# Provable Target Sample Complexity Improvements as Pre-Trained Models Scale

**Kazuto Fukuchi**
University of Tsukuba
RIKEN AIP

**Ryuichiro Hataya**
SB Intuitions Corp.
Kyoto University

**Kota Matsui**
Kyoto University
Shiga University
Institute of Science Tokyo

## Abstract

Pre-trained models have become indispensable for efficiently building models across a broad spectrum of downstream tasks. The advantages of pre-trained models have been highlighted by empirical studies on scaling laws, which demonstrate that larger pre-trained models can significantly reduce the sample complexity of downstream learning. However, existing theoretical investigations of pre-trained models lack the capability to explain this phenomenon. In this paper, we provide a theoretical investigation by introducing a novel framework, *caulking*, inspired by parameter-efficient fine-tuning (PEFT) methods such as adapter-based fine-tuning, low-rank adaptation, and partial fine-tuning. Our analysis establishes that improved pre-trained models provably decrease the sample complexity of downstream tasks, thereby offering theoretical justification for the empirically observed scaling laws relating pre-trained model size to downstream performance, a relationship not covered by existing results.

## 1 INTRODUCTION

The utilization of pre-trained models across diverse domains has become a prevalent strategy for developing models tailored to specific applications. This approach enables the construction of highly accurate models even in scenarios where domain-specific data is limited. For instance, in medical image recognition, numerous pre-trained models have been developed for a variety of tasks, such as disease diagnosis and the identification of diseased regions (Wen et al., 2021). More recently, foundation models that handle different modalities, such as chest X-ray and brain CT images, within a unified framework have also been developed (Azad et al., 2023; Wang et al., 2025). Moreover, in fields such as drug discovery and materials science, pre-trained and foundation models capable of handling chemical structures are becoming powerful tools (Xia et al., 2022; Pyzer-Knapp et al., 2025).

The advantage of leveraging large pre-trained models has been underscored by empirical studies on *scaling laws* (Henighan et al., 2020; Mikami et al., 2023). Scaling laws were first conceptualized by Kaplan et al. (2020) in the context of large language models (LLMs), demonstrating that the performance of LLMs scales with model size, dataset size, and the amount of compute used for training. The scaling laws of pre-trained models were further investigated by Henighan et al. (2020) in the context of pre-trained autoregressive models, showing that larger pre-trained models can significantly reduce the sample complexity for fine-tuning downstream tasks. Mikami et al. (2023) also demonstrated analogous scaling laws for pre-trained models in the context of synthetic-to-real transfer learning, showing that the performance of pre-trained models scales with the amount of pre-training data. A special attention of this paper is *data scaling laws* of pre-trained models, where the performance for downstream tasks scales with the amount of pre-training data.

The effectiveness of pre-trained models is also supported by substantial theoretical research, including work on few-shot learning (Du et al., 2020) and in-context learning (Bai et al., 2023b; Kim et al., 2024). Most of these studies attempt to demonstrate the advantage of pre-trained models by establishing an upper bound on the error of the learning algorithm, characterized by the source sample size $m$ (used to construct the pre-trained model) and the target sample size $n$ (sample size for the downstream task). Specifi-

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. 

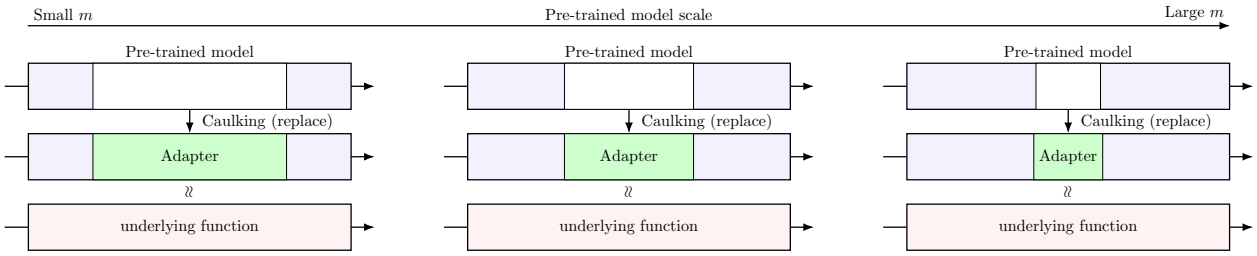


Figure 1: A conceptual illustration of caulking. Blue boxes represent pre-trained models, and red boxes represent underlying functions. The horizontal axis represents the source sample size $m$, which corresponds to the scale of the pre-trained model.

cally, the error rate obtained in these analyses is often expressed as:

$$\mathbb{E}[E_Q(f_n)] \leq m^{-\alpha} + n^{-\beta}, \tag{1}$$

where $E_Q(f_n)$ denotes the error of the estimated regression function $f_n$ under the target distribution $Q$ of the downstream task, and $\alpha, \beta > 0$ are constants that depend on the complexity of the underlying function class and the learning algorithm. If $\beta$ is larger than the rate achievable by training from scratch, Eq. (1) implies that the pre-trained model is beneficial, provided that $m$ is sufficiently large such that $m^{-\alpha} \leq n^{-\beta}$.

However, these theoretical results have limited capacity to explain the data scaling laws of pre-trained models. Within the framework of the theoretical analyses in Eq. (1), the data scaling law for pre-trained models can be interpreted as the increase in source sample size $m$ (corresponding to the scale of the pre-trained model) leading to improved sample complexity with respect to the target sample size $n$ (the sample complexity of the downstream task). Yet, the sample complexity for $n$ is characterized by $\beta$ in Eq. (1), which remains invariant to the scale of the pre-trained model, $m$. Thus, the existing results showing Eq. (1) do not fully capture the data scaling laws of pre-trained models.

To address this gap, this paper aims to explore the following critical research question:

> What properties of a pre-trained model can ensure improved sample complexity for the target sample size as the source sample size increases?

Identifying such characteristics of pre-trained models can lead to a more nuanced understanding of their recent success, particularly the data scaling laws of pre-trained models, and may result in the development of more efficient methodologies for constructing pre-trained models.

**Contributions.** The primary contribution of this paper is to clarify the properties of pre-trained models that give rise to the data scaling laws observed in practice. Specifically, we show that under a certain assumption on the pre-trained model, the error rate satisfies

$$\mathbb{E}[E_Q(f_n)] \leq n^{-\beta+o(1)}, \tag{2}$$

for some constant $\beta > 0$, where $o(1)$ is a finite term that decreases with $m$ and vanishes as $m \to \infty$. The rate in Eq. (2) demonstrates that the sample complexity with respect to $n$ improves as the source sample size $m$ increases, since the exponent of $n$ decreases with $m$. Thus, the bound in Eq. (2) provides a theoretical explanation for the data scaling laws of pre-trained models.

To achieve the rate in Eq. (2), we introduce a novel concept called *caulkability*, which is conceptually illustrated in Fig. 1. Caulkability requires that the pre-trained model (blue boxes in Fig. 1) possesses a hierarchical structure, such that the underlying function (red boxes in Fig. 1) can be approximated by inserting another function, termed the *adapter model* (green boxes in Fig. 1), into this hierarchy. We demonstrate that if the complexity of the adapter model decreases as the source sample size $m$ increases (i.e., the depth of the adapter model shrinks along the horizontal axis representing $m$ in Fig. 1), it is possible to construct a learning algorithm, called *empirical caulking*, that achieves the error rate in Eq. (2). This result shows that reducing the complexity of the adapter model is a sufficient condition for attaining the rate in Eq. (2), thereby answering the research question above. This also highlights that parameter-efficient fine-tuning (PEFT) methods can be effective in learning with pre-trained models, as we can interpret leveraging low complexity adapter models as a form of PEFT.

We further provide empirical evidence that supports our theoretical results. In particular, our experiments

on fine-tuning CNNs and integrating vision capabilities into LLMs demonstrate that larger pre-trained models can be adapted to downstream tasks using shallower adapter models. These findings reinforce our theoretical claim that adapting larger pre-trained models leads to a reduction in the sample complexity required for downstream tasks.

All missing proofs are deferred to Appendix A.

## 2 RELATED WORK

**Domain Adaptation Theory.** Domain adaptation is a major subfield of transfer learning that addresses the scenario where the training (source) data and test (target) data come from different distributions even though the learning task is the same (Redko et al., 2020). Domain adaptation is, in essence, a transfer-learning approach that leverages abundant source-domain data together with a small amount of explicitly provided target-domain data (either labeled or unlabeled), and theoretical analyses are often conducted under this assumption. A rich theory has been developed to understand the generalization performance under domain shift. Seminal work by Ben-David et al. (2006, 2010) introduced generalization bounds for domain adaptation that relate the target error to three components, namely source domain error, domain discrepancy, and an irreducible term (known as the joint error). The general bound can be stated informally as: for any hypothesis $h$ in a given class, the target risk $R_T(h)$ is bounded by the source risk $R_S(h)$ plus a discrepancy between the domains and a constant term for labeling mismatch. This result implies that if two domains are very similar (small discrepancy) and also share an underlying label function, then a model will generalize well across domains, whereas large distribution gaps or conflicting label definitions will lead to a larger generalization error. Numerous theoretical results have refined these ideas. Researchers have proposed various discrepancy metrics for different settings, e.g., the symmetric difference $\mathcal{H}$-divergence for classification and regression (Ben-David et al., 2010), and extended the theory to multi-source adaptation and other scenarios (Mansour et al., 2008). Other work analyzes, e.g., maximum mean discrepancy (Redko et al., 2019) or the Wasserstein distance (Shen et al., 2018), with their own theoretical guarantees.

Apart from discrepancy-based theories, alternative approaches have also been explored. For instance, Kpotufe (2017); Ma et al. (2023); Feng et al. (2023) investigate upper bounds on the error of methods that account for distribution shift between the source and target domains using density ratios. In addition, Kpotufe and Martinet (2021); Pathak et al. (2022); Galbraith and Kpotufe (2023); Fujikawa et al. (2025) attempt to analyze error bounds using a similarity measure based on hyperspheres in a metric space. While these analyses appear to yield sound error bounds, several limitations can be pointed out. For example, in density-ratio–based methods, the analysis is often carried out under the assumption that the true density ratio is known; in practice, it is unknown, so the estimation error is not taken into account. Moreover, in both approaches, when the supports of the source and target data distributions are significantly mismatched, the similarity measure can diverge. Therefore, existing domain adaptation theory is insufficient as a tool for analyzing the error of pre-trained models; new tools need to be developed.

**Deep Learning Theory for Pre-trained Models.** There has been substantial theoretical progress in understanding the statistical performance of deep learning models (Schmidt-Hieber, 2020; Suzuki and Nitanda, 2021; Kohler and Langer, 2021; Suh et al., 2022; Nishikawa et al., 2025; Damian et al., 2022, 2023). This line of research has been further extended to analyze the statistical properties of pre-trained models, particularly in establishing the error rate in Eq. (1). For instance, Du et al. (2020) study few-shot learning problems within the linear feature class and linear outcome model, demonstrating the error rate of the form in Eq. (1). Bai et al. (2023b) investigate the statistical performance of transformers for in-context learning under the linear regression model, also obtaining the form in Eq. (1). Kim et al. (2024) analyze transformers for in-context learning in the Besov space, again showing the error rate of the form in Eq. (1). As discussed in the introduction, these results share the form in Eq. (1) and therefore do not account for the scaling laws observed in pre-trained models.

**Large Language Models.** The proposed concept of caulking is closely related to the recent advances in large language models (LLMs). In particular, recent training methods for multimodal language models (Li et al., 2023; Bai et al., 2023a; Liu et al., 2023; Dai et al., 2023; Gosthipaty et al., 2024; Liu et al., 2025) and parameter-efficient fine-tuning (PEFT) techniques (Houlsby et al., 2019; Guo et al., 2021; Hu et al., 2021) can be viewed as analogous to caulking. For instance, when constructing a vision-language model using a pre-trained visual feature extractor and an LLM, a typical strategy is to learn a mapping from visual features to textual features, referred to as an adapter model, using an image-text pair dataset. The final vision-language model is then formed by inserting the learned adapter model between the visual feature extractor and the LLM, which is analogous to the caulking process. Similarly, PEFT methods involve replacing or augmenting the

pre-trained model with a component of lower complexity, which is also analogous to caulking.

## 3 LEARNING WITH PRE-TRAINED MODEL VIA CAULKING

**Notations.** Let $\mathcal{E} \in \mathfrak{Z}$ be an event in a probability space $(\mathcal{Z}, \mathfrak{Z}, \nu)$. The complement of $\mathcal{E}$ is denoted by $\mathcal{E}^c$, and its probability by $\mathbb{P}_\nu\{\mathcal{E}\}$. For a random variable $X$ defined on $(\mathcal{Z}, \mathfrak{Z}, \nu)$ with values in a measurable space $(\mathcal{X}, \mathfrak{X})$, its expectation is denoted by $\mathbb{E}_\nu[X]$. For $p \in [1, \infty)$, the $L^p$-norm of a function $f : \mathcal{Z} \to \mathbb{R}$ is denoted by $\|f\|_{L^p(\mathcal{Z})} = (\int_{\mathcal{Z}} |f|^p d\lambda)^{1/p}$, where $\mathcal{Z} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ and $\lambda$ is the Lebesgue measure. For a measure $\nu$ on a measurable space $(\mathcal{Z}, \mathfrak{Z})$, the $L^p(\nu)$-norm of $f : \mathcal{Z} \to \mathbb{R}$ is defined as $\|f\|_{L^p(\nu)} = (\int |f|^p d\nu)^{1/p}$ for $p \in [1, \infty)$ and $\|f\|_{L^\infty(\nu)} = \inf\{b > 0 : \nu(\{z \in \mathcal{Z} : |f(z)| \le b\}) = 1\}$.

**Problem Setup.** We consider a nonparametric regression problem in the presence of a pre-trained model. Let $X \in \mathcal{X}$ and $Y \in \Omega \subset \mathbb{R}$ denote the covariates and outcome, respectively, and let $P$ and $Q$ denote the source and target distributions. Under the target distribution $Q$, assume that $X$ and $Y$ follow the nonparametric regression model:

$$Y = f^*(X) + \xi,$$

where $f^*$ is the regression function of interest and $\xi$ is a zero-mean noise term independent of $X$. Suppose the learner has access to a pre-trained model $f_{\text{pre}}$ constructed from a source sample of size $m$ drawn from $P$. The learner's objective is to estimate $f^*$ using a target sample of size $n$ from $Q$ in conjunction with the pre-trained model $f_{\text{pre}}$. Let $f_n$ denote the estimated regression function. Then, the accuracy of $f_n$ is evaluated via the $L^2(Q_X)$ distance from $f^*$, defined as

$$E_Q(f) = \|f - f^*\|^2_{L^2(Q_X)},$$

where $Q_X$ is the marginal distribution of $X$ under $Q$.

**Motivating Example.** Fig. 2 illustrates a conceptual example that demonstrates the effective use of a pre-trained model, using images inspired by the Office-Home dataset (Venkateswara et al., 2017). The figure consists of two manifolds, each representing the probability that an image is classified as Bed ($f^*$), based on the feature space for two domains: Real-world ($P$) and Clipart ($Q$). Points on the manifolds correspond to features, with images connected by dashed lines. The manifolds differ only by rotations and translations. A pre-trained model is constructed to capture the complex structure of the left manifold using real-world images (source sample). Subsequently, an accurate estimate of $f^*$ for the clipart images (target sample) on the left manifold can be obtained by aligning the right manifold with the left one through appropriate transformations. Learning the complex manifold structure from scratch requires a larger sample size than learning only the rotations and translations (i.e., linear transformations). Therefore, leveraging the pre-trained model enables high predictive accuracy with a small target sample, highlighting the potential effectiveness of pre-trained model utilization.

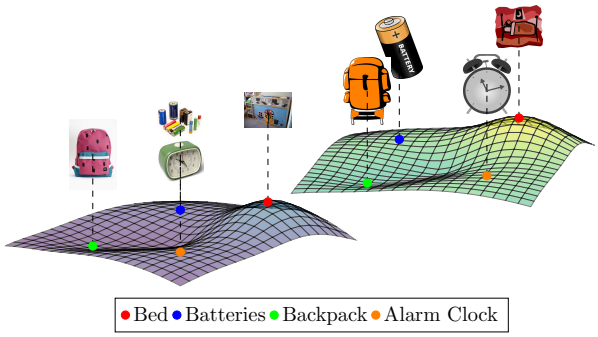


Figure 2: A motivating example illustrating the successful utilization of a pre-trained model.

**Caulking.** The scenario depicted in Fig. 2 can be described by a hierarchical structure, formalized by the concept of caulkability.

**Definition 1** (Caulkability). *Let $\mathcal{G}$ be a class of functions $g : \mathcal{Z}_i \to \mathcal{Z}_o$, where $\mathcal{Z}_i$ and $\mathcal{Z}_o$ are arbitrary domains. Given $\epsilon > 0$ and $\mathcal{G}$, a regressor $f^* : \mathcal{X} \to \mathbb{R}$ is $(\epsilon, \mathcal{G})$-caulkable by $(g_h, g_e)$, where $g_e : \mathcal{X} \to \mathcal{Z}_i$ and $g_h : \mathcal{Z}_o \to \mathbb{R}$, with respect to the $L^2(Q_X)$-norm if*

$$\inf_{g_a \in \mathcal{G}} \|f^* - g_h \circ g_a \circ g_e\|_{L^2(Q_X)} \le \epsilon. \tag{3}$$

To successfully leverage the pre-trained model, we assume that the ideal regressor $f^*$ is caulkable by the pre-trained model $f_{\text{pre}} = (g_h, g_e)$. In Def. 1, $g_e$ denotes a mapping from the image to features, corresponding to the dashed lines in Fig. 2; the manifold in Fig. 2 is parameterized by these features, where the height corresponds to the output of $g_h$ and the spatial position corresponds to its input; and $g_a$ denotes a domain-specific transformation, corresponding to the rotations and translations for the left and right manifolds in Fig. 2, respectively. Eq. (3) indicates that the ideal regressor $f^*$ can be (approximately) recovered by inserting a target-specific transformation $g_a$ into the pre-trained model $f_{\text{pre}}$. We refer to this property of the pre-trained model as caulkability, expressing the situation where the ideal regressor $f^*$ is recovered by filling the gap between $g_h$ and $g_e$. We refer to $g_h$ and $g_e$ as the *head model* and *feature extractor model*, respectively, and to $g_a$ as the *adapter model*.

**Empirical Caulking.** We propose a learning framework that leverages the pre-trained model, termed *empirical caulking*. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a target sample drawn from $Q$. Suppose that the ideal regressor $f^*$ is $(\epsilon_n, \mathcal{G})$-caulkable by the pre-trained model $f_{\mathrm{pre}} = (g_h, g_e)$, where $\epsilon_n$ is a constant possibly depending on the target sample size $n$, and $\mathcal{G}$ is a given class of functions. We design a learning algorithm for $f_n$ by inserting the fine-tuned adapter model $g_a$ into the pre-trained model $f_{\mathrm{pre}}$. Given a class $\mathcal{G}_n \subseteq \mathcal{G}$ of possible $g_a$, potentially depending on $n$, define $\mathcal{F}_n = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}_n\}$. $\mathcal{F}_n$ thus represents the set of functions obtainable by inserting an adapter model $g_a \in \mathcal{G}_n$ into the pre-trained model $f_{\mathrm{pre}}$. The estimated regressor $f_n$ is then defined as

$$f_n = \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

This learning algorithm is equivalent to the least squares estimator, but restricted to the specific function class $\mathcal{F}_n$.

**Remark 1.** *Our current formulation of empirical caulking covers the most parameter-efficient fine-tuning methods (PEFTs), whereas it cannot directly cover the case of LoRA (Hu et al., 2021). Nevertheless, LoRA can be interpreted as a method for inserting several layers into the pre-trained model. Specifically, letting $g_H \circ \cdots \circ g_1$ be an $H$-layer pre-trained model, consider the modified version $h_{H+1} \circ h_H \circ \cdots \circ h_1 \circ h_0$, where $h_i(x, y) = (g_i(x), y)$ for $i = 1, ..., H$, $h_0(x) = (x, x)$, and $h_{H+1}(x, y) = x$. Then, the composition of the $h_i$ is equivalent to the composition of the $g_i$. LoRA can be seen as inserting adapter layers $h_{1+1/2}, ..., h_{H+1/2}$ into this modified pre-trained model. Specifically, we can recover LoRA by setting $h_{i+1/2}(x, y) = (x + \Delta W y, x + \Delta W y)$, for $i = 1, ..., H$, where $\Delta W$ is low rank. Hence, extending our result to cover the case of inserting multiple adapter layers would cover LoRA, which is a crucial future direction of this work.*

## 4 ERROR ANALYSIS OF EMPIRICAL CAULKING

In this section, we provide a rigorous analysis of the error associated with empirical caulking, utilizing an appropriate complexity measure for the class $\mathcal{G}$. Throughout this section, we assume that $\mathcal{Z}_o$ is equipped with a norm $\|\cdot\|$, and we define the $L^\infty$-norm of an adapter model $g_a : \mathcal{Z}_i \to \mathcal{Z}_o$ as $\|g_a\|_{L^\infty} = \sup_{z \in \mathcal{Z}_i} \|g_a(z)\|$.

**Assumption.** We make the following assumptions throughout our analysis:

**Assumption 1.** *The following conditions hold:*

1. *The noise $\xi$ is sub-Gaussian with variance proxy $\sigma^2 < \infty$, i.e., $\mathbb{E}[\exp(\lambda \xi)] \leq \exp(\lambda^2 \sigma^2 / 2)$ for all $\lambda \in \mathbb{R}$.*
2. *$\Omega$ is bounded, that is, there exists a constant $\Delta_\Omega > 0$ such that $\sup_{x,y \in \Omega} |y - x| \leq \Delta_\Omega$.*

Asm. 1 is standard in the literature on nonparametric regression and statistical learning theory. Similar assumptions are made in prior works such as Schmidt-Hieber (2020) and Suzuki and Nitanda (2021), which analyze the theoretical properties of deep learning and nonparametric regression methods.

**Complexity of $\mathcal{G}$.** To characterize the error of empirical caulking, we adopt a complexity measure for $\mathcal{G}$ based on both the approximation error and the covering number. Given $\delta > 0$, a class $\mathcal{G}$, and a norm $\|\cdot\|$ defined on $\mathcal{G}$, the $\delta$-covering number $N(\delta, \mathcal{G}, \|\cdot\|)$ is the minimal number of balls of radius $\delta$ (with respect to $\|\cdot\|$) needed to cover $\mathcal{G}$. The following definition formalizes the notion of complexity.

**Definition 2** ($\beta$-complexity)**.** *Let $\beta > 0$. A class $\mathcal{G}$ of functions $g : \mathcal{Z}_i \to \mathcal{Z}_o$ is $\beta$-complex if, for each $J \in \mathbb{N}$, there exists a class $\mathcal{G}_J$ such that*

$$\ln(N(\delta, \mathcal{G}_J, \|\cdot\|_{L^\infty})) \leq cJ\mathrm{polylog}(J, 1/\delta),$$

*and*

$$\sup_{g^* \in \mathcal{G}} \inf_{g \in \mathcal{G}_J} \|g - g^*\|_{L^\infty} \leq c' J^{-\beta},$$

*for some constants $c, c' > 0$, where* polylog *denotes a polylogarithmic factor in its arguments.*

Def. 2 characterizes the complexity of $\mathcal{G}$ through the trade-off between approximation capability and the complexity of the classes $\mathcal{G}_J$. This concept is also used in the analysis of nonparametric regression error bounds for classes such as sparse ReLU networks, as in (Schmidt-Hieber, 2020; Suzuki and Nitanda, 2021; Chen et al., 2022). A concrete application of Def. 2 is provided in Section 4.1.

**Hölder Continuity.** Alongside $\beta$-complexity, we employ Hölder continuity to quantify the regularity of the pre-trained model $f_{\mathrm{pre}}$.

**Definition 3** (Hölder continuity)**.** *Let $g$ be a function from $\mathcal{X}$ to $\mathcal{Z}$, where $\mathcal{X}$ and $\mathcal{Z}$ are equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Z$, respectively. For $\alpha \in (0, 1]$, $g$ is said to be $\alpha$-Hölder continuous if there exists a constant $C > 0$ such that*

$$\|g(x) - g(y)\|_Z \leq C \|x - y\|_X^\alpha,$$

*for all $x, y \in \mathcal{X}$.*

**Error Bound.** We now present an error bound for empirical caulking when $\mathcal{G}$ is $\beta$-complex and the ideal function $f^*$ is $(\epsilon_n, \mathcal{G})$-caulkable by a pre-trained model $f_{\text{pre}}$.

**Theorem 1.** *Let $\alpha \in (0,1]$ and $\beta > 0$. Suppose that $f^*$ is $(n^{-\frac{2\alpha\beta}{2\alpha\beta+1}}, \mathcal{G})$-caulkable by $f_{\text{pre}} = (g_h, g_e)$ for some $\beta$-complex class $\mathcal{G}$, and $g_h$ is $\alpha$-Hölder continuous. Let $f_n$ denote the estimated regressor obtained by empirical caulking with $\mathcal{G}_n = \mathcal{G}_J$, where $\mathcal{G}_J$ is as in Def. 2 and $J = \lceil n^{1/(2\alpha\beta+1)} \rceil$. Under Asm. 1, we have*

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C n^{-\frac{2\alpha\beta}{2\alpha\beta+1}} \text{polylog}(n),$$

*for some constant $C > 0$.*

Building on Thm. 1, we now examine the effectiveness of a pre-training algorithm. Consider a pre-training algorithm that yields a pre-trained model $f_{\text{pre},m}$ such that $f^*$ is caulkable with a sequence of adapter model classes $\mathcal{G}_m$ whose complexity decreases as $m$ increases. Specifically, suppose that for each $m$, the pre-trained model $f_{\text{pre},m} = (g_{h,m}, g_{e,m})$ enables $f^*$ to be $(\epsilon_n, \mathcal{G}_m)$-caulkable with a $(\beta - \gamma_m)$-complex class $\mathcal{G}_m$, where $\gamma_m$ is a bounded term that vanishes as $m$ grows. For such a pre-training algorithm, we obtain the following corollary.

**Corollary 1.** *Let $\alpha \in (0,1]$ and $\beta > 0$. Let $\gamma_m < \beta$ be a bounded sequence depending on $m$. Suppose that $f^*$ is $(n^{-\frac{2\alpha\beta}{2\alpha\beta+1}+\gamma_m}, \mathcal{G}_m)$-caulkable by $f_{\text{pre},m} = (g_{h,m}, g_{e,m})$ for some $(\beta - \gamma_m)$-complex class $\mathcal{G}_m$, and $g_{h,m}$ is $\alpha$-Hölder continuous. Let $f_n$ denote the estimated regressor obtained by empirical caulking with $\mathcal{G}_n = \mathcal{G}_J$, where $\mathcal{G}_J$ is as in Def. 2 and $J = \lceil n^{1/(2\alpha(\beta-\gamma_m)+1)} \rceil$. Under Asm. 1, we have*

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C n^{-\frac{2\alpha\beta}{2\alpha\beta+1}\left(1-\frac{\gamma_m}{\beta(\alpha(\beta-\gamma_m)+1)}\right)} \text{polylog}(n),$$

*for some constant $C > 0$.*

Cor. 1 indicates that empirical caulking achieves a rate of $n^{-\frac{2\alpha\beta}{2\alpha\beta+1}+o(1)}$ up to a logarithmic factor, where $o(1)$ is a bounded term that vanishes as $m$ increases, since $\frac{\gamma_m}{\beta(\alpha(\beta-\gamma_m)+1)}$ decreases as $\gamma_m$ decreases. Therefore, Cor. 1 demonstrates that empirical caulking can improve target sample complexity as the source sample size increases, provided that $f^*$ is caulkable with a sequence of classes $\mathcal{G}_m$ of decreasing complexity. This result offers partial theoretical support for the scaling law of sample complexity in downstream tasks observed in empirical studies.

### 4.1 Example: Compositional Space

We now present a concrete example illustrating the application of Thm. 1, highlighting the advantages of caulking in scenarios where the regression function $f^*$ is a composition of smooth and sparse functions (Schmidt-Hieber, 2020; Suzuki and Nitanda, 2021; Kohler and Langer, 2021). In particular, these works consider the case where $f^* \in \mathcal{F}_H = \mathcal{F}_H \circ \cdots \circ \mathcal{F}_1$, with

$$\mathcal{F}_H \circ ... \circ \mathcal{F}_1 = \{f_H \circ ... \circ f_1 : f_i \in \mathcal{F}_i, i = 1, ..., H\}, \quad (4)$$

where each $\mathcal{F}_i$ is a class of smooth and sparse functions, such as the sparse Hölder class (Schmidt-Hieber, 2020; Kohler and Langer, 2021) or the anisotropic Besov class (Suzuki and Nitanda, 2021). For clarity, we focus on the setting of Schmidt-Hieber (2020), where $\mathcal{F}_i$ consists of $\beta_i$-Hölder smooth functions, each depending on $t_i$ out of $d_i$ possible variables. We first review the relevant existing results before demonstrating how Thm. 1 applies in this context.

**Error bound without a pre-trained model.** Schmidt-Hieber (2020) established that the estimator $f_n$, obtained via empirical risk minimization over the class of sparse ReLU networks, satisfies the following error bound:

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=1,...,H} n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}} \text{polylog}(n), \quad (5)$$

for some constant $C > 0$, where $\alpha_i = \prod_{\ell=i+1}^{H} \max\{1, \beta_\ell\}$. Here, $\alpha_i$ reflects the (worst-case) Hölder smoothness among the functions in the composition $\mathcal{F}_H \circ \cdots \circ \mathcal{F}_i$. Classical nonparametric regression theory tells us that the minimax error rate for the class $\mathcal{F}_i$ is $n^{-\frac{2\beta_i}{2\beta_i+t_i}}$. The rate in Eq. (5) is thus the slowest among the component classes $\mathcal{F}_i$, with each rate further degraded by the factor $\alpha_i$ due to the compositional structure.

**Error bound with a pre-trained model (caulking).** Now consider the scenario where a pre-trained model $f_{\text{pre}}$ is available, and $f^*$ is $(1/n, \mathcal{F}_{i_h} \circ \cdots \circ \mathcal{F}_{i_e})$-caulkable by $f_{\text{pre}} = (g_{h,m}, g_{e,m})$ for some $1 < i_e < i_h < H$. The complexity of the class $\mathcal{F}_{i_h} \circ \cdots \circ \mathcal{F}_{i_e}$ can be effectively controlled.

**Theorem 2** (Schmidt-Hieber (2020))**.** *$\mathcal{F}_{i_h} \circ ... \circ \mathcal{F}_{i_e}$ is $\min_{i=i_h,...,i_e} \frac{\alpha_i\beta_i}{\alpha_{i_h} t_i}$-complex in terms of Def. 2.*

By applying Thm. 1 together with Thm. 2, we obtain the following error bound for empirical caulking:

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=i_e,...,i_h} n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}} \text{polylog}(n), \quad (6)$$

for some constant $C > 0$. A comparison of Eq. (5) and Eq. (6) yields several important insights:

- Both bounds take the maximum over $n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}}$, but with a pre-trained model, the range of the index

$i$ is restricted to a narrower subset. This demonstrates the improved error rate that can be achieved by leveraging a pre-trained model through empirical caulking.

- The result in Eq. (6) can be interpreted as allowing the learner to bypass the complexities associated with learning the feature extractor part $(\mathcal{F}_{i_e-1}, ..., \mathcal{F}_1)$ and the head part $(\mathcal{F}_H, ..., \mathcal{F}_{i_h+1})$ by utilizing the pre-trained model. Moreover, the greater the complexity of the pre-trained model (increasing $i_e$ and decreasing $i_h$), the greater the improvement in the error rate. This partially supports the scaling law of the sample complexity of the downstream task observed in empirical studies.
- The influence of the feature extractor classes $\mathcal{F}_{i_e-1}, ..., \mathcal{F}_1$ is entirely eliminated from the error bound. In fact, the error rate in Eq. (6) remains unchanged even if these classes are replaced by arbitrary alternatives, as long as caulkability is maintained.
- The dependence on the head classes $\mathcal{F}_H, ..., \mathcal{F}_{i_h+1}$ persists, but only through the Hölder smoothness parameter $\alpha_i$.

More details for this example can be found in Appendix B.

## 5 EXTENSION TO CLASSIFICATION

In this section, we present the extension of our theoretical results to binary classification problems.

**Setup.** Let $X \in \mathcal{X}$ and $Y \in \{0, 1\}$ denote the covariate and label, respectively. In place of the regression function used in the regression setting, we consider the score function $f^*(x) = \mathbb{P}_Q\{Y = 1 | X = x\}$. Accordingly, we assume that the learner has access to a pre-trained model $f_{\mathrm{pre}}$ for the score function $f^*$, constructed from a source sample of size $m$ drawn from $P$. Given the pre-trained model $f_{\mathrm{pre}}$ and a target sample of size $n$ drawn from $Q$, the learner's objective is to construct a classifier $h : \mathcal{X} \to \{0, 1\}$ that minimizes the classification error on the target distribution, i.e., $\mathbb{E}_Q[\mathbb{1}\{h(X) \neq Y\}]$. Let $h_n$ denote the resulting estimated classifier.

The accuracy of a classifier $h$ is evaluated by its excess error. Let $h^*$ denote the Bayes optimal classifier that minimizes the classification error, which is given by $h^*(x) = \mathbb{1}\{f^*(x) > 1/2\}$. The excess error of a classifier $h$ is then defined as

$$E_Q(h) = \mathbb{E}_Q[\mathbb{1}\{h(X) \neq Y\}] - \mathbb{E}_Q[\mathbb{1}\{h^*(X) \neq Y\}].$$

**Algorithm.** We describe the plug-in estimator based on the estimation of the score function $f^*$. The procedure first estimates the score function $f^*$, denoted by $f_n$, and then constructs the estimated classifier as $h_n(x) = \mathbb{1}\{f_n(x) > 1/2\}$. The classification error of the plug-in estimator can be controlled via the regression error bound. Specifically, we have

$$\mathbb{E}_{Q^n}[E_Q(h_n)] \leq 2\sqrt{\mathbb{E}_{Q^n}\left[\|f^* - f_n\|^2_{L^2(Q_X)}\right]}.$$

Therefore, we employ empirical caulking with the squared error loss for estimating $f_n$. Suppose that the ideal score function $f^*$ is $(\epsilon_n, \mathcal{G})$-caulkable by the pre-trained model $f_{\mathrm{pre}} = (g_h, g_e)$. Given a class $\mathcal{G}_n \subseteq \mathcal{G}$, and letting $\mathcal{F}_n = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}_n\}$, the estimated score function is given by

$$f_n = \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

**Error analysis.** We establish an analog of Thm. 1 for the classification setting.

**Theorem 3.** *Let $\alpha \in (0, 1]$ and $\beta > 0$. Suppose that $f^*$ is $(n^{-\frac{2\alpha\beta}{2\alpha\beta+1}}, \mathcal{G})$-caulkable by $f_{\mathrm{pre}} = (g_h, g_e)$ for some $\beta$-complex class $\mathcal{G}$, and $g_h$ is $\alpha$-Hölder continuous. Let $f_n$ denote the estimated score function obtained by empirical caulking with $\mathcal{G}_n = \mathcal{G}_J$, where $\mathcal{G}_J$ is as in Def. 2 and $J = \lceil n^{1/(2\alpha\beta+1)} \rceil$. Let $h_n(x) = \mathbb{1}\{f_n(x) > 1/2\}$. Under Asm. 1, we have*

$$\mathbb{E}_{Q^n}[E_Q(h_n)] \leq C n^{-\frac{\alpha\beta}{2\alpha\beta+1}} \operatorname{polylog}(n),$$

*for some constant $C > 0$.*

## 6 PROOF SKETCH

In this section, we provide an outline of the proof for the main result, Thm. 1. Complete proofs of Thm. 1 and other omitted results are given in the supplemental material. The proof of Thm. 1 relies on the following error bound, which depends on both the approximation capability and the covering number of the class $\mathcal{F}_n$.

**Theorem 4** (Schmidt-Hieber (2020); Hayakawa and Suzuki (2020))**.** *Under the same assumptions as Thm. 1, we have*

$$\mathbb{E}_{Q^n}\left[\|f_n - f^*\|^2_{L^2(Q_X)}\right] \leq C\left(\inf_{f \in \mathcal{F}_n} \|f - f^*\|^2_{L^2(Q_X)} + \frac{\ln\left(N\left(n^{-1}, \mathcal{F}_n, \|\cdot\|_{L^\infty(Q_X)}\right)\right)}{n}\right),$$

*for some constant $C > 0$ depending on $\sigma_\xi$ and $\Delta_\Omega$ in Asm. 1.*

To utilize Thm. 4, it is necessary to control both the approximation capability and the covering number of $\mathcal{F}_n$ in terms of the $\beta$-complexity of the class $\mathcal{G}_N$. For this purpose, we establish the following two propositions.

**Proposition 1.** *Given a class of functions $\mathcal{G} = \{g : \mathcal{Z}_i \to \mathcal{Z}_o\}$, define $\mathcal{F} = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}\}$ for some functions $g_h : \mathcal{Z}_o \to \mathbb{R}$ and $g_e : \mathcal{X} \to \mathcal{Z}_i$. Assume that $g_h$ is $\alpha$-Hölder continuous with respect to a norm $\|\cdot\|$ on $\mathcal{Z}_o$ for some $\alpha \in (0, 1]$. Then, for any measure $\nu$ on $\mathcal{X}$ and any $\delta > 0$, we have*

$$N(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu)}) \leq N\left(\left(\frac{\delta}{C_\alpha}\right)^{1/\alpha}, \mathcal{G}, \|\cdot\|_{L^\infty}\right),$$

*where $C_\alpha$ is a constant for the Hölder continuity of $g_h$.*

**Proposition 2.** *Given a class of functions $\mathcal{G} = \{g : \mathcal{Z}_i \to \mathcal{Z}_o\}$, define $\mathcal{F} = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}\}$ for some functions $g_h : \mathcal{Z}_o \to \mathbb{R}$ and $g_e : \mathcal{X} \to \mathcal{Z}_i$. Assume that $g_h$ is $\alpha$-Hölder continuous with respect to a norm $\|\cdot\|$ on $\mathcal{Z}_o$ for some $\alpha \in (0, 1]$. For any $f^* = g_h \circ g_a^* \circ g_e$ with $g_a^*$ possibly not in $\mathcal{G}$, we have*

$$\inf_{f \in \mathcal{F}} \|f - f^*\|_{L^2(Q_X)} \leq C_\alpha \inf_{g_a \in \mathcal{G}} \|g_a - g_a^*\|_{L^\infty}^\alpha,$$

*where $C_\alpha$ is a constant for the Hölder continuity of $g_h$.*

By applying Prop. 1 and Prop. 2 to the first and second terms of Thm. 4, and using the $\beta$-complexity of the class $\mathcal{G}_J$, we obtain $\mathbb{E}_{Q^n}\left[\|f_n - f^*\|_{L^2(Q_X)}^2\right] = O(J^{-2\alpha\beta} + J/n)$. Choosing $J$ as in the statement yields the desired result.

# 7 EMPIRICAL EVALUATION OF CAULKING

We conduct two types of experiments to support our theoretical claims[1].

## 7.1 Fine-tuning of CNNs

We first fine-tune ResNet-50 (26M parameters, He et al. (2016)) and Wide ResNet-50-2 (68M parameters, Zagoruyko and Komodakis (2016)) with adapters on the clipart domain of the Office-Home dataset (Venkateswara et al., 2017) after pre-training on other domains. Fig. 3 shows the relationship between the error rates and the depth of adapters, where the adapters consist of linear layers forming MLPs. Due to the high similarity of the model architectures between ResNet and Wide ResNet, the results can suggest that the larger model achieves higher performance with a single-layer adapter, whereas the smaller model requires a more complex adapter. These findings are consistent with our theoretical results, which suggest that a larger pre-trained model can be effectively adapted with a simpler adapter.

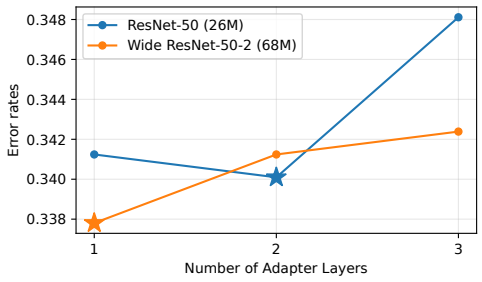


Figure 3: The relationship between the depth of adapters and the error rate on the target domain. Minimum error rates for each model are marked by $\star$.

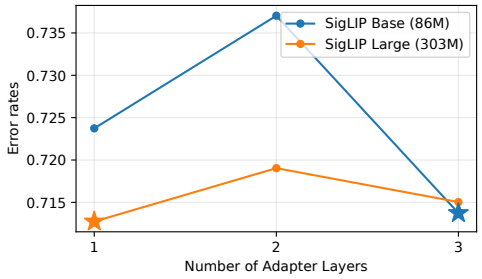


Figure 4: The relationship between the depth of adapters and the error rate on the MMStar dataset (Chen et al., 2024).

## 7.2 Integrating Vision Capabilities to LLMs

We then incorporate vision capabilities into a pre-trained language model by integrating pre-trained vision encoders and adapters. Specifically, we use a Llama-3-style Transformer model (135M parameters, Allal et al. (2024)) together with SigLIP visual feature extractors of varying sizes (86M and 303M parameters, Tschannen et al. (2025)) and train them as Vision Language Models following the training recipe of Gosthipaty et al. (2024) with a slight modification. Fig. 4 illustrates the relationship between the adapter depth and error rates. Once again, the results are consistent with our theoretical findings.

Further experimental details can be found in Appendix C.

# 8 CONCLUSION

This paper presents a theoretical analysis of the advantages of utilizing pre-trained models for downstream tasks, with particular emphasis on the scaling laws

[1]Code available at https://github.com/moskomule/calking_aistats26

associated with pre-trained models. These scaling laws indicate that larger pre-trained models can significantly reduce the sample complexity required for downstream tasks. We rigorously explain this phenomenon through the novel concept of *caulkability*. Our experimental results further corroborate our theoretical findings: in particular, larger pre-trained models can be effectively adapted to downstream tasks using simpler adapters. These insights clarify the benefits of employing larger pre-trained models for downstream applications.

**Open question.** The primary focus of this work is to identify the properties of pre-trained models that underlie the scaling laws observed in practice, and we show that caulkability by an adapter model class of decreasing complexity is a sufficient condition for these scaling laws. However, we do not propose a concrete learning algorithm for constructing a pre-trained model that exhibits the caulkability property. Designing such a learning algorithm remains an important and intriguing open question.

### Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP23K13011 to K.F., JP23K28146 and JP24K20836 to K.M, and JST BOOST Grant Number JPMJBY24G2 to R.H.

### References

Allal, L. B., Lozhkov, A., Bakouch, E., von Werra, L., and Wolf, T. (2024). Smollm-blazingly fast and remarkably powerful. *Hugging Face Blog*.

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. (2023). Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023a). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. (2023b). Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Advances in Neural Information Processing Systems*, volume 36, pages 57125–57211.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1):151–175.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., and Zhao, F. (2024). Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chen, M., Jiang, H., Liao, W., and Zhao, T. (2022). Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. (2023). InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267.

Damian, A., Lee, J., and Soltanolkotabi, M. (2022). Neural Networks can Learn Representations with Gradient Descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 5413–5452. PMLR.

Damian, A., Nichani, E., Ge, R., and Lee, J. D. (2023). Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models. *Advances in Neural Information Processing Systems*, 36:752–784.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2020). Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representations*.

Feng, X., He, X., Wang, C., Wang, C., and Zhang, J. (2023). Towards a unified analysis of kernel-based methods under covariate shift. *Advances in Neural Information Processing Systems*, 36:73839–73851.

Fujikawa, M., Akimoto, Y., Sakuma, J., and Fukuchi, K. (2025). Harnessing the Power of Vicinity-Informed Analysis for Classification under Covariate Shift. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 226–234. PMLR.

Galbraith, N. R. and Kpotufe, S. (2023). Classification tree pruning under covariate shift. *IEEE Transactions on Information Theory*, 70(1):456–481.

Gosthipaty, A. R., Wiedmann, L., Marafioti, A., Paniego, S., Noyan, M., Cuenca, P., and Srivastav, V. (2024). nanoVLM: The simplest repository to train your VLM in pure PyTorch. *Hugging Face Blog*.

Guo, D., Rush, A., and Kim, Y. (2021). Parameter-Efficient Transfer Learning with Diff Pruning. In Zong, C., Xia, F., Li, W., and Navigli, R., editors,

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4884–4896, Online. Association for Computational Linguistics.

Hayakawa, S. and Suzuki, T. (2020). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361.

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

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. (2020). Scaling Laws for Autoregressive Generative Modeling.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q. D., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models.

Kim, J., Nakamaki, T., and Suzuki, T. (2024). Transformers are Minimax Optimal Nonparametric In-Context Learners. In *Advances in Neural Information Processing Systems*, volume 37, pages 106667–106713.

Kohler, M. and Langer, S. (2021). On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates. *The Annals of Statistics*, 49(4):2231–2249.

Kpotufe, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320–1328. PMLR.

Kpotufe, S. and Martinet, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al. (2025). Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134.

Ma, C., Pathak, R., and Wainwright, M. J. (2023). Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2008). Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21.

Mikami, H., Fukumizu, K., Murai, S., Suzuki, S., Kikuchi, Y., Suzuki, T., Maeda, S.-i., and Hayashi, K. (2023). A Scaling Law for Syn2real Transfer: How Much Is Your Pre-training Effective? In Amini, M.-R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., and Tsoumakas, G., editors, *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*, pages 477–492, Cham. Springer Nature Switzerland.

Nishikawa, N., Song, Y., Oko, K., Wu, D., and Suzuki, T. (2025). Nonlinear transformers can perform inference-time feature learning. In *Forty-Second International Conference on Machine Learning*.

Pathak, R., Ma, C., and Wainwright, M. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR.

Pyzer-Knapp, E. O., Manica, M., Staar, P., Morin, L., Ruch, P., Laino, T., Smith, J. R., and Curioni, A. (2025). Foundation models for materials discovery–current state and future directions. *Npj Computational Materials*, 11(1):61.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Suh, N., Zhou, T.-Y., and Huo, X. (2022). Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks. In *The Eleventh International Conference on Learning Representations*.

Suzuki, T. and Nitanda, A. (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, volume 34, pages 3609–3621. Curran Associates, Inc.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al. (2025). Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep Hashing Network for Unsupervised Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394.

Wang, J., Wang, K., Yu, Y., Lu, Y., Xiao, W., Sun, Z., Liu, F., Zou, Z., Gao, Y., Yang, L., et al. (2025). Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, 31(2):609–617.

Wen, Y., Chen, L., Deng, Y., and Zhou, C. (2021). Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, 78:103145.

Wu, Y. and He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Xia, J., Zhu, Y., Du, Y., and Li, S. Z. (2022). A systematic survey of chemical pre-trained models. *arXiv preprint arXiv:2210.16484*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association.

## 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] See Section 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Sections 4 and 5 for our theoretical analyses of the empirical caulking algorithms.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] See the supplemental.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Asm. 1 and Defs. 1 to 3.
   (b) Complete proofs of all theoretical results. [Yes] See the supplemental.
   (c) Clear explanations of any assumptions. [Yes] See the discussions following Asm. 1 and Defs. 1 to 3.
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] See the supplemental.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See the supplemental.
   (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] See the supplemental.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See the supplemental.
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] See Section 7.
   (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]

# A MISSING PROOFS

## A.1 Proof of Thm. 1

As shown in Section 6, the proof of Thm. 1 follows a similar approach to that of Schmidt-Hieber (2020); Hayakawa and Suzuki (2020), utilizing Thm. 4. For clarity, we present here a more rigorous statement of Thm. 4:

**Theorem 5.** *Let $\sigma > 0$ and $F > 0$ be constants. Let $X$, $Y$, and $\xi$ be random variables defined on a probability space $(\mathcal{Z}, \mathfrak{Z}, \nu)$, and let $\nu_X$ be the marginal distribution of $X$. Given a measurable function $f^* : \mathcal{X} \to \mathbb{R}$ such that $\|f^*\|_{L^\infty(\nu_X)} \leq F$, suppose that these variables follow the regression model $Y = f^*(X) + \xi$. Suppose also that $\xi$ is independent of $X$ and sub-Gaussian with variance-proxy $\sigma^2$, i.e., $\mathbb{E}_\nu\left[e^{\lambda\xi}\right] \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$. Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $\sup_{f\in\mathcal{F}}\|f\|_{L^\infty(\nu_X)} \leq F$. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be $n$ i.i.d. copies of $(X, Y)$, and define $f_n \in \mathcal{F}$ as a function that attains the following infimum:*

$$\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

*Then, for any $\delta > 0$ satisfying $\ln(N(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu_X)})) \geq 1$, we have*

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \leq C\left(\inf_{f\in\mathcal{F}}\|f - f^*\|^2_{L^2(\nu_X)} + \frac{(F^2+\sigma^2)\ln\left(N\left(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu)}\right)\right)}{n} + (F+\sigma)\delta\right),$$

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

The proof of Thm. 5 is given in Appendix A.2. Thm. 5 characterizes the error bound of the least-squares estimator $f_n$ over the class $\mathcal{F}$ in terms of the approximation error of $f^*$ by $\mathcal{F}$ and the covering number of $\mathcal{F}$. To control these two terms, we utilize Prop. 1 and Prop. 2. Recall the statements of Prop. 1 and Prop. 2:

**Proposition 3.** *Given a class of functions $\mathcal{G} = \{g : \mathcal{Z}_i \to \mathcal{Z}_o\}$, define $\mathcal{F} = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}\}$ for some functions $g_h : \mathcal{Z}_o \to \mathbb{R}$ and $g_e : \mathcal{X} \to \mathcal{Z}_i$. Assume that $g_h$ is $\alpha$-Hölder continuous with respect to a norm $\|\cdot\|$ on $\mathcal{Z}_o$ for some $\alpha \in (0, 1]$. Then, for any measure $\nu$ on $\mathcal{X}$ and any $\delta > 0$, we have*

$$N(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu)}) \leq N\left(\left(\frac{\delta}{C_\alpha}\right)^{1/\alpha}, \mathcal{G}, \|\cdot\|_{L^\infty}\right),$$

*where $C_\alpha$ is a constant for the Hölder continuity of $g_h$.*

**Proposition 4.** *Given a class of functions $\mathcal{G} = \{g : \mathcal{Z}_i \to \mathcal{Z}_o\}$, define $\mathcal{F} = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}\}$ for some functions $g_h : \mathcal{Z}_o \to \mathbb{R}$ and $g_e : \mathcal{X} \to \mathcal{Z}_i$. Assume that $g_h$ is $\alpha$-Hölder continuous with respect to a norm $\|\cdot\|$ on $\mathcal{Z}_o$ for some $\alpha \in (0, 1]$. For any $f^* = g_h \circ g_a^* \circ g_e$ with $g_a^*$ possibly not in $\mathcal{G}$, we have*

$$\inf_{f\in\mathcal{F}}\|f - f^*\|_{L^2(Q_X)} \leq C_\alpha \inf_{g_a\in\mathcal{G}}\|g_a - g_a^*\|^\alpha_{L^\infty},$$

*where $C_\alpha$ is a constant for the Hölder continuity of $g_h$.*

The proofs of Prop. 3 and Prop. 4 are given in Appendix A.3 and Appendix A.4, respectively. Building upon Thm. 5, Prop. 3, and Prop. 4, we can prove Thm. 1:

*Proof of Thm. 1.* We apply Thm. 5 with $\mathcal{F} = \mathcal{F}_n$, where $\mathcal{F}_n = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}_n\}$ for a sequence of classes $\mathcal{G}_n$ in Def. 2 corresponding to a $\beta$-complex class $\mathcal{G}$. With $\delta = 1/n$, we have

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \leq C\left(\inf_{f\in\mathcal{F}_n}\|f - f^*\|^2_{L^2(\nu_X)} + \frac{(F^2+\sigma^2)\ln\left(N\left(n^{-1}, \mathcal{F}_n, \|\cdot\|_{L^\infty(\nu)}\right)\right)}{n} + \frac{(F+\sigma)}{n}\right).$$

Since $\mathcal{G}_n$ is a sequence of classes in Def. 2 corresponding to a $\beta$-complex class $\mathcal{G}$, by Props. 3 and 4, we have

$$N(n^{-1}, \mathcal{F}_n, \|\cdot\|_{L^\infty(\nu)}) \leq N\left(\left(\frac{n^{-1}}{C_\alpha}\right)^{1/\alpha}, \mathcal{G}_n, \|\cdot\|_{L^\infty}\right) \leq cJ\mathrm{polylog}(J, (C_\alpha n)^{-1/\alpha}),$$

and

$$\inf_{f\in\mathcal{F}_n}\|f-f^*\|^2_{L^2(\nu_X)} \le C_\alpha \inf_{g_a\in\mathcal{G}_n}\|g_a-g_a^*\|^\alpha_{L^\infty} \le c' J^{-2\alpha\beta},$$

where $J$ is defined in the statement. Hence, we have

$$\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|^2_{L^2(\nu_X)}\right] \le C\left(J^{-2\alpha\beta}+\frac{J\text{polylog}(J,(C_\alpha n)^{-1/\alpha})}{n}+\frac{1}{n}\right). \tag{7}$$

Eq. (7) is minimized with $J$ defined in the statement, which yields the desired claim. □

### A.2 Proof of Thm. 5

We begin by introducing notation that will be used throughout the proof for clarity and conciseness. Let $\xi_1,...,\xi_n$ be i.i.d. copies of $\xi$. For any function $f:\mathcal{X}\to\mathbb{R}$, define the empirical $L^2$ norm of $f$ and noises $\xi_i$ as

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f^2(X_i), \quad \|\xi\|_n^2 := \frac{1}{n}\sum_{i=1}^n \xi_i^2,$$

and the empirical inner product with the noise variables as

$$\langle f,\xi\rangle_n := \frac{1}{n}\sum_{i=1}^n \xi_i f(X_i).$$

For the notational convenience, we use $\mathbb{V}_\nu[X]$ to denote the variance of a random variable $X$ defined on a probability space $(\mathcal{Z},\mathfrak{Z},\nu)$.

The proof consists of two steps. In the first step, we bound the expected error of $f_n$ by its empirical error. In the second step, we give a bound on the empirical error of $f_n$. Specifically, we show the following two lemmas.

**Lemma 1.** *Under the same assumptions as in Thm. 5, we have*

$$\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|^2_{L^2(\nu_X)}\right] \le \inf_{\zeta>0}\left((1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|_n^2\right]+\left(2+3\zeta+\frac{1}{2\zeta}\right)\frac{512F^2\ln(N_\delta)}{n}+\left(1+\zeta+\frac{1}{4\zeta}\right)\frac{9F^2}{n}+4(2+3\zeta)F\delta\right).$$

*where* $N_\delta = N(\delta,\mathcal{F},\|\cdot\|_{L^\infty(\nu_X)})$.

**Lemma 2.** *Under the same assumptions as in Thm. 5, we have*

$$\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|_n^2\right] \le \inf_{\zeta>0}\left((1+\zeta)\inf_{f\in\mathcal{F}}\|f-f^*\|^2_{L^2(\nu_X)}+\left(1+\frac{\zeta}{2}+\frac{1}{2\zeta}\right)\frac{8\sigma^2\ln(3N_\delta)}{n}+\left((1+\zeta)\sqrt{2\pi}\sigma+4F\zeta\right)\delta\right),$$

*where* $N_\delta = N(\delta,\mathcal{F},\|\cdot\|_{L^\infty(\nu_X)})$.

Combining Lem. 1 and Lem. 2 immediately yields Thm. 5.

*Proof of Thm. 5.* Combining Lem. 1 and Lem. 2 and taking $\zeta$ in these lemmas as universal constants, we obtain

$$\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|^2_{L^2(\nu_X)}\right] \le C\left(\inf_{f\in\mathcal{F}}\|f-f^*\|^2_{L^2(\nu_X)}+\frac{(\sigma^2+F^2)\ln(N_\delta)}{n}+(\sigma+F)\delta\right),$$

for some universal constant $C>0$. □

In the following subsections, we present the proofs of Lem. 1 and Lem. 2. The bounds in both lemmas are determined by the covering number of the class $\mathcal{F}$. For notational convenience, let $f_k$ denote the centers of balls in a $\delta$-covering of $\mathcal{F}_n$ with respect to the norm $\|\cdot\|_{L^\infty(\nu_X)}$. Define $g_k = f_k - f^*$, and let $K = \arg\min_k \|f_n - f_k\|_{L^\infty(\nu_X)}$.

### A.2.1 Proof of Lem. 1

*Proof of Lem. 1.* Let $\tau > 0$ be a constant to be specified later. We have

$$\begin{aligned}
&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] \\
=&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2 - \|f_n - f^*\|_n^2\right] \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 8F\delta + \mathbb{E}_{\nu^n}\left[\|g_K\|_{L^2(\nu_X)}^2 - \|g_K\|_n^2\right] \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 8F\delta + \mathbb{E}_{\nu^n}\left[\max\left\{\tau^2, \|g_K\|_{L^2(\nu_X)}^2\right\}\right]^{1/2} \mathbb{E}_{\nu^n}\left[\frac{\left(\|g_K\|_{L^2(\nu_X)}^2 - \|g_K\|_n^2\right)^2}{\max\left\{\tau^2, \|g_K\|_{L^2(\nu_X)}^2\right\}}\right]^{1/2},
\end{aligned} \tag{8}$$

where the third inequality follows from the Cauchy-Schwarz inequality, and in the second inequality we use the following:

$$\begin{aligned}
&\|f_n - f^*\|_{L^2(\nu_X)}^2 - \|f_n - f^*\|_n^2 \\
=&\|g_K\|_{L^2(\nu_X)}^2 + \mathbb{E}_\nu[(f_K(X) - f_n(X))(f_K(X) + f_n(X) - 2f^*(X))] \\
&\qquad - \|g_K\|_n^2 - \frac{1}{n}\sum_{i=1}^n (f_K(X_i) - f_n(X_i))(f_K(X_i) + f_n(X_i) - 2f^*(X_i)) \\
\leq&\|g_K\|^2 - \|g_K\|_n^2 + 8F\delta,
\end{aligned}$$

almost surely.

For the last term in Eq. (8), we have

$$\begin{aligned}
\mathbb{E}_{\nu^n}\left[\max\left\{\tau^2, \|g_K\|_{L^2(\nu_X)}^2\right\}\right] \leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] + \mathbb{E}_{\nu^n}\left[\|g_K\|_{L^2(\nu_X)}^2 - \|f_n - f^*\|_{L^2(\nu_X)}^2\right] + \tau^2 \\
=&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] + \mathbb{E}_{\nu^n}[\mathbb{E}_\nu[(f_K(X) - f_n(X))(f_K + f_n - 2f^*)]] + \tau^2 \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] + 4F\delta + \tau^2.
\end{aligned} \tag{9}$$

Furthermore,

$$\mathbb{E}_{\nu^n}\left[\frac{\left(\|g_K\|_{L^2(\nu_X)}^2 - \|g_K\|_n^2\right)^2}{\max\left\{\tau^2, \|g_K\|_{L^2(\nu_X)}^2\right\}}\right] = \mathbb{E}_{\nu^n}\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{\left(\|g_K\|_{L^2(\nu_X)}^2 - g_K^2(X_i)\right)}{\max\left\{\tau^2, \|g_K\|_{L^2(\nu_X)}\right\}}\right)^2\right].$$

Define $Z_{k,i} = \frac{(g_k^2(X_i) - \|g_k\|_{L^2(\nu_X)}^2)}{\max\left\{\tau^2, \|g_k\|_{L^2(\nu_X)}\right\}}$. Then $Z_{k,i}$ is zero-mean and

$$\begin{aligned}
\mathbb{E}_\nu\left[Z_{k,i}^2\right] =&\mathbb{E}_\nu\left[\frac{(g_k(X_i) - g_k(X_i'))^2(g_k(X_i) + g_k(X_i'))^2}{\tau^2 \vee \|g_k\|_{L^2(\nu_X)}^2}\right] \\
\leq&8\|g_k\|_{L^\infty(\nu_X)}^2 \frac{\mathbb{V}_\nu[g_k(X_i)]}{\max\left\{\tau^2, \|g_k\|_{L^2(\nu_X)}^2\right\}} \leq 8\|g_k\|_{L^\infty(\nu_X)}^2 \leq 32F^2,
\end{aligned}$$

where $X_i'$ are independent copies of $X_i$, and $|Z_{k,i}| \leq \|g_k\|_{L^\infty(\nu_X)}^2/\tau \leq 4F^2/\tau$ almost surely. By the Bernstein condition for bounded random variables, for all $k$ and $\lambda \in (0, 3\tau/4F^2)$,

$$\mathbb{E}_{\nu^n}\left[\exp\left(\lambda\sum_{i=1}^n Z_{k,i}\right)\right] \leq \exp\left(\frac{32n\lambda^2F^2}{2(1 - 4\lambda F^2/3\tau)}\right).$$

Applying Lem. 3, for any $\gamma > 0$,

$$\mathbb{E}_{\nu^n}\left[\max_k\left(\sum_{i=1}^n Z_{k,i}\right)^2\right] \leq 512\left(nF^2\ln(N_\delta/\gamma) + \gamma\max\left\{nF^2, \frac{F^4}{9\tau^2}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^4}{9\tau^2}\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\frac{\left(\|g_K\|^2_{L^2(\nu_X)} - \|g_K\|_n^2\right)^2}{\max\left\{\tau^2, \|g_K\|^2_{L^2(\nu_X)}\right\}}\right] &\leq \frac{1}{n^2}\mathbb{E}_{\nu^n}\left[\max_k\left(\sum_{i=1}^n Z_{k,i}\right)^2\right] \\ &\leq \frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right). \qquad (10)\end{aligned}$$

Substituting Eq. (9) and Eq. (10) into Eq. (8), we obtain

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] &\leq \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 8F\delta \\ &+ \left(\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] + 4F\delta + \tau^2\right)^{1/2}\left(\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right)\right)^{1/2}. \qquad (11)\end{aligned}$$

Now, for $a, b, c > 0$, if $x^2 \leq a + b\sqrt{x^2 + c^2}$, then $x^2 \leq a$ or

$$x^4 - 2x^2\left(a + \frac{b^2}{2}\right) + \left(a^2 - b^2c^2\right) \leq 0,$$

which implies that

$$x^2 \leq a + \frac{b^2}{2} + b\sqrt{a + \frac{b^2}{4} + c^2}. \qquad (12)$$

Therefore, from Eq. (11), we have

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] &\leq \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 8F\delta \\ &+ \frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right) \\ &+ \left(\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right)\right)^{1/2} \\ &\cdot \left(\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + \tau^2 + \frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right) + 12F\delta\right)^{1/2}.\end{aligned}$$

By the AM-GM inequality, for any $\zeta > 0$, we have

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] &\leq (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta \\ &+ \left(1 + \zeta + \frac{1}{4\zeta}\right)\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right) + \zeta\tau^2.\end{aligned}$$

Now set

$$\tau^2 = \frac{512F^2\ln(N_\delta/\gamma)}{n}.$$

Then,

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] \leq (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta + \left(1+\zeta+\frac{1}{4\zeta}\right)\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\left(1+\frac{9}{512\ln(N_\delta/\gamma)}\right)\right) + \frac{512\zeta F^2\ln(N_\delta/\gamma)}{n}.$$

Simplifying, we obtain

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] \leq (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta + \left(1+2\zeta+\frac{1}{4\zeta}\right)\frac{512F^2\ln(N_\delta/\gamma)}{n} + \left(1+\zeta+\frac{1}{4\zeta}\right)\frac{512F^2\gamma}{n}\left(1+\frac{9}{512\ln(N_\delta/\gamma)}\right).$$

Setting $\gamma = \ln(N_\delta)$, we have

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] \leq (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta + \left(2+3\zeta+\frac{1}{2\zeta}\right)\frac{512F^2\ln(N_\delta)}{n} + \left(1+\zeta+\frac{1}{4\zeta}\right)\frac{9F^2}{n}.$$

The arbitrariness of $f \in \mathcal{F}$ and $\zeta$ gives the desired result. □

### A.2.2 Proof of Lem. 2

*Proof of Lem. 2.* Starting from the regression model $Y = f^*(X) + \xi$, we obtain

$$\frac{1}{n}\sum_{i=1}^n (f_n(X_i) - Y_i)^2 = \|f_n - f^*\|_n^2 - 2\langle f_n - f^*, \xi\rangle_n + \|\xi\|_n^2.$$

Since $\|\xi\|_n^2$ is independent of $f_n$, for any $f \in \mathcal{F}_n$ that is independent of the sample, it follows that

$$\|f_n - f^*\|_n^2 - 2\langle f_n - f^*, \xi\rangle_n \leq \|f - f^*\|_n^2 - 2\langle f - f^*, \xi\rangle_n.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] =& \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2 - 2\langle f_n - f^*, \xi\rangle_n + 2\langle f_n - f^*, \xi\rangle_n\right] \\
\leq& \|f - f^*\|_{L^2(\nu_X)}^2 + 2\mathbb{E}_{\nu^n}[\langle f_n - f_K, \xi\rangle_n] + 2\mathbb{E}_{\nu^n}[\langle g_K, \xi\rangle_n] \\
\leq& \|f - f^*\|_{L^2(\nu_X)}^2 + 2\delta\mathbb{E}_\nu[|\xi|] + 2\mathbb{E}_{\nu^n}\left[\|g_K\|_n^2\right]^{1/2}\mathbb{E}_{\nu^n}\left[\left|\left\langle \frac{g_K}{\|g_K\|_n}, \xi\right\rangle_n\right|^2\right]^{1/2}. \qquad (13)
\end{aligned}$$

For the second term in Eq. (13), the subgaussianity of $\xi$ together with Markov's inequality yields

$$\begin{aligned}
\mathbb{E}_\nu[|\xi|] =& \int_0^\infty (\mathbb{P}_\nu\{\xi > t\} + \mathbb{P}_\nu\{-\xi > t\})dt \\
\leq& 2\int_0^\infty \inf_{\lambda>0}\exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right)dt \\
=& 2\int_0^\infty \exp\left(-\frac{t^2}{2\sigma^2}\right)dt \\
=& \sqrt{2\pi}\sigma. \qquad (14)
\end{aligned}$$

For the last term in Eq. (13), by the triangle inequality, we have

$$\|g_K\|_n^2 = \|f_n - f^*\|_n^2 + \|g_K\|_n^2 - \|f_n - f^*\|_n^2$$

$$\begin{aligned}
&\leq \|f_n - f_K\|_n^2 + \frac{1}{n}\sum_{i=1}^{n}(f_K(X_i) - f_n(X_i))(f_K(X_i) + f_n(X_i) - 2f^*(X_i)) \\
&\leq \|f_n - f^*\|_n^2 + 4F\delta, \qquad (15)
\end{aligned}$$

almost surely.

Conditioned on $X_1, ..., X_n$, $Z_k = \left\langle \frac{g_k}{\|g_k\|_n}, \xi \right\rangle_n$ is zero-mean and sub-Gaussian with variance proxy $\sigma^2/n$. Thus,

$$\begin{aligned}
\mathbb{E}_{\nu^n}\left[\exp\left(\lambda Z_k^2\right)\right] =& 1 + \int_1^\infty \mathbb{P}_{\nu^n}\left\{e^{\lambda Z_k^2} > t\right\} dt \\
=& 1 + \int_1^\infty \inf_{\kappa>0} \mathbb{P}_{\nu^n}\left\{e^{\kappa\sqrt{\lambda}|Z_k|} > e^{\kappa \ln^{1/2}(t)}\right\} dt \\
\leq& 1 + 2\int_1^\infty \inf_{\kappa>0} \exp\left(\frac{\lambda\sigma^2\kappa^2}{2n} - \kappa \ln^{1/2}(t)\right) dt \\
=& 1 + 2\int_1^\infty \exp\left(-\frac{n\ln(t)}{2\lambda\sigma^2}\right) dt \\
=& 1 + 2\int_1^\infty t^{-\frac{n}{2\lambda\sigma^2}} dt \\
=& 1 + \frac{2}{\frac{n}{2\lambda\sigma^2} - 1},
\end{aligned}$$

provided that $2\lambda\sigma^2 < n$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\nu^n}\left[\exp\left(\lambda \max_k Z_k^2\right)\right] &\leq \sum_k \mathbb{E}_{\nu^n}\left[\exp\left(\lambda Z_k^2\right)\right] \\
&\leq N_\delta\left(1 + \frac{2}{\frac{n}{2\lambda\sigma^2} - 1}\right).
\end{aligned}$$

By Jensen's inequality, it follows that

$$\mathbb{E}_{\nu^n}\left[\lambda \max_k Z_k^2\right] \leq \ln\left(\mathbb{E}_{\nu^n}\left[\exp\left(\lambda \max_k Z_k^2\right)\right]\right) \leq \ln(N_\delta) + \ln\left(1 + \frac{2}{\frac{n}{2\lambda\sigma^2} - 1}\right).$$

Setting $\lambda = \frac{n}{4\sigma^2}$, we obtain

$$\mathbb{E}_{\nu^n}\left[\max_k Z_k^2\right] \leq \frac{4\sigma^2}{n}\ln(3N_\delta). \qquad (16)$$

Substituting Eq. (14), Eq. (15), and Eq. (16) into Eq. (13), we have

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq \|f - f^*\|_{L^2(Q_X)}^2 + \sqrt{2\pi}\sigma\delta + 4\sqrt{\left(\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4F\delta\right)\frac{\sigma^2}{n}\ln(3N_\delta)}. \qquad (17)$$

Applying Eq. (12) to Eq. (17) yields

$$\begin{aligned}
\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq& \|f - f^*\|_{L^2(\nu_X)}^2 + \sqrt{2\pi}\sigma\delta \\
&+ \frac{8\sigma^2\ln(3N_\delta)}{n} + \left(\frac{16\sigma^2\ln(3N_\delta)}{n}\right)^{1/2}\left(\|f - f^*\|_{L^2(\nu_X)}^2 + \frac{4\sigma^2\ln(3N_\delta)}{n} + \left(\sqrt{2\pi}\sigma + 4F\right)\delta\right)^{1/2}.
\end{aligned}$$

By the AM-GM inequality, for any $\zeta > 0$, we have

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq (1+\zeta)\|f - f^*\|_{L^2(\nu_X)}^2 + \left((1+\zeta)\sqrt{2\pi}\sigma + 4F\zeta\right)\delta + \left(1 + \frac{\zeta}{2} + \frac{1}{2\zeta}\right)\frac{8\sigma^2\ln(3N_\delta)}{n}.$$

The arbitrariness of $\zeta$ gives the desired result. □

### A.3 Proof of Prop. 3

*Proof of Prop. 3.* Let $f = g_h \circ g_a \circ g_e \in \mathcal{F}$ and $f' = g_h \circ g'_a \circ g_e \in \mathcal{F}$. Then, we have

$$\begin{aligned}\|f - f'\|_{L^\infty(\nu)} =& \|g_h \circ g_a \circ g_e - g_h \circ g'_a \circ g_e\|_{L^\infty(\nu)} \\ =& \sup_{z \in g_e \circ \mathcal{X}} |g_h(g_a(z)) - g_h(g'_a(z))| \\ \le& C_\alpha \sup_{z \in \mathcal{Z}_i} \|g_a(z) - g'_a(z)\|^\alpha = C_\alpha \|g_a - g'_a\|^\alpha_{L^\infty},\end{aligned}$$

where $g_e \circ \mathcal{X} = \{g_e(x) : x \in \mathcal{X}\} \subseteq \mathcal{Z}_i$. Let $g_1, ..., g_N$ be a $\left(\frac{\delta}{C_\alpha}\right)^{1/\alpha}$-net for $\mathcal{G}$ with respect to $\|\cdot\|_{L^\infty}$. Then, for any $f = g_h \circ g'_a \circ g_e \in \mathcal{F}$, we have

$$\min_{i \in \{1,...,N\}} \|f - g_h \circ g_i \circ g_e\|_{L^\infty(\nu)} \le C_\alpha \min_{i \in \{1,...,N\}} \|g_a - g_i\|^\alpha_{L^\infty} \le \delta.$$

Hence, $\{g_h \circ g_i \circ g_e : i \in \{1, ..., N\}\}$ is a $\delta$-net for $\mathcal{F}$ with respect to $\|\cdot\|_{L^\infty(\nu)}$. Therefore, $N$ is greater than or equal to the covering number of $\mathcal{F}$ for the $\|\cdot\|_{L^\infty(\nu)}$-norm. □

### A.4 Proof of Prop. 4

*Proof of Prop. 4.* Let $f = g_h \circ g_a \circ g_e \in \mathcal{F}$ and $f^* = g_h \circ g^*_a \circ g_e$. Then, we have

$$\begin{aligned}\|f - f^*\|^2_{L^2(Q_X)} =& \int ((g_h \circ g_a \circ g_e)(x) - (g_h \circ g^*_a \circ g_e)(x))^2 Q_X(dx) \\ \le& C^2_\alpha \int (g_a(g_e(x)) - g^*_a(g_e(x)))^{2\alpha} Q_X(dx) \\ \le& C^2_\alpha \int \|g_a - g^*_a\|^{2\alpha}_{L^\infty} Q_X(dx) = C^2_\alpha \|g_a - g^*_a\|^{2\alpha}_{L^\infty}.\end{aligned}$$

Taking the square root of both sides, we get the desired claim. □

### A.5 Auxiliary Lemmas

#### A.5.1 Maximal Inequality for Finite Sum of Squares

We present here a maximal inequality that is useful for controlling the supremum of a finite collection of random variables, which arises in the analysis of covering numbers and empirical processes. The following lemma provides an upper bound on the expected maximum of the squared values of a finite sequence of independent random variables, under suitable moment generating function conditions.

**Lemma 3.** *Let $Z_1, ..., Z_N$ be independent random variables on a probability space $(\mathcal{Z}, \mathfrak{Z}, \nu)$ such that $Z_i$ are sub-Gamma with variance proxy $\sigma^2$ and scaling factor $F > 0$, i.e., for all $\lambda \in (0, 1/F)$, $\mathbb{E}_\nu[e^{\lambda |Z_k|}] \le \exp\left(\frac{\lambda^2 \sigma^2}{2(1 - F\lambda)}\right)$. Then,*

$$\mathbb{E}_\nu\left[\max_{i \in \{1,...,N\}} Z_i^2\right] \le 16 \inf_{\gamma > 0} \left(\sigma^2 \ln(N/\gamma) + \gamma \max\left\{\sigma^2, 2F^2 \ln(N/\gamma)\right\} + 2\gamma F^2\right).$$

*Proof of Lem. 3.* The expectation can be expressed in terms of probability as

$$\mathbb{E}_\nu\left[\max_{k \in \{1,...,N\}} Z_k^2\right] = \int_0^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,...,N\}} Z_k^2 > t\right] dt.$$

For any $z > 0$, we have

$$\begin{aligned}\int_0^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,...,N\}} Z_k^2 > t\right] dt \le& z^2 + \int_{z^2}^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,...,N\}} |Z_k| > \sqrt{t}\right] dt \\ =& z^2 + \int_{z^2}^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,...,N\}} \exp(\lambda |Z_k|) > \exp\left(\lambda \sqrt{t}\right)\right] dt\end{aligned}$$

$$
\begin{aligned}
&\leq z^2 + \sum_{i=1}^{N} \int_{z^2}^{\infty} \mathbb{P}_\nu \Big[ \exp(\lambda |Z_k|) > \exp\Big(\lambda \sqrt{t}\Big) \Big] dt \\
&\leq z^2 + \sum_{i=1}^{N} \int_{z^2}^{\infty} \inf_{\lambda \in (0, 1/F)} \mathbb{E}_\nu [\exp(\lambda |Z_k|)] e^{-\lambda \sqrt{t}} dt \\
&\leq z^2 + N \int_{z^2}^{\infty} \inf_{\lambda \in (0, 1/F)} \exp\left( \frac{\lambda^2 \sigma^2}{2(1 - F\lambda)} - \lambda \sqrt{t} \right) dt \\
&\leq z^2 + N \int_{z^2}^{\infty} \exp\left( - \frac{t}{2(\sigma^2 + F\sqrt{t})} \right) dt,
\end{aligned}
$$

where the fourth inequality uses Markov's inequality, and the last inequality follows from the standard proof for the Bernstein inequality. Regarding the integral part, we have

$$
\begin{aligned}
\int_{z^2}^{\infty} \exp\left( - \frac{t}{2(\sigma^2 + F\sqrt{t})} \right) dt &\leq \max\left\{ \int_{z^2}^{\infty} \exp\left( - \frac{t}{4\sigma^2} \right) dt, \int_{z^2}^{\infty} \exp\left( - \frac{\sqrt{t}}{4F} \right) dt \right\} \\
&= \max\Big\{ 4\sigma^2 e^{-z^2/4\sigma^2}, \big(8Fz + 32F^2\big) e^{-z/4F} \Big\}.
\end{aligned}
$$

Setting $z = 4 \max\Big\{ \sigma \ln^{1/2}(N/\gamma), F \ln(N/\gamma) \Big\}$, we have

$$
\begin{aligned}
&\mathbb{E}_\nu \left[ \max_{i \in \{1, \ldots, N\}} Z_i^2 \right] \\
&\leq 16 \ln(N/\gamma) \max\big\{ \sigma^2, F^2 \ln(N/\gamma) \big\} + \gamma \max\Big\{ 4\sigma^2, 32F(\sigma \ln^{1/2}(N/\gamma) + F), 32F(F \ln(N/\gamma) + F) \Big\}.
\end{aligned}
$$

If $\sigma^2 \leq F^2 \ln(N/\gamma)$, we have

$$
\begin{aligned}
&\mathbb{E}_\nu \left[ \max_{i \in \{1, \ldots, N\}} Z_i^2 \right] \\
&\leq 16\sigma^2 \ln(N/\gamma) + \gamma \max\big\{ 4F^2 \ln(N/\gamma), 32F(F \ln(N/\gamma) + F), 32F(F \ln(N/\gamma) + F) \big\} \\
&= 16\sigma^2 \ln(N/\gamma) + 32\gamma F(F \ln(N/\gamma) + F) \\
&= 16\big(\sigma^2 + 2\gamma F^2\big) \ln(N/\gamma) + 32\gamma F^2.
\end{aligned}
$$

If $\sigma^2 > F^2 \ln(N/\gamma)$, we have

$$
\begin{aligned}
&\mathbb{E}_\nu \left[ \max_{i \in \{1, \ldots, N\}} Z_i^2 \right] \\
&\leq 16\sigma^2 \ln(N/\gamma) + \gamma \max\big\{ 4\sigma^2, 32(\sigma^2 + F^2), 32(\sigma^2 + F^2) \big\} \\
&= 16\sigma^2 \ln(N/\gamma) + 32\gamma(\sigma^2 + F^2).
\end{aligned}
$$

Hence, we have

$$
\mathbb{E}_\nu \left[ \max_{i \in \{1, \ldots, N\}} Z_i^2 \right] \leq 16\sigma^2 \ln(N/\gamma) + 16\gamma \max\big\{ \sigma^2, 2F^2 \ln(N/\gamma) \big\} + 32\gamma F^2,
$$

which is equivalent to the desired result. □

# B DETAILED EXAMPLE: COMPOSITIONAL SPACE

In this section, we present a detailed application of Thm. 1 to the compositional spaces introduced in Section 4.1. Recall that the compositional space is defined as

$$
\mathcal{F}_H \circ \ldots \circ \mathcal{F}_1 = \{ f_H \circ \ldots \circ f_1 : f_i \in \mathcal{F}_i, i = 1, \ldots, H \}.
$$

Below, we provide rigorous definitions for the compositional sparse Hölder space (Schmidt-Hieber, 2020; Kohler and Langer, 2021) and the compositional anisotropic Besov space (Suzuki and Nitanda, 2021).

**Compositional Sparse Hölder Space.** The compositional sparse Hölder space corresponds to the case where each $\mathcal{F}_i$ is a sparse Hölder space. For a multi-index $\alpha \in \mathbb{N}^d$, we denote $\partial^\alpha = \partial^{\alpha_1} ... \partial^{\alpha_d}$ and $|\alpha| = \|\alpha\|_1$. For $\beta > 0$, the Hölder space is defined as

$$H^\beta([0,1]^d) = \left\{ f : [0,1]^d \to [0,1] : \|f\|_{H^\beta([0,1]^d)} < \infty \right\}.$$

where $\|f\|_{H^\beta([0,1]^d)} = \max_{\alpha \in \mathbb{N}^d : |\alpha| \leq \beta} \|\partial^\alpha f\|_{L^\infty([0,1]^d)} + \max_{\alpha \in \mathbb{N}^d : |\alpha| = \lfloor \beta \rfloor} |\partial^\alpha f|_{H^{\beta - \lfloor \beta \rfloor}([0,1]^d)}$,

$$|f|_{H^\beta([0,1]^d)} = \sup_{x,y \in (0,1)^d : x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|^\beta}.$$

The compositional sparse Hölder space is given by Eq. (4) with

$$\mathcal{F}_i = \mathcal{H}_i = \left\{ f = (f_j)_j : [0,1]^{d_j} \to [0,1]^{d_{j+1}} : f_j \in H^{\beta_j}([0,1]^{t_j}), \|f_j\|_{H^{\beta_j}([0,1]^{t_j})} \leq 1 \right\},$$

for some $\beta_j > 0$ and $1 \leq t_j \leq d_j$.

**Compositional Anisotropic Besov Space.** The compositional anisotropic Besov space arises when each $\mathcal{F}_i$ is an anisotropic Besov space. For a function $f : [0,1]^d \to [0,1]$, the $r$th finite difference is defined recursively as

$$\Delta_h^r(f)(x) = \Delta_h^{r-1}(f)(x+h) - \Delta_h^{r-1}(f)(x), \Delta_h^0(x) = f(x),$$

where $x \in [0,1]^d$ and $x + rh \in [0,1]^d$, and otherwise $\Delta_h^r(f)(x) = 0$. For a function $f \in L^p([0,1]^d)$ and $p \in (0, \infty]$, the $r$th modulus of smoothness of $f$ is defined as

$$w_{r,p}(f,t) = \sup_{h \in \mathbb{R}^d : |h_i| \leq t_i} \|\Delta_h^r(f)(x)\|_{L^p([0,1]^d)},$$

where $t = (t_1, ..., t_d)^\top \in \mathbb{R}^d, t_i > 0$. For $\beta = (\beta_1, ..., \beta_d)^\top \in \mathbb{R}^d$, where $\beta_i > 0$, the anisotropic Besov space is defined by

$$B_{p,q}^\beta([0,1]^d) = \{ f \in L^p : \|f\|_{B_{p,q}^\beta([0,1]^d)} < \infty \},$$

where $\|f\|_{B_{p,q}^\beta([0,1]^d)} = \|f\|_{L^p} + |f|_{B_{p,q}^\beta([0,1]^d)}$, and

$$|f|_{B_{p,q}^\beta([0,1]^d)} = \begin{cases} \left( \sum_{k=0}^\infty \left( 2^k w_{r,p}\left( f, \left( 2^{-k/\beta_1}, ..., 2^{-k/\beta_d} \right) \right) \right)^q \right)^{1/q} & \text{if } q < \infty, \\ \sup_{k \geq 0} 2^k w_{r,p}\left( f, \left( 2^{-k/\beta_1}, ..., 2^{-k/\beta_d} \right) \right) & \text{if } q = \infty. \end{cases}$$

The compositional anisotropic Besov space is given by Eq. (4) with

$$\mathcal{F}_i = \mathcal{B}_i = \left\{ f : [0,1]^{d_i} \to [0,1]^{d_{i+1}}, f_j \in B_{p,q}^{\beta_i}([0,1]^{d_i}), \|f\|_{B_{p,q}^{\beta_i}([0,1]^{d_i})} \leq 1 \right\},$$

where $\beta_i > 0$.

**Sparse ReLU Network.** Sparse ReLU networks are used in Schmidt-Hieber (2020); Kohler and Langer (2021); Suzuki and Nitanda (2021) to effectively approximate the compositional spaces. A ReLU network is a deep neural network with ReLU activation functions $\eta(x) = \max(0, x)$. For a vector $x \in \mathbb{R}^d$, $\eta(x)$ is applied element-wise. The class of ReLU networks with height $N_h$, width $N_w$, sparsity constraint $S$, and norm constraint $B$ is defined as

$$\mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B) = \Bigg\{ (W^{(N_h)} \eta(\cdot) + b^{(L)}) \circ ... \circ (W^{(1)} \eta(\cdot) + b^{(1)}) : W^{(L)} \in \mathbb{R}^{1 \times N_w}, b^{(L)} \in \mathbb{R}, W^{(1)} \in \mathbb{R}^{N_w \times d}, W^{(\ell)} \in \mathbb{R}^{N_w \times N_w}, b^{(\ell)} \in \mathbb{R}^{N_w}, \sum_{\ell=1}^{N_h} \left( \left\| W^{(\ell)} \right\|_\infty + \left\| b^{(\ell)} \right\|_\infty \right) \leq S, \max_{\ell = 1, ..., N_h} \left\| W^{(\ell)} \right\|_\infty \vee \left\| b^{(\ell)} \right\|_\infty \leq B \Bigg\}.$$

The approximation error of $\mathcal{F}_H \circ ... \circ \mathcal{F}_1$ with $H^{\beta_i}([0,1]^{t_i})$ and $B_{p,q}^{\beta_i}([0,1]^{d_i})$ by $\mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)$, as well as the covering number of $\mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)$, are established in the following theorems.

**Theorem 6** (Schmidt-Hieber (2020)). *Define* $\alpha_i = \prod_{\ell=i+1}^{H} \min\{1, \beta_\ell\}$. *For each* $J \in \mathbb{N}$, *there exist appropriate choices of* $N_h, N_w, S, B$ *such that*

$$N(\delta, \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B), \|\cdot\|_{L^\infty([0,1]^{d_1})}) \leq CJ\mathrm{polylog}(J, 1/\delta),$$

*and for any* $f^* \in \mathcal{H}_H \circ ... \circ \mathcal{H}_1$,

$$\inf_{f \in \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)} \|f - f^*\|_{L^\infty([0,1]^{d_1})} \leq C' J^{-\min_{i=1,...,H} \frac{\alpha_i \beta_i}{t_i}},$$

*for some constants* $C, C' > 0$.

**Theorem 7** (Suzuki and Nitanda (2021)). *Define* $\tilde{\beta}^{(\ell)} = (\sum_{j=1}^{d} 1/\beta_j^{(\ell)})^{-1}$ *and* $\alpha_i = \prod_{k=\ell+1}^{H} (\min\{1, (\beta_{\min}^{(\ell)} - 1/p)\})$, *where* $\beta_{\min}^{(\ell)} = \min_{j=1,...,d} \beta_j^{(\ell)}$. *Assume that* $\tilde{\beta}^{(\ell)} > 1/p$ *for all* $\ell$. *For each* $J \in \mathbb{N}$, *there exist appropriate choices of* $N_h, N_w, S, B$ *such that*

$$N(\delta, \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B), \|\cdot\|_{L^\infty(\lambda)}) \leq CJ \ln(J)\big(\ln^2(J) + \ln(1/\delta)\big),$$

*and for any* $f^* \in \mathcal{B}_H \circ ... \circ \mathcal{B}_1$,

$$\inf_{f \in \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)} \|f - f^*\|_{L^\infty(\lambda)} \leq C' J^{-\min_{\ell \in \{1,...,H\}} \alpha_i \tilde{\beta}^{(\ell)}},$$

*for some constants* $C, C' > 0$.

**Error without a pre-trained model.** By combining Thm. 6 or Thm. 7 with Thm. 1, we obtain the following error rate:

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=1,...,H} n^{-\gamma_i} \mathrm{polylog}(n), \tag{18}$$

where $\gamma_i = \frac{2\alpha_i\beta_i}{2\alpha_i\beta_i + t_i}$ for the compositional sparse Hölder space and $\gamma_i = \frac{\alpha_i \tilde{\beta}^{(\ell)}}{2\alpha_i \tilde{\beta}^{(\ell)} + 1}$ for the compositional anisotropic Besov space.

**Error with a pre-trained model (caulking).** Now, consider the scenario where a pre-trained model $f_{\mathrm{pre}}$ is available, and $f^*$ is $(1/n, \mathcal{F}_{i_h} \circ \cdots \circ \mathcal{F}_{i_e})$-caulkable by $f_{\mathrm{pre}} = (g_{h,m}, g_{e,m})$ for some $1 < i_e < i_h < H$, with $\mathcal{F}_H \circ ... \circ \mathcal{F}_1$ being either the compositional sparse Hölder space or the compositional anisotropic Besov space. The complexity of the class $\mathcal{F}_{i_h} \circ ... \circ \mathcal{F}_{i_e}$ is also established by Thm. 6 or Thm. 7. Consequently, we have

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=i_e,...,i_h} n^{-\gamma_i} \mathrm{polylog}(n), \tag{19}$$

Comparing Eq. (18) and Eq. (19), we observe that the range of the index $i$ is restricted to a narrower subset by leveraging a pre-trained model, demonstrating the improved error rate achievable through empirical caulking with a pre-trained model.

# C DETAILS OF EXPERIMENTS

## C.1 Finetuning of CNNs

We use ResNet-50 (26M parameters, He et al. (2016)) and Wide ResNet-50-2 (68M parameters, Zagoruyko and Komodakis (2016)), pre-trained on ImageNet. Their BatchNorm layers are replaced with GroupNorm (Wu and He, 2018). These CNNs with GroupNorm are trained on other domains than the target domain (clipart) of Office-Home dataset (Venkateswara et al., 2017) and then finetune on the target domain. Each domain is split into 80% for training and 20% for validation.

When fine-tuning, adapters are inserted in between the image-feature extractor and the last classifier. Adapters are multi-layer perceptron with the ReLU activation. We adopt SGD with a learning rate of $10^{-3}$, a momentum of 0.9, a weight decay of $10^{-4}$ for pre-training on the other domain and AdamW with a learning rate of $10^{-5}$ and a weight decay of $10^{-2}$ for fine-tuning of the adapters. At each stage, the models were updated for 50 epochs. A single NVIDIA's H100 GPU in our internal cluster was used to run each trial. The best results on the validation split over five different random seeds were reported.

The attached script is to reproduce the experiments. To run it, `uv`[2] is required.

[2]See `https://docs.astral.sh/uv/` for the installation instruction.

### C.2 Integrating Vision Capabilities to LLMs

We use `nanovlm v0.2`[3] for this experiment and follow its settings other than our problem-specific settings described below. A Llama-3-style Transformer model, `SmolLM2-135M` (Allal et al., 2024), is adopted as the base language model, and the vision encoders of SigLIP2 Tschannen et al. (2025), `siglip2-base-patch16-224` and `siglip2-base-patch16-224`, are used as the visual feature extractors. The extracted visual features are projected to the language embedding space after applying pixel shuffle and adapters.

The adapters are optimized with AdamW with a learning rate of $10^{-2}$ and a weight decay of $10^{-2}$, and the language model is also optimized with AdamW with a learning rate of $10^{-5}$ and a weight decay of $10^{-2}$ for 64k iterations. The weights of visual feature extractors are fixed. To run each trial, we used eight H100 GPUs in our internal cluster. The best test accuracy on MMStar was reported.

[3] https://github.com/huggingface/nanoVLM/tree/v0.2