{V}\in\mathcal{V}}\mathbf{v}^\top\left(\frac{1}{n}\sum_{i}^{n}\sigma_i^j h_{\mathbf{w}}(\mathbf{x}_i)\right)\right], \tag{18}$$

$$\mathfrak{R}_j(\mathbf{v},\mathbf{w}) := \mathbf{v}^\top\left(\frac{1}{n}\sum_{i}^{n}\sigma_i^j h_{\mathbf{w}}(\mathbf{x}_i).\right) \tag{19}$$

Also recall the definition of Ψ's Moreau Envelope:

$$\Psi_{1/4L}(\mathbf{w}) := \min_{\mathbf{w}' \in \mathcal{W}} \Psi(\mathbf{w}') + 2L \left\| \mathbf{w} - \mathbf{w}' \right\|^2 .$$

We define the proximal solution as:

$$\hat{\mathbf{w}}^t := \arg\min_{\mathbf{w}' \in \mathcal{W}} \Psi(\mathbf{w}') + 2L \left\| \mathbf{w}^t - \mathbf{w}' \right\|^2 .$$

With all aforementioned definitions are in place, we proceed to proving the lemma. First, since $\hat{\mathbf{w}}^{t-1}$ is not minimizer of $\Psi(\cdot) + 2L \left\| \cdot - \mathbf{w}^t \right\|^2$ we have

$$\mathbb{E}[\Psi_{1/4L}(\mathbf{w}^t)] \leq \mathbb{E}[\Psi(\hat{\mathbf{w}}^{t-1})] + 2L \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^t \right\|^2$$

Recall the updating rule for $\mathbf{w}$ and $\mathbf{v}_j$ as stated below:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \left(\frac{1}{n'} \sum_{i=1}^{n'} \nabla \mathcal{L}(\mathbf{w}^t; \tilde{\mathbf{x}}_i^t) + \lambda \frac{1}{B} \sum_{j=1}^{B} \frac{1}{n'} \sum_{i=1}^{n'} \nabla_{\mathbf{w}} \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t; \tilde{\mathbf{x}}_i^t) \right),$$

$$\mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \gamma\lambda \frac{1}{B} \sum_{j=1}^{B} \frac{1}{n'} \sum_{i=1}^{n'} \nabla_{\mathbf{v}} \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t; \tilde{\mathbf{x}}_i^t).$$

Hence we can get the following relation by completing the square trick:

$$\begin{aligned} \mathbb{E} \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^t \right\|^2 &= \mathbb{E} \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\|^2 \\ &\quad + 2\eta \mathbb{E} \left\langle \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1},\, \nabla \mathcal{L}(\mathbf{w}^{t-1}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \nabla_{\mathbf{v}} \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right\rangle + \eta^2 (G^2 + \frac{\delta^2}{n'}). \end{aligned}$$

According to L-smoothness of $\mathcal{L}$ and $\mathfrak{R}_j$, we can re-write the inner product term as:

$$\begin{aligned} &\mathbb{E} \left\langle \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1},\, \nabla \mathcal{L}(\mathbf{w}^{t-1}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \nabla_{\mathbf{v}} \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right\rangle \\ &\leq \mathbb{E} \left[\mathcal{L}(\hat{\mathbf{w}}^{t-1}) - \mathcal{L}(\mathbf{w}^{t-1}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \left(\mathfrak{R}_j(\mathbf{v}_j^{t-1}, \hat{\mathbf{w}}^{t-1}) - \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right) \right] + L \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\| \end{aligned}$$

Notice the following fact about Ψ, $\Psi_{1/4L}$:

$$\mathcal{L}(\hat{\mathbf{w}}^{t-1}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \hat{\mathbf{w}}^{t-1}) \leq \Psi(\hat{\mathbf{w}}^{t-1}) \leq \Psi_{1/4L}(\mathbf{w}^{t-1}) - 2L \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\|^2 .$$

The last inequality is because $\hat{\mathbf{w}}^{t-1}$ is the minimizer of $\Psi(\cdot) + 2L \left\| \cdot - \mathbf{w}^{t-1} \right\|^2$. As a result, the inner product is bounded by:

$$\mathbb{E} \left\langle \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1},\, \nabla \mathcal{L}(\mathbf{w}^{t-1}) + \lambda \frac{1}{B} \sum_{j=1}^{B} \nabla_{\mathbf{v}} \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right\rangle \leq \mathbb{E} \left[\Psi(\mathbf{w}^{t-1}) - F(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right] - L\mathbb{E} \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\| .$$

Finally, putting pieces together and using the fact that $\nabla \Psi_{1/4L}(\mathbf{w}^{t-1}) = \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\| / 4L$ will conclude the proof:

$$\begin{aligned} \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^t)] &\leq \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^{t-1})] + 4\eta L \mathbb{E} \left[\Psi(\mathbf{w}^{t-1}) - F(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right] - 4\eta L \mathbb{E} \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\| + 2\eta^2 L (G^2 + \frac{\delta^2}{n'}) \\ &= \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^{t-1})] + 4\eta L \frac{1}{B} \sum_{j=1}^{B} \mathbb{E} \left[\mathfrak{R}_j(\mathbf{v}^*(\mathbf{w}^{t-1}), \mathbf{w}^{t-1}) - \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1}) \right] \\ &\quad - 4\eta L^2 \mathbb{E} \left\| \hat{\mathbf{w}}^{t-1} - \mathbf{w}^{t-1} \right\| + 2\eta^2 L (G^2 + \frac{\delta^2}{n'}). \end{aligned}$$

□

Lemma 13 (Lemma D4 in [Lin et al., 2019]). *If $\mathfrak{R}_j(\mathbf{v}, \mathbf{w})$ is convex and smooth in $\mathbf{v}$, L smooth and G Lipschitz in $\mathbf{w}$, then under the dynamic of stochastic gradient descent ascent on $\mathbf{v}$, we have the following statement holding:*

$$\begin{aligned}
&\mathbb{E}[\mathfrak{R}_j(\mathbf{v}^*(\mathbf{w}^{t-1}), \mathbf{w}^{t-1}) - \mathfrak{R}_j(\mathbf{v}_j^{t-1}, \mathbf{w}^{t-1})] \\
&\leq \eta G\sqrt{G^2 + \delta^2/n'}(2t - 2s - 1) + \frac{1}{2\gamma}\left(\mathbb{E}\left\|\mathbf{v}^*(\mathbf{w}^s) - \mathbf{v}_j^{t-1}\right\|^2 - \mathbb{E}\left\|\mathbf{v}^*(\mathbf{w}^s) - \mathbf{v}_j^{t}\right\|^2\right) \\
&\quad + \mathbb{E}[\mathfrak{R}_j(\mathbf{v}^t, \mathbf{w}^t) - \mathfrak{R}_j(\mathbf{v}^{t-1}, \mathbf{w}^{t-1})] + \frac{\gamma\delta^2}{2n'}
\end{aligned}$$

and

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\mathfrak{R}_j(\mathbf{v}^*(\mathbf{w}^t), \mathbf{w}^t) - \mathfrak{R}_j(\mathbf{v}_j^t, \mathbf{w}^t)] \leq \eta G S^2\sqrt{G^2 + \sigma^2} + \frac{D^2}{2S\gamma} + \frac{\gamma\delta^2}{2n'} + \frac{\max_{\mathbf{v}}\mathfrak{R}(\mathbf{v}, \mathbf{w}^0) - \mathfrak{R}(\mathbf{v}^0, \mathbf{w}^0)}{T+1}.$$

where $D = \max_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|$,

E.2 Proof of Theorem 8

Now we are ready to present proof of Theorem 8 by putting the above results together.

Proof. Summing Lemma 12 from $t = 0$ to T yields:

$$\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla\Psi(\mathbf{w}^t)\right\|^2 &= \frac{\mathbb{E}[\Psi_{1/2L}(\mathbf{w}^0)] - \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^T)]}{\eta(T+1)} + 4L\left(\eta G S\sqrt{G^2 + \sigma^2} + \frac{D^2}{2S\gamma} + \frac{\gamma\delta^2}{2}\right) \\
&\quad + 4\frac{1}{B}\sum_{j=1}^{B}\frac{L(\max_{\mathbf{v}}\mathfrak{R}_j(\mathbf{v}, \mathbf{w}^0) - \mathfrak{R}_j(\mathbf{v}^0, \mathbf{w}^0))}{T+1} + 4\eta L\sqrt{G^2 + \delta^2/n'}.
\end{aligned}$$

Setting $S = \frac{D}{2}\sqrt{\frac{1}{\eta\gamma G\sqrt{G^2+\delta^2/n'}}}$ yields:

$$\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla\Psi(\mathbf{w}^t)\right\|^2 &= O\left(\frac{\mathbb{E}[\Psi_{1/2L}(\mathbf{w}^0)] - \mathbb{E}[\Psi_{1/2L}(\mathbf{w}^T)]}{\eta(T+1)}\right) + O\left(LD\sqrt{\frac{\eta G\sqrt{G^2+\delta^2}}{\gamma}} + \frac{L\gamma\delta^2}{2}\right) \\
&\quad + O\left(\frac{L(\max_{\mathbf{v}}\mathfrak{R}(\mathbf{v}, \mathbf{w}^0) - \mathfrak{R}(\mathbf{v}^0, \mathbf{w}^0))}{T+1} + \eta L\sqrt{G^2+\delta^2}\right).
\end{aligned}$$

Finally, by choosing $\eta = \Theta\left(\frac{\epsilon^6}{L^3D^2G}\right)$ and $\gamma = \Theta\left(\frac{\epsilon^2}{L\delta^2}\right)$, we can guarantee the stationary of past iterates:

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\Psi_{1/4L}(\mathbf{w}^t)\right\| \leq \epsilon,$$

with the gradient complexity bounded by

$$O\left(\frac{BL^3(G^2 + \delta^2/n')D^2\delta^2\Delta_{\Psi_{1/4L}}}{\epsilon^8}\right).$$

as stated. □

F Experiment Details

Recall we utilize the Masked AutoEncoder (MAE) [He et al., 2022] as the base unsupervised pre-training method. For models, we use the Tiny Vision Transform (TinyViT) [Wu et al., 2022] as the backbone for pre-training

and use a 10-way linear classifier on top of the encoder for fine-tuning. The encoder h sequentially contains one convolutional layer, 12 192-head attention blocks, and one layer-normalization layer. The decoder g for reconstructing images in MAE includes 4 192-head attention blocks followed by one linear layer.Details of hyperparameters for the experiments reported in Table 1 are included in Table 2. For RadReg, we sample σ for 50 times and solve the inner maximization by Adam optimizer with a learning rate of 0.001 and a weight decay of 5×10^{-4}.

Config	**Value**
Optimizer	AdamW
Base learning rate	1.5×10^{-4}
Optimizer momentum	$\beta = 0.9, 0.95$
Batch size	4096
Learning rate schedule	cosine decay
Warmup epochs	200
Augmentation	RandomResizedCrop
Masking ratio	75%
Pre-training epochs	2000
Fine-tuning epochs	300

Table 2: Pre-training setting. Fine-tuning follows the same setting except for the number of epochs.