Provable Data Scaling Law for Meta Learning via Complexity Minimization

Anonymous authors
Paper under double-blind review

Abstract

Pre-training has become a fundamental paradigm in modern machine learning, with one of its key empirical benefits being reduced downstream sample complexity as the scale of pre-training data increases. However, existing theoretical frameworks for pre-training do not fully explain this phenomenon. In this paper, we introduce complexity minimization, a novel meta-representation learning framework designed to enable theoretical analysis of this scaling behavior, which learns representations by evaluating the downstream model complexity best suited to each domain and minimizing the worst-case such complexity across source domains. Our end-to-end theoretical analysis, spanning pre-training through downstream regression, shows that this framework provably captures this scaling behavior; in particular, we show that the error rate of few-shot adaptation improves as the amount of meta-training data grows. Empirically, we demonstrate that incorporating complexity regularization into existing meta-learning methods consistently improves downstream sample efficiency.

1 Introduction

Pre-training, encompassing self-supervised learning, representation learning, and meta-learning, is now a fundamental component of modern machine learning, as demonstrated by the recent success of foundation models, large models pre-trained on massive datasets. For example, in natural language processing, large-scale pre-trained language models such as BERT and GPT-3 have shown that representations learned from broad text corpora can be adapted to a wide range of downstream tasks, including natural language inference, question answering, and text generation (Devlin et al., 2019; Brown et al., 2020). In computer vision and vision–language learning, contrastive and promptable pre-training has led to models such as CLIP and the Segment Anything Model, which exhibit strong transferability across image classification, retrieval, and segmentation tasks with little or no task-specific supervision (Radford et al., 2021; Kirillov et al., 2023). A similar paradigm has also emerged in scientific and domain-specific applications: AlphaFold has transformed protein-structure prediction by leveraging large-scale biological sequence and structural information (Jumper et al., 2021), while medical large language models such as Med-PaLM illustrate the potential of pre-training and instruction tuning for clinical question answering (Singhal et al., 2023). In robotics, large transformer-based policies trained on diverse robot interaction data have shown improved generalization to new objects, environments, and instructions (Brohan et al., 2022).

The theoretical study of pre-training, including analyses for few-shot learning (Du et al.), in-context learning (Bai et al.; Li et al.; Kim et al.), and meta-learning (Collins et al.; Aliakbarpour et al., 2024; Huang et al., 2022; Wu et al.; Ding et al.; Block et al.), has revealed the advantage of pre-training in terms of the sample complexity of the downstream learning task. For example, Du et al. showed that the existence of a common linear representation shared across source and downstream tasks yields a reduction in sample complexity. Pre-training has also been shown to reduce downstream sample complexity in in-context learning under generalized linear models (Bai et al.), nonparametric regression models (Kim et al.), and a hypothesis class with bounded algorithmic stability (Li et al.). Furthermore, many researchers have shown that the meta-learning algorithms provably

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reduce the downstream sample complexity (Collins et al.; Aliakbarpour et al., 2024; Huang et al., 2022; Wu et al.; Ding et al.; Block et al.).

These results, however, are inconsistent with the empirical phenomenon known as the *scaling law*, first introduced by Kaplan et al.. The *data scaling law* for pre-trained models, in particular, shows that pre-training on more data leads to better error rate of the downstream learning task (Henighan et al.; Mikami et al., 2022). The aforementioned theoretical results cannot explain this empirical finding, since the downstream error rates they establish are independent of the pre-training data size.

Recently, Fukuchi et al. (2026) has provided a theoretical framework that can explain the data scaling law for pre-trained models. Their framework, caulking, adapts the pre-trained model to the downstream task by inserting an adapter, as in parameter-efficient fine-tuning (PEFT) methods. Their analysis establishes that training the pre-trained model so that the complexity of the adapter decreases as the pre-training data size grows provably reduces the sample complexity of the downstream task.

However, their results lack an end-to-end analysis from pre-training to downstream learning, leaving unclear the training strategy that achieves the data scaling law. Developing a pre-training algorithm that provably achieves the data scaling law is important not only for the theoretical understanding of recent advances in foundation models, but also for guiding the practical development of pre-trained models.

Our contribution The main contribution of this paper is a meta-representation learning algorithm together with its theoretical analysis, proving the achievability of the data scaling law. Our contributions are summarized as follows:

- We propose *complexity minimization*, a novel meta-representation learning framework that selects a feature extractor by minimizing the worst-case best model complexity across observed source domains. The best model complexity serves as a proxy for the convergence rate of the downstream excess error: for instance, when the underlying regression function is sparse, the sparsity level governs the downstream convergence rate, so minimizing it directly reduces downstream sample complexity.
- To instantiate complexity minimization, we construct a novel estimator of the best model complexity using Lepski's method (Lepskii), a principled adaptive model selection procedure that identifies the optimal complexity level from a sample without prior knowledge of the underlying complexity parameters.
- We provide an end-to-end theoretical analysis spanning meta-training through downstream learning and prove that the downstream error rate exhibits the data scaling law. Specifically, the downstream excess error achieves the rate

$$(n/\ln n)^{-\beta^* + O(\ln^{-\gamma} m)}, \tag{1}$$

  where m and n are the meta-learning and downstream sample sizes, respectively, $\beta^* > 0$ is the ideal downstream convergence exponent, and $\gamma > 0$ is a constant. The exponent in Eq. (1) approaches β^* as $m \to \infty$, meaning the downstream error decays faster with n as the meta-training sample size grows, which is precisely the data scaling law.
- We empirically verify that adding a norm-based complexity regularizer to standard meta-learning algorithms consistently improves downstream sample efficiency across multiple baselines and datasets.

All missing proofs are left to the appendix.

2 Problem Formulation

Notation For a positive integer m, let $[m] = \{1, \dots, m\}$. For sequences a_n and b_n, we write $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) if $a_n \leq C b_n$ (resp. $a_n \geq C b_n$) for some $C > 0$ and all n; $a_n \asymp b_n$ means both hold. For a vector $x \in \mathbb{R}^d$ and a function $f \colon \mathcal{X} \to \mathbb{R}$, we write $\|x\|_p$ for the ℓ^p-norm and $\|f\|_{L^p}$ for the L^p-norm. We write $\mathbb{P}$ and $\mathbb{E}$ for probability and expectation. For a measurable function $f \colon \mathcal{X} \to \mathbb{R}$ and a random variable X taking values in $\mathcal{X}$, we define $\|f\|_{L^p(X)} = (\mathbb{E}[|f(X)|^p])^{1/p}$ for $p \in [1, \infty)$. For a set A endowed with a metric ρ and $\epsilon > 0$,

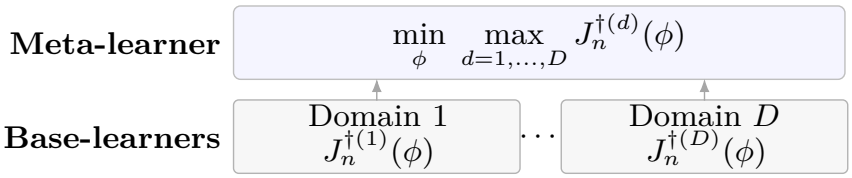

Figure 1: Conceptual diagram of *complexity minimization*. Each base-learner (bottom) assesses the best model complexity $J_n^{\dagger(d)}(\phi)$ for its domain d and reports it to the meta-learner. The meta-learner (top) selects ϕ to minimize the worst-case best model complexity over the observed domains, thereby reducing the downstream sample complexity across all domains.

$N(\epsilon, A, \rho)$ denotes the ϵ-covering number of A; for A endowed with a norm $\|\cdot\|$, we write $N(\epsilon, A, \|\cdot\|) = N(\epsilon, A, \rho_{\|\cdot\|})$ where $\rho_{\|\cdot\|}(x, y) = \|x - y\|$. Additional notation used in the proofs is collected in Appendix A.

Meta-representation learning problem Consider a representation learning problem with samples from multiple domains. Let $\mathcal{P}^*$ be the set of all pairs (X, Y) of random variables, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a feature and $Y \in \mathbb{R}$ is an outcome. Throughout, we assume $\mathbb{E}[Y|X] \in [0, 1]$ almost surely. Let $\mathcal{P} \subseteq \mathcal{P}^*$ denote the subset of feature-outcome pairs associated with all domains of interest, and let $\mathfrak{P} \subseteq 2^{\mathcal{P}^*}$ be the collection of all possible realizations of $\mathcal{P}$. The learner knows $\mathcal{P}^*$ but not $\mathcal{P}$. Let $(X^{(1)}, Y^{(1)}), \ldots, (X^{(D)}, Y^{(D)}) \in \mathcal{P}$ be the feature-outcome pairs for D observed domains, drawn i.i.d. from a distribution over $\mathcal{P}$. The learner observes m i.i.d. copies of each $(X^{(d)}, Y^{(d)})$, denoted $(X_1^{(d)}, Y_1^{(d)}), \ldots, (X_m^{(d)}, Y_m^{(d)})$. The goal is to learn a feature extractor $\phi\colon \mathcal{X} \to \mathbb{R}^p$ that minimizes the sample complexity of learning a regressor of the form $f \circ \phi$ for some $f\colon \mathbb{R}^p \to [0, 1]$ from an additional sample drawn from some $(X, Y) \in \mathcal{P}$, which we refer to as the downstream learning task.

Remark 1 (Intuition behind $\mathcal{P}^*$ and $\mathcal{P}$)**.** *The distinction between $\mathcal{P}^*$ and $\mathcal{P}$ is central to characterizing the conditions for successful meta representation learning. We assume that a single common feature representation performs well for all domains in $\mathcal{P}$, so that obtaining such a representation minimizes the sample complexity of the downstream task. In other words, $\mathcal{P}$ shares a common feature representation, whereas $\mathcal{P}^*$ encompasses all possible feature-outcome distributions over a variety of representations. Since the learner does not know $\mathcal{P}$, they do not know this common representation either. Identifying $\mathcal{P}$ from pre-training data is therefore valuable for reducing downstream sample complexity.*

Downstream regression problem In the downstream task, the learner receives an additional sample from some $(X, Y) \in \mathcal{P}$ and a pre-trained feature extractor $\phi\colon \mathcal{X} \to \mathbb{R}^p$, and finds a head function $f\colon \mathbb{R}^p \to [0, 1]$ such that $f \circ \phi$ is an accurate regressor for (X, Y). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n i.i.d. copies of (X, Y). The quality of f is measured by the expected squared error

$$E_\phi(f; X, Y) = \mathbb{E}\Big[((f \circ \phi)(X) - Y)^2\Big].$$

Equivalently, f minimizes the excess error $\bar{E}_\phi(f; X, Y) = \|f \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2(X)}$, where $\mathbb{E}[Y|X]$ is the Bayes optimal regressor for (X, Y). Let f_n denote the head function learned from the sample. The sample complexity of the downstream task is characterized by the rate at which $\bar{E}_\phi(f_n; X, Y)$ decreases as n grows.

3 Complexity Minimization

We propose *complexity minimization* (Fig. 1), a meta-representation learning framework following the meta-learner/base-learner architecture of existing approaches (Hospedales et al.; Finn et al.; Wu et al.; Ding et al.). The meta-learner maintains the feature extractor ϕ as its meta-parameter (Fig. 1 top); each base-learner is associated with one observed domain and evaluates ϕ by a domain-specific criterion (Fig. 1 bottom). Many existing meta-learning

algorithms, including MAML (Finn et al.), instantiate this criterion as the downstream regression error.

Our key departure is to replace the regression error with the *best model complexity* of the head function, which serves as a proxy for the convergence rate of the downstream excess error. Formally, let $\mathcal{F}_J$ be a sequence of increasing classes of head functions $f\colon \mathbb{R}^p \to [0,1]$, indexed by a complexity parameter $J \in \mathbb{N}$, where model complexity (e.g., the number of non-zero weights) increases with J. Letting $f_{n,J}$ denote the head function learned over $\mathcal{F}_J$ from a downstream sample of size n, the best model complexity for domain d under ϕ is

$$J_n^{\dagger(d)}(\phi) = \arg\min_{J\in\mathbb{N}} \bar{E}_\phi(f_{n,J}; X^{(d)}, Y^{(d)}).$$

When, for instance, the head function is truly sparse, the minimal sufficient sparsity level governs the downstream convergence rate; a smaller best model complexity therefore implies faster downstream learning.

The meta-learner collects these criteria from every base-learner and selects ϕ to minimize the worst-case value across all observed domains:

$$\min_{\phi\in\Phi} \max_{d=1,\dots,D} J_n^{\dagger(d)}(\phi), \tag{2}$$

where Φ is a class of feature extractors. Because $J_n^{\dagger(d)}(\phi)$ depends on the unknown downstream distribution, it must be estimated from pre-training samples in practice. This estimation step is precisely what connects complexity minimization to the data scaling law: larger pre-training samples yield more accurate complexity estimates, resulting in a smaller selected model complexity and therefore a faster downstream convergence rate across all domains in $\mathcal{P}$.

4 Provable Scaling Laws via Complexity Minimization

In this section, we present a concrete instantiation of the complexity minimization framework and establish the data scaling law for the resulting algorithm.

Downstream rate We employ a specific characterization of the downstream error rate $\bar{E}_\phi(f_{n,J}; X, Y)$ to build the concrete algorithm. Specifically, the downstream error is characterized by two quantities: the approximation error and the Minkowski–Bouligand dimension. This characterization is applicable to deep neural network based estimators (Schmidt-Hieber; Imaizumi & Schmidt-Hieber; Suzuki; Nishimura & Suzuki; Hayakawa & Suzuki; Kohler & Langer; Chen et al.) and hence covers modern machine learning algorithms.

We first introduce these two quantities and then present the characterization on the downstream error. Given $\phi \in \Phi$ and $(X, Y) \in \mathcal{P}^*$, the approximation error of the regression function under ϕ is

$$A_J(\phi; X, Y) = \min_{f_J\in\mathcal{F}_J} \|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2(X)}.$$

The Minkowski–Bouligand dimension of a set S with respect to a norm $\|\cdot\|$ is defined as

$$d_{\mathrm{MB}}(S; \|\cdot\|) = \limsup_{\epsilon\to 0} \frac{\ln N(\epsilon, S, \|\cdot\|)}{\ln(1/\epsilon)}.$$

Building from these definitions, we have the following theorem.

Theorem 1 (Based on Schmidt-Hieber; Hayakawa & Suzuki)**.** *Fix $(X, Y) \in \mathcal{P}^*$ and $\phi \in \Phi$. Let $\mathcal{F}_J$ be a sequence of increasing classes of functions $f : \mathbb{R}^p \to [0,1]$ such that $d_{\mathrm{MB}}(\mathcal{F}_J; \|\cdot\|_{L^\infty}) \le J$ for any $J \in \mathbb{N}$. Then, there is a learning algorithm $f_{n,J}$ such that with high probability,*

$$\bar{E}_\phi(f_{n,J}; X, Y) \lesssim A_J(\phi; X, Y) + \frac{J\ln(n)}{n}. \tag{3}$$

The convergence rate induced from Thm. 1 is determined by the decreasing rate of the approximation error as J grows. For example, if $A_J(\phi; X, Y) \asymp J^{-2\alpha}$ for some $\alpha > 0$, then the convergence rate is $\asymp (n/\ln n)^{-2\alpha/(2\alpha+1)}$ with $J \asymp (n/\ln n)^{1/(2\alpha+1)}$, derived by optimizing the right hand side of Eq. (3) for J. As the choice of J depends on the unknown parameter α, we refer to this rate as the oracle rate.

Ideal downstream rate We introduce the ideal decreasing rate of the downstream error. In our analysis, we focus only on the polynomial decreasing rate of A_J.

Assumption 1 (Polynomial decreasing rate of A_J)**.** *There exists a functional $\alpha(\phi; X, Y) \in (0, \infty)$ for $(X, Y) \in \mathcal{P}^*$ and $\phi \in \Phi$ such that $A_J(\phi; X, Y) \asymp J^{-2\alpha(\phi;X,Y)}$. Additionally, there exists a constant $\bar{\alpha} < \infty$ such that $\alpha(\phi; X, Y) \le \bar{\alpha}$ for any $\phi \in \Phi$ and $(X, Y) \in \mathcal{P}^*$.*

We write $\beta(\phi; X, Y) = 2\alpha(\phi; X, Y)/(2\alpha(\phi; X, Y) + 1)$ and $\bar{\beta} = 2\bar{\alpha}/(2\bar{\alpha} + 1)$. From Thm. 1, $\bar{E}_\phi(f_{n,J}; X, Y) \lesssim (n/\ln n)^{-\beta(\phi;X,Y)}$ with appropriately chosen J for fixed ϕ and $(X, Y) \in \mathcal{P}$. The ideal convergence exponent for a given $\mathcal{P} \in \mathfrak{P}$ is therefore

$$\beta^*_{\mathcal{P}} := \sup_{\phi\in\Phi} \inf_{(X,Y)\in\mathcal{P}} \beta(\phi; X, Y).$$

Technical assumptions For our main theorem, we need several technical assumptions. First, we introduce an assumption about the complexity of the class of feature extractors, Φ.

Assumption 2 (Complexity of Φ)**.** *There exist $\beta_0 > 0$ with $\beta_0 + \bar{\beta} < 1$ and $\gamma \in (0, 1]$ such that for any $\mathcal{P} \in \mathfrak{P}$, for all $m \ge 1$,*

$$\sum_{J\in[m]:J\le\frac{m}{\ln m}} N(V_{m,J}, \Phi, \rho_J) \lesssim m^{(m/\ln m)^{\beta_0}},$$

and

$$\ln N(\ln^{-\gamma} m, \Phi, \rho_{\beta,\mathcal{P}}) \lesssim \ln m,$$

where $\rho_J(\phi, \phi') = \sup_{f\in\mathcal{F}_J} \|f \circ \phi - f \circ \phi'\|_{L^\infty}$ and $\rho_{\beta,\mathcal{P}}(\phi, \phi') = \sup_{(X,Y)\in\mathcal{P}} |\beta(\phi; X, Y) - \beta(\phi'; X, Y)|$.

Asm. 2 requires that two types of metric entropy conditions on Φ hold simultaneously. Constructing concrete families Φ satisfying Asm. 2 is an important open problem.

Next, we introduce an assumption about the distribution over the domains.

Assumption 3 (Uniform domain sampling)**.** *For each $\mathcal{P} \in \mathfrak{P}$, $(X, Y) \in \mathcal{P}$ is distributed by the domain distribution. There exists $\nu > 0$ such that for any $\delta > 0$ and any $\phi \in \Phi$, the domain-distributed (X', Y') satisfies $|\beta(\phi; X, Y) - \beta(\phi; X', Y')| \le \delta$ with probability at least $C\delta^\nu$ for some universal constant $C > 0$.*

Since the learner observes samples from only finitely many domains, these domains must collectively represent all of $\mathcal{P}$ for the meta-learned representation to generalize. Asm. 3 imposes a near-uniform condition on the domain distribution, ensuring that the observed domains approximately cover $\mathcal{P}$ in terms of β when D is sufficiently large.

Lastly, we introduce a mild assumption about the noise in Y.

Assumption 4 (Sub-gaussian noise)**.** *For any $(X, Y) \in \mathcal{P}^*$, conditioned on X, $Y - \mathbb{E}[Y|X]$ is sub-gaussian with variance proxy at most σ^2; namely, for all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda(Y - \mathbb{E}[Y|X]))|X] \le e^{\sigma^2\lambda^2/2}$ almost surely.*

Asm. 4 is a standard assumption employed in a broad literature (Schmidt-Hieber; Imaizumi & Schmidt-Hieber; Suzuki; Nishimura & Suzuki; Hayakawa & Suzuki; Kohler & Langer; Chen et al.).

Main result We present our main theoretical result, exhibiting meta-learning and downstream algorithms that provably achieve the data scaling law.

Theorem 2 (Main theorem). *Assume asm. 1 to 4. There exist a meta-learning algorithm and a downstream learning algorithm such that, if* $\ln^{1+\nu\gamma} m \lesssim D$ *and* $\ln D \lesssim \ln m$*, then with probability at least* $1 - O(m^{-1}) - O(n^{-1})$*,*

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim \left(\frac{n}{\ln(n)}\right)^{-\beta_{\mathcal{P}}^* + O(\frac{1}{\ln^{\gamma} m})}.$$

The error rate in Thm. 2 exhibits the data scaling law: the rate at which the downstream error decreases in n improves as the meta-learning sample size m grows. To the best of our knowledge, this is the first end-to-end theoretical analysis establishing the data scaling law for pre-training. The proof constructs concrete meta-learning and downstream algorithms and establishes that both achieve the rate stated in Thm. 2. The analyses of our meta-learning and downstream algorithms for proving Thm. 2 are found in Appendix A.

4.1 Base-learner

Our meta-learning algorithm consists of interacting base-learner and meta-learner as described in Section 3, and in this subsection, we describe the concrete construction of the base-learner. In the complexity minimization framework, the base-learner assesses the best model complexity $J_n^{\dagger(d)}(\phi)$. Because this quantity depends on the unknown downstream distribution, each base-learner must estimate it from pre-training samples. Our idea in estimating the best model complexity is to employ the *Lepski's method* (Lepskii), which selects the model complexity adaptively without knowledge of the underlying complexity parameters.

Lepski's method Lepski's method (Lepskii) is a powerful tool for adaptive model selection in nonparametric statistics and can automatically find the optimal model complexity from a sample without prior knowledge of the underlying complexity parameters. For example, it builds estimators for nonparametric regression within smooth function classes such as Hölder, Sobolev, and Besov spaces, achieving convergence rates determined by the smoothness parameter without prior knowledge of it (Lepskii; Lepski et al.).

We now instantiate Lepski's method using the oracle rate from Thm. 1 with fixed ϕ and $(X, Y) \in \mathcal{P}^*$. The idea is to estimate $A_J(\phi; X, Y)$ and select J so that the estimated A_J and the term $\frac{J \ln(n)}{n}$ are balanced. Let $f^*_{Y|X,\phi,J}$ be the best regressor in $\mathcal{F}_J$ such that

$$\left\| f^*_{Y|X,\phi,J} \circ \phi - \mathbb{E}[Y|X] \right\|^2_{L^2(X)} = \min_{f_J \in \mathcal{F}_J} \|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2(X)},$$

where $f^*_{Y|X,\phi,J}$ is an arbitrary one if the tie occurs. We omit the first and second subscripts to denote f^*_J if (X, Y) and ϕ are clear from the context. Then, the algorithm selects J following the Lepski's rule, defined as

$$J_n^*(\phi; X, Y) = \min\left\{ J \in [m] : \forall J \le J' \le \frac{m}{\ln m}, \|f^*_J \circ \phi - f^*_{J'} \circ \phi\|^2_{L^2(X)} \le \rho V_{n,J'} \right\}, \quad (4)$$

where $V_{n,J} = \frac{J \ln(n)}{n}$ is referred to as a variance term or a majorant, and $\rho > 0$ is a constant chosen as specified in the analyses. The intuition behind the Lepski's rule is that if f^*_J sufficiently approximates the regression function, then increasing J does not significantly deviate f^*_J up to the variance.

Empirical estimation The base-learner for domain d estimates the Lepski rule value $J_m^*(\phi; X^{(d)}, Y^{(d)})$ from Eq. (4) as a proxy for the best model complexity $J_n^{\dagger(d)}(\phi)$ from Eq. (2). For each J, a sieved least-squares estimator is computed, yielding the regressor

$$\hat{f}^{(d)}_{m,J} = \arg\min_{f_J \in \mathcal{F}_J} \frac{1}{m} \sum_{i=1}^{m} \left((f_J \circ \phi)(X_i^{(d)}) - Y_i^{(d)} \right)^2.$$

Then, the estimated complexity is obtained as

$$\hat{J}^{(d)}(\phi) = \min\left\{ J \in [m] : \forall J \le J' \le \frac{m}{\ln m},\right.$$

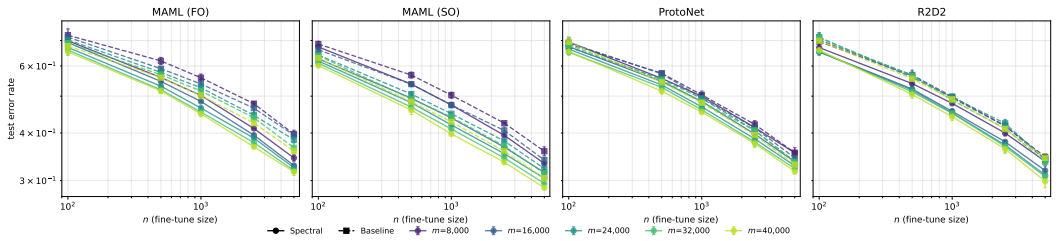

Figure 2: **Regularizing model complexity (parameter spectral norms) improves downstream sample efficiency.** Test error rate on CIFAR-10 (log scale) vs. fine-tuning size (n, log scale) with different pre-training size m for four meta-learning algorithms trained on Mini-ImageNet in the 5-way 1-shot setting with and without model complexity regularization.

$$\frac{1}{m}\sum_{i=1}^{m}\Big((\hat{f}^{(d)}_{m,J}\circ\phi)\Big(X^{(d)}_i\Big)-(\hat{f}^{(d)}_{m,J'}\circ\phi)\Big(X^{(d)}_i\Big)\Big)^2\le\rho V_{m,J'}\Bigg\}.$$

4.2 Meta-learner

The meta-learner collects the estimated complexity $\hat{J}^{(d)}(\phi)$ from all D base-learners and selects ϕ to minimize the worst-case estimated complexity, forming the empirical counterpart of Eq. (2). Specifically, the estimated feature extractor is defined as

$$\hat{\phi}=\arg\min_{\phi\in\Phi}\max_{d\in[D]}\hat{J}^{(d)}(\phi). \tag{5}$$

Minimizing the worst case ensures that $\hat{\phi}$ simultaneously reduces the estimated downstream complexity across all observed domains, yielding a feature extractor whose downstream performance generalizes uniformly over $\mathcal{P}$.

4.3 Downstream Algorithm

At downstream time, we again carry out Lepski's method to construct the learned regressor. Specifically, define

$$\hat{f}_{n,J}=\arg\min_{f_J\in\mathcal{F}_J}\frac{1}{n}\sum_{i=1}^{n}\Big((f_J\circ\hat{\phi})(X_i)-Y_i\Big)^2.$$

Then, the complexity is estimated as

$$\hat{J}_n=\min\Bigg\{J\in[n]:\forall J\le J'\le\frac{n}{\ln n},\frac{1}{n}\sum_{i=1}^{n}\Big((\hat{f}_{n,J}\circ\hat{\phi})(X_i)-(\hat{f}_{n,J'}\circ\hat{\phi})(X_i)\Big)^2\le\rho V_{n,J'}\Bigg\}.$$

Consequently, the learned regressor is given by $\hat{f}_n=\hat{f}_{n,\hat{J}_n}$.

5 Experiments

We empirically validate the effect of complexity minimization by adding a complexity regularization term to meta-learning algorithms. The experiments demonstrate how the pre-training size m affects the downstream scaling behavior in terms of the fine-tuning size n. Four representative meta-learning algorithms are adopted, which are discussed in Section 6: first- and second-order MAML (Finn et al.), Prototypical Networks (Snell et al., 2017), and R2-D2 (Bertinetto et al., 2019). The meta-losses of these algorithms are augmented with a spectral norm-based regularizer on the model parameters. Further experimental details are described in Appendix C.

Fig. 2 reports the downstream test error rates on CIFAR-10 (Krizhevsky, 2009) for a CNN, whose feature extractor is pre-trained with each meta-learning algorithm on Mini-ImageNet (Ravi & Larochelle, 2017) with m samples in the 5-way 1-shot setting and then

fine-tuned on a subset of the CIFAR-10 training dataset with n samples. Regularizing the spectral norm of the parameters as a measure of model complexity yields a clear improvement in downstream performance. Additional experiments in Appendix C show that complexity minimization works on other datasets and meta-learning settings, and regularizing other norms, such as the ℓ_1 norm, also improves downstream sample efficiency. Together, these algorithm-agnostic results provide empirical support for our theoretical claims.

6 Related Work

In this section, we briefly review prior works closely related to our study. A more comprehensive literature survey is provided in Appendix D.

Meta-Learning Methodologies. Meta-learning aims to acquire a learning procedure that can rapidly adapt to unseen tasks from limited samples by exploiting experiences from a collection of past tasks drawn from a task distribution (Hospedales et al., 2021; Wang et al., 2020). A standard taxonomy divides meta-learning methods into metric-based, optimization-based, and model-based approaches. **Metric-based methods** classify queries by proximity, attention, or comparison in an embedding space. Representative examples include Matching Networks, which introduced one-shot classification via attention over a support set and formalized episodic training (Vinyals et al., 2016); Prototypical Networks, which classify queries by distances to class-wise mean embeddings and provide a clear view of meta-representation learning (Snell et al., 2017); and Relation Networks, which learn the comparison function itself using a neural network (Sung et al., 2018). **Optimization-based methods** learn an initialization, update rule, or inner-loop adaptation mechanism such that a few optimization steps on a new task yield strong performance. MAML established a model-agnostic framework based on inner-loop gradient descent and outer-loop optimization of post-adaptation performance (Finn et al.). Meta-SGD further learns the initialization, update directions, and learning rates (Li et al., 2017). R2-D2 replaces iterative inner-loop adaptation with a differentiable closed-form ridge-regression base learner on top of learned embeddings, occupying an intermediate position between metric-based classifiers and gradient-based adaptation methods (Bertinetto et al., 2019). **Model-based methods** implement adaptation within the network architecture itself, using memory, hypernetworks, or learned optimizers. Memory-Augmented Neural Networks use external memory for rapid one/few-shot adaptation (Santoro et al., 2016), while learning-to-learn approaches such as Optimization as a Model learn update rules using recurrent architectures (Ravi & Larochelle, 2017; Hochreiter et al., 2001; Andrychowicz et al., 2016). SNAIL further combines temporal convolutions and attention as a general-purpose meta-learner across supervised and reinforcement learning domains (Mishra et al., 2017).

Meta-Representation Learning: Sharing Representations Across Tasks Meta-learning is closely related to transfer learning, since both transfer information from previous tasks to unseen ones (Pan & Yang, 2009). Representation learning motivates the acquisition of shared latent features that facilitate learning across tasks (Bengio et al., 2013), and meta-representation learning specializes such shared representations for few-shot task adaptation with statistical and computational efficiency. Classically, Baxter's inductive bias learning model formalized meta-generalization as learning a good hypothesis space from multiple tasks sampled from a task environment (Baxter, 2000). Subsequent work on Multi-Task Representation Learning established generalization bounds showing the benefit of learning low-dimensional dictionaries or feature maps shared across tasks (Maurer et al., 2016; 2013). More recent theory studies sample-efficient estimation and transfer of shared low-dimensional linear representations across linear regression tasks (Tripuraneni et al., 2021), as well as the role of overparameterization in enabling few-shot adaptation with large-scale models (Sun et al., 2021).

Learning Theory of (Deep) Meta-Learning Learning theory for meta-learning must handle a dual-sampling structure: tasks are sampled from a task distribution, and data points are sampled within each task (Baxter, 2000; Hospedales et al., 2021). Recent studies typically decompose excess risk into statistical estimation, optimization, and approximation

errors, often through the meta-generalization gap between the expected risk on unseen tasks and the empirical meta-objective (Rezazadeh, 2022; Wang & Arora, 2024). This line of work has clarified how representation sharing and the number of inner-loop adaptation steps affect sample efficiency and stability (Huang et al., 2022; Chen et al., 2022). Several theoretical frameworks have been developed. Algorithmic stability, including meta-stability for both inner and outer loops, yields realistic bounds for gradient-based and non-convex meta-learning algorithms (Wang & Arora, 2024; Bousquet & Elisseeff, 2002). PAC-Bayes theory introduces hierarchical meta-priors and task-specific posteriors to obtain bounds depending on both the number of tasks and within-task sample size (Pentina & Lampert, 2014; Amit & Meir, 2018; Rezazadeh, 2022). Information-theoretic analyses bound generalization via mutual information between algorithm outputs and data (Chen et al., 2021). Uniform convergence remains a classical approach, but its bounds are often loose for deep learning and meta-learning, motivating the recent shift toward data-dependent analyses (Nagarajan & Kolter, 2019). Crucially, these results may not explain the data scaling law for pre-training, as their error rates with respect to the downstream sample size are independent of the pre-training sample size.

7 Conclusion

We introduced complexity minimization, a meta representation learning framework that provably achieves the data scaling law for pre-training. The framework instructs each base-learner to estimate the best model complexity for its domain via Lepski's method, and the meta-learner selects the feature extractor that minimizes the worst-case complexity across all observed domains. Our end-to-end theoretical analysis establishes that the downstream excess error decays faster with the downstream sample size n as the meta-training size m grows, formally capturing the data scaling law. Empirically, augmenting standard meta-learning algorithms with a norm-based complexity regularizer consistently improves downstream sample efficiency across multiple algorithms and datasets, supporting the theoretical claims.

Limitations and broader impacts While the complexity assumptions on Φ in Asm. 2 are reasonable, exhibiting a concrete class Φ that satisfies them remains open, and constructing such feature families is an important direction for future work. Our theoretical analysis likewise focuses on regression; extending it to classification and structured prediction remains open. We build on standard meta-learning and do not propose a qualitatively new paradigm with novel direct societal risks. As in broader work on large-scale pre-training and foundation models, familiar considerations regarding computational cost and equitable access to data and compute nonetheless apply.

AI use statement

In this work, we used generative AI tools for assisting in the writing of proofs, implementing methods, assistting with translation, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refineing hypotheses. Additionally, we used generative AI tools for identifying relevant literature, drafting parts of a research paper, editing a research paper to improve readability. We have reviewed all AI-assisted work. We draft the content of the paper by authors and collaborate with AI tools to improve their quality. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Maryam Aliakbarpour, Konstantina Bairaktari, Gavin Brown, Adam Smith, Nathan Srebro, and Jonathan Ullman. Metalearning with very few samples per task. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 46–93. PMLR, 2024.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pp. 205–214. PMLR, 2018.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211. URL https://papers.nips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

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.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Jacob L. Block, Sundararajan Srinivasan, Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Provable meta-learning with low-rank adaptations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. URL https://openreview.net/forum?id=QUN6uidabr.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8): 084002, 2025.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. Jax: composable transformations of python+ numpy programs, 2018. URL http://github.com/jax-ml/jax.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, and others. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Lisha Chen, Songtao Lu, and Tianyi Chen. Understanding benign overfitting in gradient-based meta learning. *Advances in neural information processing systems*, 35:19887–19899, 2022.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. 11(4):1203–1253. ISSN 2049-8772. doi: 10.1093/imaiai/iaac001. URL https://doi.org/10.1093/imaiai/iaac001.

Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 34:25878–25890, 2021.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Liam Collins, Aryan Mokhtari, Sewoong Oh, and Sanjay Shakkottai. Maml and anil provably learn representations. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 4238–4310. PMLR. URL https://proceedings.mlr.press/v162/collins22a.html.

DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The deepmind jax ecosystem, 2020. URL http://github.com/google-deepmind.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Wenjun Ding, Jingling Liu, Lixing Chen, Xiu Su, Tao Sun, Fan Wu, and Zhe Qu. On the stability and generalization of meta-learning: The impact of inner-levels. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. URL https://openreview.net/forum?id=l1L0Yhh6x6.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

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

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135. PMLR. URL https://proceedings.mlr.press/v70/finn17a.html.

Kazuto Fukuchi, Ryuichiro Hataya, and Kota Matsui. Provable target sample complexity improvements as pre-trained models scale. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

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.

Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. 123:343–361. ISSN 0893-6080. doi: 10.1016/j.neunet.2019.12.014. URL https://www.sciencedirect.com/science/article/pii/S089360801930406X.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. URL https://doi.org/10.48550/arXiv.2010.14701.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv:2102.01293*, 2021.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv:1712.00409*, 2017.

Sepp Hochreiter, A. Steven Younger, and Peter R. Conwell. Learning to learn using gradient descent. In *International conference on artificial neural networks*, pp. 87–94. Springer, 2001.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and others. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. 44(09):5149–5169. ISSN 0162-8828. doi: 10.1109/TPAMI.2021.3079209. URL https://www.computer.org/csdl/journal/tp/2022/09/09428530/1twaJR3AcJW.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Yu Huang, Yingbin Liang, and Longbo Huang. Provable generalization of overparameterized meta-learning trained with sgd. *Advances in Neural Information Processing Systems*, 35: 16563–16576, 2022.

Masaaki Imaizumi and Johannes Schmidt-Hieber. On generalization bounds for deep networks based on loss surface implicit regularization. 69(2):1203–1223. ISSN 1557-9654. doi: 10.1109/TIT.2022.3215088. URL https://ieeexplore.ieee.org/document/9919858.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and others. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873): 583–589, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. URL https://doi.org/10.48550/arXiv.2001.08361.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in jax via callable pytrees and filtered transformations. In *Differentiable Programming Workshop at Neural Information Processing Systems*, 2021.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106667–106713. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c11daad0a48ea5f3c5c6390c7b060720-Abstract-Conference.html.

Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, and others. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. 49(4):2231–2249. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2034. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-4/On-the-rate-of-convergence-of-fully-connected-deep-neural/10.1214/20-AOS2034.full.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. 25(3):929–947. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1069362731. URL https://projecteuclid.org/journals/annals-of-statistics/volume-25/issue-3/Optimal-spatial-adaptation-to-inhomogeneous-smoothness--an-approach-based/10.1214/aos/1069362731.full.

O. V. Lepskii. On a problem of adaptive estimation in gaussian white noise. 35(3):454–466. ISSN 0040-585X. doi: 10.1137/1135065. URL https://epubs.siam.org/doi/10.1137/1135065.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19565–19594. PMLR. URL https://proceedings.mlr.press/v202/li23l.html.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv:1707.09835*, 2017.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16167–16180, Suzhou, China, Nov 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.877. URL https://aclanthology.org/2025.findings-emnlp.877/.

Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *International conference on machine learning*, pp. 343–351. PMLR, 2013.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Hiroaki Mikami, Kenji Fukumizu, Shogo Murai, Shuji Suzuki, Yuta Kikuchi, Taiji Suzuki, Shin-ichi Maeda, and Kohei Hayashi. A scaling law for syn2real transfer: How much is your pre-training effective? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 477–492. Springer, 2022.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv:1707.03141*, 2017.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in neural information processing systems*, 32, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

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.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv:1803.02999*, 2018.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *Forty-second International Conference on Machine Learning*, 2025.

Yuto Nishimura and Taiji Suzuki. Minimax optimality of convolutional neural networks for infinite dimensional input-output problems and separation from kernel methods. In *The Twelfth International Conference on Learning Representations*. URL https://openreview.net/forum?id=EW8ZExRZkJ.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Kenta Oono and Taiji Suzuki. Approximation and non-parametric estimation of resnet-type convolutional neural networks. In *International conference on machine learning*, pp. 4922–4931. PMLR, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pp. 991–999. PMLR, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.

Arezou Rezazadeh. A unified view on pac-bayes bounds for meta-learning. In *International Conference on Machine Learning*, pp. 18576–18595. PMLR, 2022.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850. PMLR, 2016.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pp. 5628–5637. PMLR, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. 48(4):1875–1897. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1875. URL https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-4/Nonparametric-regression-using-deep-neural-networks-with-ReLU-activation-function/10.1214/19-AOS1875.full.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and others. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Yue Sun, Adhyyan Narang, Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. *Advances in Neural Information Processing Systems*, 34:28156–28168, 2021.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=H1ebTsActm.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. *Advances in Neural Information Processing Systems*, 36:34536–34556, 2023.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, pp. 33416–33447. PMLR, 2023.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International conference on machine learning*, pp. 10434–10443. PMLR, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, and others. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Yunjuan Wang and Raman Arora. On the stability and generalization of meta-learning. *Advances in Neural Information Processing Systems*, 37:83665–83710, 2024.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Shiguang Wu, Yaqing Wang, Yatao Bian, and Quanming Yao. Learning to learn with contrastive meta-objective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. URL https://openreview.net/forum?id=s6YHno8Ke3.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94:103–114, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A Analyses

Additional notations We fix a common probability space $(\Omega, \mathcal{F}, \mu)$ and identify all random variables with measurable maps on it. For a random variable $X\colon \Omega \to \mathbb{R}$, we set $\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$ for $p \in [1,\infty)$ and $\|X\|_\infty = \inf\{C > 0 : \mathbb{P}[|X| \le C] = 1\}$. For a measurable function $f\colon \mathcal{X} \to \mathbb{R}$ and a random variable X on $\mathcal{X}$, we set $\|f\|_{L^\infty(X)} = \inf\{C > 0 : \mathbb{P}[|f(X)| \le C] = 1\}$. For a set A, $|A|$ denotes its cardinality; $\mathbb{1}$ denotes the indicator function. For real values a, b, define $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

We write $\alpha^*_{\mathcal{P}} = \sup_{\phi\in\Phi} \inf_{(X,Y)\in\mathcal{P}} \alpha(\phi; X, Y)$, $\alpha_{\mathcal{P}}(\phi) = \inf_{(X,Y)\in\mathcal{P}} \alpha(\phi; X, Y)$, and $\beta_{\mathcal{P}}(\phi) = \inf_{(X,Y)\in\mathcal{P}} \beta(\phi; X, Y)$.

Given a function $h : \mathcal{X} \to \mathbb{R}$ and $k \in \mathbb{N}$, define the empirical $L^p(X)$ norm of h for $p \in [1,\infty)$ and a random variable $X \in \mathcal{X}$ by

$$\|h\|^p_{L^p_k(X)} = \frac{1}{k}\sum_{i=1}^{k} h^p(X_i),$$

where $X_1, \ldots, X_k$ are i.i.d. copies of X. Given two random variables $X, Y \in \mathbb{R}$ and $k \in \mathbb{N}$, we use the empirical inner product and the empirical L^p norm of X for $p \in [1,\infty]$ defined as

$$\langle X, Y\rangle_k = \frac{1}{k}\sum_{i=1}^{k} X_i Y_i \text{ and } \|X\|^p_{L^p_k} = \frac{1}{k}\sum_{i=1}^{k} X_i^p,$$

where $(X_1, Y_1), \ldots, (X_k, Y_k)$ are i.i.d. copies of (X, Y). Let

$$B_{J,J'}(\phi; X, Y) = \|f^*_J \circ \phi - f^*_{J'} \circ \phi\|^2_{L^2(X)} \text{ and } \hat{B}_{J,J'}(\phi; X, Y) = \left\|\hat{f}_J \circ \phi - \hat{f}_{J'} \circ \phi\right\|^2_{L^2_m(X)},$$

where we refer to these quantities as bias terms. Given a feature extractor ϕ and $k \in \mathbb{N}$, the ideal and empirical best complexities are defined as

$$J^*_k(\phi) = \sup_{(X,Y)\in\mathcal{P}} J^*_k(\phi; X, Y) \text{ and } \hat{J}_k(\phi) = \max_{d\in[D]} \hat{J}_k(\phi; X^{(d)}, Y^{(d)}).$$

Here, $\hat{J}_k(\phi; X, Y)$ (two arguments) denotes the empirical Lepski complexity for a specific domain (X, Y) and sample size k, as in Thm. 5, while $\hat{J}_k(\phi)$ (one argument) is its worst-case value over the observed domains. The ideal feature extractor ϕ^* is such that $J^*_m(\phi^*) = \inf_{\phi\in\Phi} J^*_m(\phi)$ and the estimated extractor $\hat{\phi}$ satisfies $\hat{J}_m(\hat{\phi}) = \min_{\phi\in\Phi} \hat{J}_m(\phi)$, consistent with Eq. (5). We define the empirical counterpart of $f^*_{Y|X,\phi,J}$ as

$$\hat{f}_{Y|X,\phi,k,J} = \arg\min_{f_J\in\mathcal{F}_J} \frac{1}{k}\sum_{i=1}^{k} ((f_J \circ \phi)(X_i) - Y_i)^2,$$

where $(X_1, Y_1), \ldots, (X_k, Y_k)$ are i.i.d. copies of (X, Y). We also omit the first and second subscripts in $\hat{f}_{Y|X,\phi,k,J}$ if (X, Y) and ϕ are clear from the context.

Meta-learning and downstream learning analyses We analyze the proposed meta-learning and downstream learning algorithms and establish the performance guarantees. Specifically, we show the following two theorems corresponding to meta-learning and downstream learning, respectively.

Theorem 3 (Meta-learning performance guarantee)**.** *Assume asm. 1 to 4. Suppose that* $\ln^{1+\nu\gamma} m \lesssim D$ *and* $\ln D \lesssim \ln m$*. For any* $\mathcal{P} \in \mathfrak{P}$*, the feature extractor* $\hat{\phi}$ *in Eq. (5) satisfies that there exists a constant* $C > 0$ *such that with probability at least* $1 - O(m^{-1})$*,*

$$\beta_{\mathcal{P}}(\hat{\phi}) \geq \beta^*_{\mathcal{P}} - \frac{C}{\ln^{\gamma} m}.$$

Theorem 4 (Downstream learning performance guarantee)**.** *Let* $\mathcal{P} \in \mathfrak{P}$ *be the target distribution. Suppose that the learned feature extractor* $\hat{\phi}$ *is independent of the downstream sample. Assume Asm. 1. Then, the learned regressor* $\hat{f}_n$ *in Section 4.3 satisfies that with probability at least* $1 - n^{-1}$*,*

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim (n/\ln n)^{-\beta_{\mathcal{P}}(\hat{\phi})}.$$

Combining Thm. 3 and Thm. 4 immediately leads to the main theorem Thm. 2.

To prove thm. 3 and 4, we first derive an error bound on the best complexity estimator $\hat{J}_m$ with a fixed feature extractor ϕ in Appendix A.1. Then, we extend it to the learned feature extractor $\hat{\phi}$ in Appendix A.3. Building from these analyses, we prove thm. 3 and 4 in Appendix A.4 by appropriately handling the effect of finite observation of domains.

A.1 Best Complexity Estimation with Fixed Feature Extractor and Distribution

In this subsection, we analyze the best complexity estimator $\hat{J}_m(\phi; X, Y)$ with a fixed feature extractor ϕ and a fixed distribution $(X, Y) \in \mathcal{P}$. Specifically, we prove the following theorem.

Theorem 5. *Fix* ϕ *and* $(X, Y) \in \mathcal{P}$*. Suppose that* $A_J(\phi; X, Y) \asymp J^{-2\alpha}$ *for some* $\alpha > 0$*. Assume Asm. 4. If* $\rho > 0$ *is a sufficiently large constant, there exist constants* $c, C > 0$ *such that with probability at least* $1 - m^{-\Omega((m/\ln m)^{1/(2\alpha+1)})}$*,*

$$cJ^*_m(\phi; X, Y) \leq \hat{J}_m(\phi; X, Y) \text{ and } \hat{J}_m(\phi; X, Y) \leq CJ^*_m(\phi; X, Y),$$

or equivalently, $\hat{J}_m(\phi; X, Y) \asymp (m/\ln m)^{1/(2\alpha+1)}$*.*

Thm. 5 states that while the learner does not use the unknown parameter α in the complexity estimation, the resulting best complexity estimator $\hat{J}_m(\phi; X, Y)$ is equivalent to the ideal complexity depending on α up to multiplicative constants.

The key ingredient to prove Thm. 5 is the upper and lower bounds on $\hat{B}_{J,J'}$ via the approximation error $A_J(\phi; X, Y)$.

Theorem 6. *Fix* ϕ*,* $(X, Y) \in \mathcal{P}$*, and* $J, J' \in [m]$ *such that* $J' \geq J$*. Assume Asm. 4. Then, with probability at least* $1 - e^{-t}$*,*

$$\hat{B}_{J,J'}(\phi; X, Y) \lesssim A_J(\phi; X, Y) + V_{m,J'} + \frac{t}{m}.$$

Moreover, with probability at least $1 - e^{-t}$*,*

$$\hat{B}_{J,J'}(\phi; X, Y) \gtrsim A_J(\phi; X, Y) - A_{J'}(\phi; X, Y) - V_{m,J'} - \frac{t}{m}.$$

Thm. 6 shows that the empirical bias term $\hat{B}_{J,J'}$ is characterized dominantly by the approximation error A_J for a small complexity J. In particular, when $A_J \asymp J^{-2\alpha}$, the empirical comparison $\hat{B}_{J,J'}$ between regressors at complexity levels J and J' faithfully reflects the underlying approximation structure: it is large when A_J is large (meaning complexity J is insufficient) and small when $A_J \asymp V_{m,J'}$ (meaning the Lepski stopping criterion is met). This fidelity is what enables the Lepski method to identify the optimal complexity from data without prior knowledge of α.

We also use the following property of the approximation error.

Lemma 1. *For fixed ϕ and $(X,Y) \in \mathcal{P}$, assume $A_J(\phi; X, Y) \asymp J^{-2\alpha}$ for some $\alpha \in (0, \infty)$. Then, we have*

$$A_{J^*} \asymp V_{m,J^*} \asymp (m/\ln m)^{-2\alpha/(2\alpha+1)} \text{ and } J^* \asymp (m/\ln m)^{1/(2\alpha+1)},$$

where $J^* = J_m^*(\phi; X, Y)$.

Lem. 1 states that J^* achieves the ideal complexity of $(m/\ln m)^{1/(2\alpha+1)}$, and A_{J^*} and V_{m,J^*} are equivalent up to multiplicative constants.

Now, we prove Thm. 5.

Proof of Thm. 5. Let $J^* = J_m^*(\phi; X, Y)$, $\hat{J} = \hat{J}_m(\phi; X, Y)$, $\hat{B}_{J,J'} = \hat{B}_{J,J'}(\phi; X, Y)$, and $A_J = A_J(\phi; X, Y)$. Applying the union bound over $J, J' \in [m]$ into Thm. 6 gives that with probability at least $1 - e^{-t}$, for all $J, J' \in [m]$ where $J' \geq J$,

$$\hat{B}_{J,J'} \lesssim A_J + V_{m,J'} + \frac{t}{m}, \tag{6}$$

and

$$\hat{B}_{J,J'} \gtrsim A_J - A_{J'} - V_{m,J'} - \frac{t}{m}, \tag{7}$$

where we use $\frac{\ln m}{m} \lesssim V_{m,J} \lesssim V_{m,J'}$. Let $\mathcal{E}_t$ be this event; hence, $\mathbb{P}\{\mathcal{E}_t\} \geq 1 - e^{-t}$.

Upper bound. Let $J_{\text{up}} \in [m]$ such that $CJ^* \geq J_{\text{up}} \geq CJ^* - 1$ for some $C > 1$. Suppose that $\mathcal{E}_t$ occurs. Assume that $\hat{J} > J_{\text{up}}$. Then, J_{up} must fail the empirical Lepski condition; hence, for some $J' \geq J_{\text{up}}$, we have

$$\hat{B}_{J_{\text{up}},J'} > \rho V_{m,J'}.$$

By Eq. (6), there exist constants $C_1, C_2 > 0$ such that

$$\rho V_{m,J'} < C_1\big(A_{J_{\text{up}}} + V_{m,J'}\big) + C_2\frac{t}{m}.$$

By Lem. 1, we have

$$A_{J_{\text{up}}} \leq A_{J^*} \lesssim V_{m,J^*} \leq V_{m,J'}.$$

Hence, there are constants $C_1, C_2 > 0$ such that

$$\frac{C_2 t}{m} > (\rho - C_1) V_{m,J'},$$

where the sufficiently large ρ ensures $\rho - C_1 > 0$. The above inequality is contradicting if

$$t \leq \frac{m(\rho - C_1)V_{m,J'}}{C_2}.$$

For such t, $\hat{J} \leq J_{\text{up}}$ under the event $\mathcal{E}_t$. Choosing $t = \ln m \cdot \kappa(m/\ln m)V_{m,J^*}$ for $\kappa = (\rho - C_1)/C_2$ yields the contradictory t. Consequently, with probability at least $\mathbb{P}\{\mathcal{E}_t\} \geq 1 - m^{-\kappa(m/\ln m)V_{m,J^*}}$, we have $\hat{J} \leq J_{\text{up}} \leq CJ^*$.

Lower bound. Let $J_{\text{lo}} \in [m]$ such that $cJ^* + 1 \geq J_{\text{lo}} \geq cJ^*$ for some $c \in (0, 1)$. Suppose that $\mathcal{E}_t$ occurs. Assume that $\hat{J} < J_{\text{lo}}$. Then, since $\hat{J}$ satisfies the empirical Lepski condition, we have

$$\hat{B}_{\hat{J},J^*} \leq \rho V_{m,J^*}.$$

By Eq. (7), there exist constants $C_1, C_2, C_3 > 0$ such that

$$\rho V_{m,J^*} \geq C_1 A_{\hat{J}} - C_2(A_{J^*} + V_{m,J^*}) - \frac{C_3 t}{m}.$$

By the assumption of $A_J \asymp J^{-2\alpha}$ and Lem. 1, we have

$$A_{\hat{J}} \gtrsim c^{-2\alpha} A_{J^*} \gtrsim c^{-2\alpha} V_{m,J^*},$$

and $A_{J^*} \lesssim V_{m,J^*}$. Hence, there exist constants $C_1, C_2, C_3 > 0$ such that

$$\frac{C_3 t}{m} \geq \big(C_1 c^{-2\alpha} - C_2 - \rho\big) V_{m,J^*}.$$

A sufficiently small c ensures $C_1 c^{-2\alpha} - C_2 - \rho > 0$. The above inequality is contradicting if

$$t \leq \frac{m\big(C_1 c^{-2\alpha} - C_2 - \rho\big) V_{m,J^*}}{C_3}.$$

For such t, $\hat{J} \geq J_{\mathrm{lo}}$ under the event $\mathcal{E}_t$. Choosing $t = \ln m \cdot \kappa (m/\ln m) V_{m,J^*}$ for $\kappa = (C_1 c^{-2\alpha} - C_2 - \rho)/C_3$ yields the contradictory t. Consequently, with probability at least $\mathbb{P}\{\mathcal{E}_t\} \geq 1 - m^{-\kappa (m/\ln m) V_{m,J^*}}$, we have $\hat{J} \geq J_{\mathrm{lo}} \geq cJ^*$.

By Lem. 1, we have $(m/\ln m) V_{m,J^*} \gtrsim (m/\ln m)^{1/(2\alpha+1)}$, which gives the claim. □

A.2 Bias Analysis with Fixed Feature Extractor

In this subsection, we investigate the empirical bias term $\hat{B}_{J,J'}$ to prove Thm. 6. To derive bounds on $\hat{B}_{J,J'}$, we first derive the error upper and lower bounds on the sieved least-square estimator $\hat{f}_{m,J}$.

Sieved least-square estimator

Lemma 2. *Fix ϕ, J, and $(X,Y) \in \mathcal{P}$. Assume Asm. 4. Then, there exists universal constants $c_1, c_2 > 0$ such that for any $\eta \in (0,1)$, with probability at least $1 - 3e^{-t}$,*

$$\left\| \hat{f}_J \circ \phi - \mathbb{E}[Y|X] \right\|^2_{L^2_m(X)} \leq \frac{1}{1-\eta} \left(2(1+\eta) A_J(\phi; X, Y) + \frac{6\sigma^2 c_2 J \ln m}{\eta m} + \left(2\sqrt{2\ln(2)}\sigma c_1 + \frac{\eta c_1^2}{m} \right) \frac{1}{m} + \left(\frac{2\sqrt{2}\sigma c_1}{\ln^{1/2}(2)} + \frac{6\sigma^2}{\eta} + \frac{5(1+\eta)}{6} \right) \frac{t}{m} \right).$$

Lemma 3. *Fix ϕ, J, and $(X,Y) \in \mathcal{P}$. Assume Asm. 4. Then, there exists universal constants $c_1, c_2 > 0$ such that with probability at least $1 - e^{-t}$,*

$$\left\| \hat{f}_J \circ \phi - \mathbb{E}[Y|X] \right\|^2_{L^2_m(X)} \geq \frac{1}{4} A_J(\phi; X, Y) - \frac{2c_2 J \ln m}{3m} - \frac{3c_1^2}{2m^2} - \frac{2t}{3m}.$$

Remark 2. *Lem. 2 implies that, with probability at least $1 - e^{-t}$,*

$$\left\| \hat{f}_J \circ \phi - \mathbb{E}[Y|X] \right\|^2_{L^2_m(X)} \lesssim A_J(\phi; X, Y) + V_{m,J} + \frac{t}{m}.$$

Also, Lem. 3 implies that, with probability at least $1 - e^{-t}$,

$$\left\| \hat{f}_J \circ \phi - \mathbb{E}[Y|X] \right\|^2_{L^2_m(X)} \gtrsim A_J(\phi; X, Y) - V_{m,J} - \frac{t}{m}.$$

Now, we prove Thm. 6.

Proof of Thm. 6. For shorthands, let $\hat{B}_{J,J'} = \hat{B}_{J,J'}(\phi; X, Y)$ and $A_J = A_J(\phi; X, Y)$. By the triangle and reverse triangle inequalities, we have

$$\left(\left\| \hat{f}_{m,J} \circ \phi - \mathbb{E}[Y|X] \right\|_{L^2_m(X)} - \left\| \hat{f}_{m,J'} \circ \phi - \mathbb{E}[Y|X] \right\|_{L^2_m(X)} \right)^2 \leq \hat{B}_{J,J'} \leq \left(\left\| \hat{f}_{m,J} \circ \phi - \mathbb{E}[Y|X] \right\|_{L^2_m(X)} + \left\| \hat{f}_{m,J'} \circ \phi - \mathbb{E}[Y|X] \right\|_{L^2_m(X)} \right)^2. \quad (8)$$

We now prove the upper and lower bounds separately.

Upper bound Application of Lem. 2 to both terms in the right-hand side of Eq. (8) yields that with probability at least $1-2e^{-t}$,

$$\left\|\hat{f}_{m,J}\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\lesssim A_J+V_{m,J}+\frac{t}{m}$$
$$\left\|\hat{f}_{m,J'}\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\lesssim A_{J'}+V_{m,J'}+\frac{t}{m}.$$

Noting that A_J is decreasing in J, $V_{m,J}$ is increasing in J, and $V_{m,J}\gtrsim\frac{1}{m}$, we have with probability at least $1-e^{-t}$,

$$\hat{B}_{J,J'}\lesssim A_J+V_{m,J'}+\frac{t}{m}.$$

Lower bound Respectively applying Lem. 3 and Lem. 2 into the first and second terms of the left-hand side of Eq. (8) gives that with probability at least $1-2e^{-t}$,

$$\left\|\hat{f}_{m,J}\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\gtrsim A_J-V_{m,J}-\frac{t}{m}$$
$$\left\|\hat{f}_{m,J'}\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\lesssim A_{J'}+V_{m,J'}+\frac{t}{m}.$$

Again, using the facts that A_J is decreasing in J, $V_{m,J}$ is increasing in J, and $V_{m,J}\gtrsim\frac{1}{m}$, we have with probability at least $1-e^{-t}$,

$$\hat{B}_{J,J'}\gtrsim A_J-A_{J'}-V_{m,J'}-\frac{t}{m}.$$

□

A.3 Best Complexity Estimator with Learned Feature Extractor

In this subsection, we prove Thm. 3. To this end, we extend Thm. 5 for the learned feature extractor $\hat{\phi}$. Specifically, we prove the following theorem.

Theorem 7. *Fix $(X,Y)\in\mathcal{P}$. Let $\hat{\phi}$ be the learned feature extractor depending on the pre-training samples. Assume asm. 1 and 2. If $\rho>0$ is a sufficiently large constant, there exist constants $c,C>0$ such that for some $\epsilon>0$, with probability at least $1-\sum_{J,J'\in[m]:J'\geq J}N(V_{m,J},\Phi,\rho_J)N(V_{m,J'},\Phi,\rho_{J'})m^{-\Omega((m/\ln m)^{1/(2\bar{\alpha}+1)})}$,*

$$cJ^*_m(\hat{\phi};X,Y)\leq\hat{J}_m(\hat{\phi};X,Y)\ \text{and}\ \hat{J}_m(\hat{\phi};X,Y)\leq CJ^*_m(\hat{\phi};X,Y),$$

or equivalently, $\hat{J}_m(\hat{\phi};X,Y)\asymp(m/\ln m)^{1/(2\alpha(\hat{\phi};X,Y)+1)}$.

Thm. 7 is an analogue of Thm. 5 with learned $\hat{\phi}$.

Following the analyses with the fixed ϕ case, we derive bounds on the bias terms with $\hat{\phi}$ to prove Thm. 7.

Theorem 8. *Fix ϕ, $(X,Y)\in\mathcal{P}$, and $J,J'\in[m]$ such that $J'\geq J$. Let $\hat{\phi}$ be the learned feature extractor depending on the pre-training samples. Assume asm. 2 and 4. Then, with probability at least $1-N(V_{m,J},\Phi,\rho_J)N(V_{m,J'},\Phi,\rho_{J'})e^{-t}$,*

$$\hat{B}_{J,J'}(\hat{\phi};X,Y)\lesssim A_J(\hat{\phi};X,Y)+V_{m,J'}+\frac{t}{m}.$$

Moreover, with probability at least $1-N(V_{m,J},\Phi,\rho_J)N(V_{m,J'},\Phi,\rho_{J'})e^{-t}$,

$$\hat{B}_{J,J'}(\hat{\phi};X,Y)\gtrsim A_J(\hat{\phi};X,Y)-A_{J'}(\hat{\phi};X,Y)-V_{m,J'}-\frac{t}{m}.$$

Based on Thm. 8, we prove Thm. 7.

Proof of Thm. 7. We invoke the same proof of Thm. 5 except leveraging Thm. 8 and choosing t as

$$t=\ln m\cdot\kappa(m/\ln m)^{1/(2\bar{\alpha}+1)},$$

where κ is an appropriate constant leading to the contradictory t. Noting that $m/\ln m>1$, such a choice of t gives contradictory t for any $\hat{\phi}$ and $(X,Y)\in\mathcal{P}^*$. □

A.4 Meta-Learning and Downstream Learning Analyses

Now, we prove Thm. 3.

Proof of Thm. 3. By the definition of $\hat{\phi}$, we have

$$\max_{d\in[D]} \hat{J}_m(\hat{\phi}; X^{(d)}, Y^{(d)}) \leq \max_{d\in[D]} \hat{J}_m(\phi^*; X^{(d)}, Y^{(d)}).$$

Application of the union bound into Thm. 7 over $d \in [D]$ yields that with probability at least $1 - D\sum_{J,J'\in[m]:J'\geq J} N(V_{m,J}, \Phi, \rho_J)N(V_{m,J'}, \Phi, \rho_{J'})m^{-\Omega((m/\ln m)^{1/(2\tilde{\alpha}+1)})}$,

$$\max_{d\in[D]} (m/\ln m)^{1/(2\alpha(\hat{\phi};X^{(d)},Y^{(d)})+1)} \lesssim \max_{d\in[D]} J^*_m(\hat{\phi}; X^{(d)}, Y^{(d)}) \lesssim \max_{d\in[D]} \hat{J}_m(\hat{\phi}; X^{(d)}, Y^{(d)})$$

Also, applying the union bound to Thm. 5 with $\phi = \phi^*$ over $d \in [D]$ gives that with probability at least $1 - Dm^{-\Omega((m/\ln m)^{1/(2\tilde{\alpha}+1)})}$,

$$\max_{d\in[D]} \hat{J}_m(\phi^*; X^{(d)}, Y^{(d)}) \lesssim \max_{d\in[D]} J^*_m(\phi^*; X^{(d)}, Y^{(d)}) \leq J^*_m(\phi^*) \lesssim (m/\ln m)^{1/(2\alpha^*_{\mathcal{P}}+1)}.$$

Note that $1/(2\alpha(\phi; X, Y) + 1) = 1 - \beta(\phi; X, Y)$. Consequently, there exists a constant $C > 1$ such that with probability at least $1 - 2D\sum_{J,J'\in[m]:J'\geq J} N(V_{m,J}, \Phi, \rho_J)N(V_{m,J'}, \Phi, \rho_{J'})m^{-\Omega((m/\ln m)^{1/(2\tilde{\alpha}+1)})}$,

$$(m/\ln m)^{1-\min_{d\in[D]}\beta(\hat{\phi};X^{(d)},Y^{(d)})} \leq C(m/\ln m)^{1-\beta^*_{\mathcal{P}}}.$$

Taking the logarithm of both sides and dividing by $\ln(m/\ln m)$, we have

$$\min_{d\in[D]} \beta(\hat{\phi}; X^{(d)}, Y^{(d)}) \geq \beta^*_{\mathcal{P}} - \frac{\ln C}{\ln(m/\ln m)}.$$

Fix an arbitrary $\phi \in \Phi$. Let $(X^{[1]}, Y^{[1]}), \ldots, (X^{[\lceil 1/\epsilon\rceil]}, Y^{[\lceil 1/\epsilon\rceil]})$ be such that for any $(X, Y) \in \mathcal{P}$, there exists i satisfying $|\beta(\phi; X^{[i]}, Y^{[i]}) - \beta(\phi; X, Y)| \leq \epsilon$. By Asm. 3, there exists $d \in [D]$ such that $|\min_i \beta(\phi; X^{[i]}, Y^{(i)}) - \beta(\phi; X^{(d)}, Y^{(d)})| \leq \epsilon$ with probability at least $1-(1-C\epsilon^\nu)^D$ for some universal constant $C > 0$. Hence, with probability at least $1 - (1 - C\epsilon^\nu)^D$,

$$\min_{d\in[D]} \beta(\phi; X^{(d)}, Y^{(d)}) \leq \inf_{(X,Y)\in\mathcal{P}} \beta(\phi; X, Y) + 2\epsilon.$$

For $t > 0$, taking $\epsilon^\nu = t/CD$ yields with probability at least $1 - e^{-t}$,

$$\min_{d\in[D]} \beta(\phi; X^{(d)}, Y^{(d)}) \leq \inf_{(X,Y)\in\mathcal{P}} \beta(\phi; X, Y) + 2\left(\frac{t}{CD}\right)^\nu.$$

Consider $\ln^{-\gamma} m$-cover of Φ in Asm. 2, denoted as $\phi_1, \ldots, \phi_{N_m}$, where $N_m = N(\ln^{-\gamma} m, \Phi, \rho_{\beta,\mathcal{P}})$. Let $\hat{\phi}_m$ be the closest ϕ_i to $\hat{\phi}$ in terms of $\rho_{\beta,\mathcal{P}}$. Then, we have with probability at least $1 - N_m e^{-t}$,

$$\begin{aligned}
\min_{d\in[D]} \beta(\hat{\phi}; X^{(d)}, Y^{(d)}) &\leq \min_{d\in[D]} \beta(\hat{\phi}_m; X^{(d)}, Y^{(d)}) + \frac{1}{\ln^\gamma m} \\
&\leq \inf_{(X,Y)\in\mathcal{P}} \beta(\hat{\phi}_m; X, Y) + 2\left(\frac{t}{CD}\right)^\nu + \frac{1}{\ln^\gamma m} \\
&\leq \beta_{\mathcal{P}}(\hat{\phi}) + 2\left(\frac{t}{CD}\right)^\nu + \frac{2}{\ln^\gamma m}.
\end{aligned}$$

By Asm. 2 and $D \gtrsim \ln^{1+\nu\gamma} m$, with $t = \ln N_m + \ln m$, we have with probability at least $1 - m^{-1}$,

$$\min_{d\in[D]} \beta(\hat{\phi}; X^{(d)}, Y^{(d)}) \leq \beta_{\mathcal{P}}(\hat{\phi}) + O\left(\frac{1}{\ln^\gamma m}\right).$$

By the union bound, we have with probability at least $1 - m^{-1} - 2m^{O(1)} \sum_{J,J' \in [m]: J' \geq J} N(V_{m,J}, \Phi, \rho_J) N(V_{m,J'}, \Phi, \rho_{J'}) m^{-\Omega((m/\ln m)^{1/(2\bar{\alpha}+1)})}$,

$$\beta_{\mathcal{P}}(\hat{\phi}) \geq \beta^*_{\mathcal{P}} - O\left(\frac{1}{\ln^{\gamma} m}\right).$$

By Asm. 2, we have

$$\sum_{J,J' \in [m]: J' \geq J} N(V_{m,J}, \Phi, \rho_J) N(V_{m,J'}, \Phi, \rho_{J'}) m^{-\Omega((m/\ln m)^{1/(2\bar{\alpha}+1)})} \lesssim m^{-\Omega((m/\ln m)^{1/(2\bar{\alpha}+1)})} \lesssim m^{-1},$$

which gives the desired result. □

Next, we prove Thm. 4.

Proof of Thm. 4. Let $J^* = J^*_n(\hat{\phi}; X, Y)$ and $\hat{J} = \hat{J}_n(\hat{\phi}; X, Y)$. Let $\hat{f}'_{n,J}$ be the closest function among $O(1/n)$-net of $\mathcal{F}_J$ to $\hat{f}_{n,J}$ in L^∞-norm. Application of the union bound over $O(1/n)$-net and $\hat{J}$ and Lem. 7 yields that with probability at least $1 - e^{-t}$,

$$\begin{aligned}
\bar{E}_{\hat{\phi}}(\hat{f}_n) &= \left\| \hat{f}_{n,\hat{J}} \circ \hat{\phi} - \mathbb{E}[Y|X] \right\|^2_{L^2(X)} \\
&\lesssim \left\| \hat{f}'_{n,\hat{J}} \circ \hat{\phi} - \mathbb{E}[Y|X] \right\|^2_{L^2(X)} + \frac{1}{n^2} \\
&\lesssim \left\| \hat{f}'_{n,\hat{J}} \circ \hat{\phi} - \mathbb{E}[Y|X] \right\|^2_{L^2_n(X)} \\
&\quad + \sqrt{V_{n,\hat{J}} + \frac{\ln n}{n} + \frac{t}{n}} \left\| \hat{f}'_{n,\hat{J}} \circ \hat{\phi} - \mathbb{E}[Y|X] \right\|_{L^2(X)} + V_{n,\hat{J}} + \frac{\ln n}{n} + \frac{t}{n}.
\end{aligned}$$

By AM-GM inequality, with probability at least $1 - e^{-t}$,

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim \left\| \hat{f}_{n,\hat{J}} \circ \hat{\phi} - \mathbb{E}[Y|X] \right\|^2_{L^2_n(X)} + V_{n,\hat{J}} + \frac{t}{n}.$$

If $\hat{J} \leq J^*$, the empirical Lepski's rule ensures that

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim \left\| \hat{f}_{n,J^*} \circ \hat{\phi} - \mathbb{E}[Y|X] \right\|^2_{L^2_n(X)} + V_{n,J^*} + \frac{t}{n}.$$

By Lem. 2, we get with probability at least $1 - e^{-t}$,

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim A_{J^*}(\hat{\phi}; X, Y) + V_{n,J^*} + \frac{t}{n}.$$

From Lem. 1, we have with probability at least $1 - e^{-t}$,

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim (n/\ln n)^{-\beta(\hat{\phi};X,Y)} + \frac{t}{n}.$$

If $\hat{J} > J^*$, there exists $J' \geq J^*$ such that J^* fails the empirical Lepski condition, i.e.,

$$\hat{B}_{J^*,J'}(\hat{\phi}; X, Y) > \rho V_{n,J'}.$$

By Thm. 6 (upper bound), with probability at least $1 - e^{-t}$,

$$\hat{B}_{J^*,J'}(\hat{\phi}; X, Y) \lesssim A_{J^*}(\hat{\phi}; X, Y) + V_{n,J'} + \frac{t}{n}.$$

From Lem. 1, $A_{J^*} \lesssim V_{n,J^*} \leq V_{n,J'}$, so there exist constants $C_1, C_2 > 0$ such that

$$\rho V_{n,J'} < C_1 V_{n,J'} + C_2 \frac{t}{n}.$$

Rearranging gives $(\rho - C_1)V_{n,J'} < C_2 t/n$. Choosing $\rho > C_1 + C_2$ and $t = \ln n$, this is contradicted since $V_{n,J'} \geq V_{n,1} = \ln(n)/n$ implies $(\rho - C_1)\ln(n)/n \leq C_2 \ln(n)/n$. Hence, by choosing ρ sufficiently large, the event $\hat{J} > J^*$ occurs with probability at most $e^{-t} = 1/n$. Combining both cases with a union bound, we have with probability at least $1 - 2/n$,

$$\bar{E}_{\hat{\phi}}(\hat{f}_n) \lesssim (n/\ln n)^{-\beta_{\mathcal{P}}(\hat{\phi})} + \frac{\ln n}{n} \lesssim (n/\ln n)^{-\beta_{\mathcal{P}}(\hat{\phi})},$$

where the last step uses $\beta_{\mathcal{P}}(\hat{\phi}) \in (0,1)$. □

A.5 Bias Analysis with Learned Feature Extractor

Here, we analyze $\hat{B}_{J,J'}$ for the learned feature extractor $\hat{\phi}$. We follow similar steps of the analyses in Appendix A.2 but introduce an approximation and the union bound due to the covering of Φ. First, we reveal the error upper and lower bounds on the sieved least-square estimator.

Lemma 4. *Fix $(X,Y) \in \mathcal{P}$, $J \in [m]$, and $\epsilon > 0$. Let $\hat{\phi}$ be the learned feature extractor depending on the pre-training samples. Assume asm. 2 and 4. Then, there exist universal constants $c_1, c_2 > 0$ such that for any $\eta \in (0,1)$, with probability at least $1 - (1 + 2N(\epsilon, \Phi, \rho_J))e^{-t}$,*

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \leq \frac{1}{1-\eta}\left(4(2+3\eta)A_J(\hat{\phi}; X, Y) + \frac{6\sigma^2 c_2 J \ln m}{\eta m}\right.$$
$$\left. + \left(2\sqrt{2\ln(2)}\sigma c_1 + \frac{3\eta c_1^2}{m}\right)\frac{1}{m} + (4+3\eta)\epsilon + \left(\frac{2\sqrt{2}\sigma c_1}{\ln^{1/2}(2)} + \frac{6\sigma^2}{\eta} + \frac{5(1+\eta)}{6}\right)\frac{t}{m}\right).$$

Lemma 5. *Fix $(X,Y) \in \mathcal{P}$, $J \in [m]$, and $\epsilon > 0$. Let $\hat{\phi}$ be the learned feature extractor depending on the pre-training samples. Assume asm. 1 and 2. Then, there exist universal constants $c_1, c_2 > 0$ such that with probability at least $1 - N(\epsilon, \Phi, \rho_J)e^{-t}$,*

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{4}A_J(\hat{\phi}; X, Y) - \frac{2c_2 J \ln m}{3m} - \frac{3\epsilon^2}{2} - \frac{3c_1^2}{2m^2} - \frac{2t}{3m}.$$

We now give a proof of Thm. 8.

Proof of Thm. 8. We follow the proof of Thm. 6 with lem. 4 and 5. For shorthands, let $\hat{B}_{J,J'} = \hat{B}_{J,J'}(\hat{\phi}; X, Y)$ and $A_J = A_J(\hat{\phi}; X, Y)$. By the triangle and reverse triangle inequalities, we have

$$\left(\left\|\hat{f}_{m,J} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|_{L^2_m(X)} - \left\|\hat{f}_{m,J'} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|_{L^2_m(X)}\right)^2 \leq \hat{B}_{J,J'} \leq$$
$$\left(\left\|\hat{f}_{m,J} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|_{L^2_m(X)} + \left\|\hat{f}_{m,J'} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|_{L^2_m(X)}\right)^2. \quad (9)$$

We now prove the upper and lower bounds separately.

Upper bound Application of Lem. 4 to both terms in the right-hand side of Eq. (9) yields that with probability at least $1 - N(V_{m,J}, \Phi, \rho_J)N(V_{m,J'}, \Phi, \rho_{J'})e^{-t}$,

$$\left\|\hat{f}_{m,J} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \lesssim A_J + V_{m,J} + \frac{t}{m}$$
$$\left\|\hat{f}_{m,J'} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \lesssim A_{J'} + V_{m,J'} + \frac{t}{m}.$$

Noting that A_J is decreasing in J, $V_{m,J}$ is increasing in J, and $V_{m,J} \gtrsim \frac{1}{m}$, we have with probability at least $1 - N(V_{m,J}, \Phi, \rho_J)N(V_{m,J'}, \Phi, \rho_{J'})e^{-t}$,

$$\hat{B}_{J,J'} \lesssim A_J + V_{m,J'} + \frac{t}{m}.$$

Lower bound Respectively applying Lem. 5 and Lem. 4 into the first and second terms of the left-hand side of Eq. (9) gives that with probability at least $1 - N(V_{m,J}, \Phi, \rho_J)N(V_{m,J'}, \Phi, \rho_{J'})e^{-t}$,

$$\left\|\hat{f}_{m,J} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \gtrsim A_J - V_{m,J} - \frac{t}{m}$$
$$\left\|\hat{f}_{m,J'} \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \lesssim A_{J'} + V_{m,J'} + \frac{t}{m}.$$

Again, using the facts that A_J is decreasing in J, $V_{m,J}$ is increasing in J, and $V_{m,J} \gtrsim \frac{1}{m}$, we have with probability at least $1 - N(V_{m,J}, \Phi, \rho_J)N(V_{m,J'}, \Phi, \rho_{J'})e^{-t}$,

$$\hat{B}_{J,J'} \gtrsim A_J - A_{J'} - V_{m,J'} - \frac{t}{m}.$$

□

B Proofs of Analyses

B.1 Proofs for Bias Analyses

Properties of Lepski's method To prove Lem. 1, we leverage the following lemma.

Lemma 6. *Fix ϕ and $(X,Y) \in \mathcal{P}$. Assume that $A_J(\phi; X, Y) \asymp J^{-2\alpha}$. For any constant $c > 1$, there exists $C > 0$ such that for any $J, J' \in \mathbb{N}$ where $J' \geq CJ$,*

$$B_{J,J'}(\phi; X, Y) \leq cA_J(\phi; X, Y).$$

Furthermore, for $c \geq 4$, this inequality is satisfied with $C = 1$. Moreover, there exists a constant $c > 0$ such that for any constant $C > 0$ and for any $J, J' \in \mathbb{N}$ where $J' \geq CJ$,

$$B_{J,J'}(\phi; X, Y) \geq \left(1 - cC^{-\alpha}\right)A_J(\phi; X, Y).$$

Proof of Lem. 6. We use the shorthands $B_{J,J'} = B_{J,J'}(\phi; X, Y)$ and $A_J = A_J(\phi; X, Y)$.

Upper bound. By the triangle inequality, for any $J, J' \in \mathbb{N}$,

$$B_{J,J'} \leq A_J\left(1 + \frac{A_{J'}^{1/2}}{A_J^{1/2}}\right)^2.$$

By the assumption of $A_J \asymp J^{-2\alpha}$, there exists a constant $C' > 0$ such that

$$\frac{A_{J'}^{1/2}}{A_J^{1/2}} \leq C'\left(\frac{J}{J'}\right)^{\alpha} \leq C'C^{-\alpha}.$$

Since $c > 1$, we can choose $C > 0$ such that $1 + C'C^{-\alpha} \leq \sqrt{c}$, confirming the upper bound. The further statement follows from the fact that for any $J' \geq J$, $A_{J'} \leq A_J$.

Lower bound. By the reverse triangle inequality, for any $J, J' \in \mathbb{N}$,

$$B_{J,J'} \geq A_J\left(1 - \frac{A_{J'}^{1/2}}{A_J^{1/2}}\right)^2.$$

By the same argument as the upper bound, there exists a constant $c > 0$ such that

$$B_{J,J'} \geq A_J\left(1 - cC^{-\alpha}/2\right)^2.$$

For $C > 0$ such that $1 - cC^{-\alpha}/2 \in (0,1)$, we have $(1 - cC^{-\alpha}/2)^2 \leq 1 - cC^{-\alpha}$, confirming the lower bound. □

Now, we prove Lem. 1.

Proof of Lem. 1. Write A_J for $A_J(\phi; X, Y)$ throughout. By assumption, there exist constants $0 < c_\ell \le c_u < \infty$ such that

$$c_\ell J^{-2\alpha} \le A_J \le c_u J^{-2\alpha} \text{ for all } J \in \mathbb{N}.$$

Define $J_0 = (m/\ln m)^{1/(2\alpha+1)}$; a direct computation gives $V_{m,J_0} = J_0 \ln(m)/m = J_0^{-2\alpha}$. Hence, with this J_0, we have $V_{m,J_0} \asymp A_{J_0}$ due to assumption.

By Lem. 6, for some constant $c \ge 4$, for any $J' \ge J$,

$$B_{J,J'} \le cA_J \le cc_u J^{-2\alpha}.$$

We can choose J such that $cc_u J^{-2\alpha} \le J_0^{-2\alpha} = V_{m,J_0}$, satisfying the Lepski condition. Hence, for such a J, we have $J_m^*(\phi; X, Y) \le J \lesssim J_0$.

By Lem. 6, for some constant $c \in (0, 1)$, there exists $C > 0$ such that for any $J, J' \in \mathbb{N}$ where $J' \ge CJ$,

$$B_{J,J'} \ge cA_J \ge cc_\ell J^{-2\alpha}.$$

We can choose J such that $cc_\ell J^{-2\alpha} \ge J_0^{-2\alpha} = V_{m,J_0}$, breaking the Lepski condition. Hence, for such a J, we have $J_m^*(\phi; X, Y) \ge J \gtrsim J_0$. □

Proofs for sieved least square estimator We use three concentration inequalities for the empirical L^2-norm $\|\cdot\|^2_{L^2_k(X)}$, the empirical inner product to the noise $\langle \cdot, \epsilon \rangle_k$, and the absolute sum of the noise $\frac{1}{k}\sum_{i=1}^k |\epsilon_i|$.

Lemma 7. *Let X be a random variable on $\mathcal{X}$. For any fixed measurable function $h\colon \mathcal{X} \to [-1, 1]$ and any $t > 0$, with probability at least $1 - e^{-t}$,*

$$\|h\|^2_{L^2_k(X)} - \|h\|^2_{L^2(X)} \le \sqrt{\frac{2t}{k}\|h\|^2_{L^2(X)}} + \frac{t}{3k}.$$

Moreover, with probability at least $1 - e^{-t}$,

$$\|h\|^2_{L^2(X)} - \|h\|^2_{L^2_k(X)} \le \sqrt{\frac{2t}{k}\|h\|^2_{L^2(X)}} + \frac{t}{3k}.$$

Lemma 8. *Let $X \in \mathbb{R}$ be a random variable. Let ϵ be sub-gaussian and mean-zero independent random variables such that their variance proxy is at most $\sigma^2 > 0$. Then, conditioned on X, with probability at least $1 - e^{-t}$,*

$$\langle X, \epsilon \rangle_k \le \sqrt{\frac{2\sigma^2 t}{k}} \|X\|_{L^2_k}.$$

Lemma 9. *Let $\epsilon_1, \ldots, \epsilon_k$ be sub-gaussian and mean-zero independent random variables such that their variance proxy is at most $\sigma^2 > 0$. Then, with probability at least $1 - e^{-t}$,*

$$\frac{1}{k}\sum_{i=1}^k |\epsilon_i| \le \sigma\left(\frac{2t}{k} + 2\ln 2\right)^{1/2}.$$

From the definition of the Minkowski–Bouligand dimension, for $c_2 > 1$, there is a sufficiently small $c_1 > 0$ such that with $\epsilon_m = \frac{c_1}{m}$, $\ln N(\epsilon_m, \mathcal{F}_J, \|\cdot\|_{L^\infty}) \le c_2 J \ln(m)$ for any $m \ge 1$. Let $N_{m,J} = N(\epsilon_m, \mathcal{F}_J, \|\cdot\|_{L^\infty})$ and $f_1, ..., f_{N_{m,J}}$ be an ϵ_m-cover of $\mathcal{F}_J$ in $\|\cdot\|_{L^\infty}$. Let $\hat{f}_{J,\epsilon_m}$ be the closest function from these to $\hat{f}_J$ in $\|\cdot\|_{L^\infty}$.

We now prove lem. 2 and 3.

Proof of Lem. 2. We can decompose the squared error as

$$\left\|\hat{f}_J \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} = \|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2_m(X)}$$

$$+\frac{2}{m}\sum_{i=1}^{m}\Big((\hat{f}_J\circ\phi)(X_i)-(f_J\circ\phi)(X_i)\Big)\epsilon_i$$
$$+\frac{1}{m}\sum_{i=1}^{m}\Big((\hat{f}_J\circ\phi)(X_i)-Y_i\Big)^2-\frac{1}{m}\sum_{i=1}^{m}((f_J\circ\phi)(X_i)-Y_i)^2,$$

where $\epsilon_i = Y_i - \mathbb{E}[Y|X]$. Since $\hat{f}_J$ is an empirical minimizer over $\mathcal{F}_J$ and $f_J \in \mathcal{F}_J$, we have

$$\left\|\hat{f}_J\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\le$$
$$\|f_J\circ\phi-\mathbb{E}[Y|X]\|^2_{L^2_m(X)}+\frac{2}{m}\sum_{i=1}^{m}\Big((\hat{f}_J\circ\phi)(X_i)-(f_J\circ\phi)(X_i)\Big)\epsilon_i.$$

For the second term in the right-hand side, we have

$$\frac{1}{m}\sum_{i=1}^{m}\Big((\hat{f}_J\circ\phi)(X_i)-(f_J\circ\phi)(X_i)\Big)\epsilon_i\le$$
$$\frac{\epsilon_m}{m}\sum_{i=1}^{m}|\epsilon_i|+\frac{1}{m}\sum_{i=1}^{m}\Big((\hat{f}_{J,\epsilon_m}\circ\phi)(X_i)-(f_J\circ\phi)(X_i)\Big)\epsilon_i.$$

By Lem. 9, Lem. 8, union bound over the ϵ_m-covers, and triangle inequality, we have with probability at least $1-2e^{-t}$,

$$\frac{\epsilon_m}{m}\sum_{i=1}^{m}|\epsilon_i|\le\sigma\epsilon_m\left(\frac{2t}{m}+2\ln 2\right)^{1/2},$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\Big((\hat{f}_{J,\epsilon_m}\circ\phi)(X_i)-(f_J\circ\phi)(X_i)\Big)\epsilon_i\le$$
$$\sqrt{\frac{2\sigma^2(\ln(N_{m,J})+t)}{m}}\left(\left\|\hat{f}_J\circ\phi-f_J\circ\phi\right\|_{L^2_m(X)}+\epsilon_m\right).$$

Consequently, we have with probability at least $1-2e^{-t}$,

$$\left\|\hat{f}_J\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\le$$
$$\|f_J\circ\phi-\mathbb{E}[Y|X]\|^2_{L^2_m(X)}+2\sigma\epsilon_m\left(\frac{2t}{m}+2\ln 2\right)^{1/2}$$
$$+\sqrt{\frac{8\sigma^2(c_2J\ln(m)+t)}{m}}\left(\left\|\hat{f}_J\circ\phi-f_J\circ\phi\right\|_{L^2_m(X)}+\epsilon_m\right).$$

By the triangle inequality, the AM-GM inequality, and the fact that $\sqrt{1+x}\le 1+x$ for $x\ge 0$, for any $\eta\in(0,1)$, we have

$$(1-\eta)\left\|\hat{f}_J\circ\phi-\mathbb{E}[Y|X]\right\|^2_{L^2_m(X)}\le$$
$$(1+\eta)\|f_J\circ\phi-\mathbb{E}[Y|X]\|^2_{L^2_m(X)}+\frac{6\sigma^2c_2J\ln(m)}{\eta m}$$
$$+\left(2\sqrt{2\ln(2)}\sigma c_1+\frac{\eta c_1^2}{m}\right)\frac{1}{m}+\left(\frac{2\sqrt{2}\sigma c_1}{\ln^{1/2}(2)}+\frac{6\sigma^2}{\eta}\right)\frac{t}{m}. \quad (10)$$

Application of Lem. 7 to the first term in the right-hand side of Eq. (10) yields with probability at least $1-e^{-t}$,

$$\|f_J\circ\phi-\mathbb{E}[Y|X]\|^2_{L^2_m(X)}\le$$

$$\|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2(X)} + \sqrt{\frac{2t}{m}\|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2(X)}} + \frac{t}{3m}.$$

The AM-GM inequality $\sqrt{2tv/m} \leq v + t/(2m)$ yields with probability at least $1 - e^{-t}$,

$$\|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2_m(X)} \leq 2\|f_J \circ \phi - \mathbb{E}[Y|X]\|^2_{L^2(X)} + \frac{5t}{6m}. \tag{11}$$

Combining Eq. (10) and Eq. (11) yields the claim. □

Proof of Lem. 3. By the triangle and reverse triangle inequalities, we have

$$\begin{aligned}
\left\|\hat{f}_J \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} &\geq \left(\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|_{L^2_m(X)} - \left\|\hat{f}_J \circ \phi - \hat{f}_{J,\epsilon_m} \circ \phi\right\|_{L^2_m(X)}\right)^2 \\
&\geq \left(\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|_{L^2_m(X)} - \epsilon_m\right)^2 \\
&\geq \frac{1}{2}\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} - \epsilon_m^2,
\end{aligned}$$

Application of the union bound to Lem. 7 gives with probability at least $1 - e^{-t}$,

$$\begin{aligned}
\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} \\
- \sqrt{\frac{2\ln N_{m,J} + 2t}{m}}\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|_{L^2(X)} - \frac{4\ln N_{m,J}}{3m} - \frac{t}{3m}.
\end{aligned}$$

The AM-GM inequality yields with probability at least $1 - e^{-t}$,

$$\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{2}\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} - \frac{4\ln N_{m,J}}{3m} - \frac{4t}{3m}.$$

By the definition of $\hat{f}_{J,\epsilon_m}$, we have

$$\left\|\hat{f}_{J,\epsilon_m} \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{2}\left(\left\|\hat{f}_J \circ \phi - \mathbb{E}[Y|X]\right\|_{L^2(X)} - \epsilon_m\right)^2 - \frac{4\ln N_{m,J}}{3m} - \frac{4t}{3m}.$$

Noting that $\hat{f}_J \in \mathcal{F}_J$ almost surely, we have

$$\left\|\hat{f}_J \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} \geq A_J(\phi; X, Y) \text{ almost surely.}$$

By the triangle inequality, we have

$$(A_J(\phi; X, Y) - \epsilon_m)^2 \geq \frac{1}{2}A_J(\phi; X, Y) - \epsilon_m^2.$$

Consequently, we have with probability at least $1 - e^{-t}$,

$$\left\|\hat{f}_J \circ \phi - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{4}A_J(\phi; X, Y) - \frac{3\epsilon_m^2}{2} - \frac{2\ln N_{m,J}}{3m} - \frac{2t}{3m}.$$

Using the upper bound on $\ln N_{m,J}$ and the definition of ϵ_m, we get the desired result. □

B.2 Proofs for Bias Analyses with Learned Feature Extractor

Proofs for sieved least-square estimator Consider ϵ-cover of Φ in Asm. 2, denoted as $\phi_1, ..., \phi_{N_\epsilon}$, where $N_\epsilon = N(\epsilon, \Phi, \rho_J)$. Let $\hat{i}_\epsilon = \arg\min_{i \in [N_\epsilon]} \rho_J(\hat{\phi}, \phi_i)$ and $\hat{\phi}_\epsilon = \phi_{\hat{i}_\epsilon}$.

From the definition of the Minkowski–Bouligand dimension, for $c_2 > 1$, there is a sufficiently small $c_1 > 0$ such that with $\epsilon_m = \frac{c_1}{m}$, $\ln N(\epsilon_m, \mathcal{F}_J, \|\cdot\|_{L^\infty}) \leq c_2 J \ln(m)$ for any $m \geq 1$. Let $N_{m,J} = N(\epsilon, \mathcal{F}_J, \|\cdot\|_{L^\infty})$ and $f_1, ..., f_{N_{m,J}}$ be a ϵ_m-cover of $\mathcal{F}_J$ in $\|\cdot\|_{L^\infty}$. Let $\hat{f}_{J,\epsilon_m}$ be the closest function from these functions to $\hat{f}_J$ in $\|\cdot\|_{L^\infty}$.

Proof of Lem. 4. We can decompose the squared error as

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} = \left\|f_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} + \frac{2}{m}\sum_{i=1}^m \Big((\hat{f}_J \circ \hat{\phi})(X_i) - (f_J \circ \hat{\phi})(X_i)\Big)\epsilon_i + \frac{1}{m}\sum_{i=1}^m \Big((\hat{f}_J \circ \hat{\phi})(X_i) - Y_i\Big)^2 - \frac{1}{m}\sum_{i=1}^m \Big((f_J \circ \hat{\phi})(X_i) - Y_i\Big)^2,$$

where $\epsilon_i = Y_i - \mathbb{E}[Y|X]$. Since $\hat{f}_J$ is an empirical minimizer over $\mathcal{F}_J$ and $f_J \in \mathcal{F}_J$, we have

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \leq \left\|f_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} + \frac{2}{m}\sum_{i=1}^m \Big((\hat{f}_J \circ \hat{\phi})(X_i) - (f_J \circ \hat{\phi})(X_i)\Big)\epsilon_i.$$

For the second term in the right-hand side, we have

$$\frac{1}{m}\sum_{i=1}^m \Big((\hat{f}_J \circ \hat{\phi})(X_i) - (f_J \circ \hat{\phi})(X_i)\Big)\epsilon_i \leq \frac{\epsilon_m + 2\epsilon}{m}\sum_{i=1}^m |\epsilon_i| + \frac{1}{m}\sum_{i=1}^m \Big((\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon)(X_i) - (f_J \circ \hat{\phi}_\epsilon)(X_i)\Big)\epsilon_i.$$

By Lem. 9, Lem. 8, union bound over the ϵ-covers, and triangle inequality, we have with probability at least $1 - (1 + N_\epsilon)e^{-t}$,

$$\frac{\epsilon}{m}\sum_{i=1}^m |\epsilon_i| \leq \sigma(\epsilon_m + 2\epsilon)\left(\frac{2t}{m} + 2\ln 2\right)^{1/2},$$

and

$$\frac{1}{m}\sum_{i=1}^m \Big((\hat{f}_{J,\epsilon} \circ \hat{\phi}_\epsilon)(X_i) - (f_J \circ \hat{\phi}_\epsilon)(X_i)\Big)\epsilon_i \leq \sqrt{\frac{2\sigma^2(\ln(N_{m,J}) + t)}{m}}\left(\left\|\hat{f}_J \circ \hat{\phi} - f_J \circ \hat{\phi}_\epsilon\right\|_{L^2_m(X)} + \epsilon_m\right).$$

Consequently, we have with probability at least $1 - (1 + N_\epsilon)e^{-t}$,

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \leq \left\|f_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} + 2\sigma(\epsilon_m + 2\epsilon)\left(\frac{2t}{m} + 2\ln 2\right)^{1/2} + \sqrt{\frac{8\sigma^2(c_2 J \ln(m) + t)}{m}}\left(\left\|\hat{f}_J \circ \hat{\phi} - f_J \circ \hat{\phi}_\epsilon\right\|_{L^2_m(X)} + \epsilon_m + \epsilon\right).$$

By the triangle inequality, the AM-GM inequality, and the fact that $\sqrt{1+x} \leq 1 + x$ for $x \geq 0$, for any $\eta \in (0, 1)$, we have

$$(1-\eta)\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \leq (2+3\eta)\left\|f_J \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} + \frac{6\sigma^2 c_2 J \ln(m)}{\eta m} + \left(2\sqrt{2\ln(2)}\sigma c_1 + \frac{3\eta c_1^2}{m}\right)\frac{1}{m} + (2+3\eta)\epsilon + \left(\frac{2\sqrt{2}\sigma c_1}{\ln^{1/2}(2)} + \frac{6\sigma^2}{\eta}\right)\frac{t}{m}. \quad (12)$$

Application of Lem. 7 with union bound over ϵ-cover of Φ to the first term in the right-hand side of Eq. (12) yields with probability at least $1 - N_\epsilon e^{-t}$,

$$\left\|f_J \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \leq \left\|f_J \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} + \sqrt{\frac{2t}{m}\left\|f_J \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2(X)}} + \frac{t}{3m}.$$

The AM-GM inequality $\sqrt{2tv/m} \leq v + t/(2m)$ yields with probability at least $1 - N_\epsilon e^{-t}$,

$$\left\|f_J \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \leq 4\left\|f_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} + 2\epsilon^2 + \frac{5t}{6m}. \tag{13}$$

Combining Eq. (12) and Eq. (13) yields the claim. □

Proof of Lem. 5. By the reverse triangle and triangle inequalities, we have

$$\begin{aligned}\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} &\geq \left(\left\|\hat{f}_J \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|_{L^2_m(X)} - \epsilon\right)^2 \\ &\geq \left(\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|_{L^2_m(X)} - \epsilon_m - \epsilon\right)^2 \\ &\geq \frac{1}{2}\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} - \epsilon_m^2 - \epsilon^2.\end{aligned}$$

Application of the union bound to Lem. 7 gives with probability at least $1 - N_\epsilon e^{-t}$,

$$\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} - \sqrt{\frac{2\ln N_{m,J} + 2t}{m}}\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|_{L^2(X)} - \frac{\ln N_{m,J}}{3m} - \frac{t}{3m}.$$

The AM-GM inequality yields with probability at least $1 - N_\epsilon e^{-t}$,

$$\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{2}\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} - \frac{4\ln N_{m,J}}{3m} - \frac{4t}{3m}.$$

By the definitions of $\hat{f}_{J,\epsilon_m}$ and $\hat{\phi}_\epsilon$, we have

$$\left\|\hat{f}_{J,\epsilon_m} \circ \hat{\phi}_\epsilon - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{2}\left(\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|_{L^2(X)} - \epsilon_m - \epsilon\right)^2 - \frac{4\ln N_{m,J}}{3m} - \frac{4t}{3m}.$$

Noting that $\hat{f}_J \in \mathcal{F}_J$ almost surely, we have

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2(X)} \geq A_J(\hat{\phi}; X, Y) \text{ almost surely.}$$

By the triangle inequality, we have

$$\left(A_J(\hat{\phi}; X, Y) - \epsilon_m - \epsilon\right)^2 \geq \frac{1}{2}A_J(\hat{\phi}; X, Y) - \epsilon_m^2 - \epsilon^2.$$

Consequently, we have with probability at least $1 - N_\epsilon e^{-t}$,

$$\left\|\hat{f}_J \circ \hat{\phi} - \mathbb{E}[Y|X]\right\|^2_{L^2_m(X)} \geq \frac{1}{4}A_J(\hat{\phi}; X, Y) - \frac{3(\epsilon_m^2 + \epsilon^2)}{2} - \frac{2\ln N_{m,J}}{3m} - \frac{2t}{3m}.$$

Using the upper bound on $\ln N_{m,J}$ and the definition of ϵ_m, we get the desired result. □

B.3 Proofs for Concentration Inequalities

Proof of Lem. 7. Let $X_1, \dots, X_k$ be i.i.d. copies of X and set $g = h^2 \colon \mathcal{X} \to [0,1]$. Then

$$\|h\|^2_{L^2_k(X)} - \|h\|^2_{L^2(X)} = \frac{1}{k}\sum_{i=1}^{k}(g(X_i) - \mathbb{E}[g(X)]).$$

The summands $g(X_i) - \mathbb{E}[g(X)]$ are i.i.d., bounded in $[-1,1]$, and satisfy

$$\mathbb{V}[g(X)] \le \mathbb{E}[g(X)^2] = \mathbb{E}[h(X)^4] \le \mathbb{E}[h(X)^2] = \|h\|^2_{L^2(X)},$$

where $h^4 \le h^2$ holds pointwise since $|h| \le 1$. Applying Bernstein's inequality to the i.i.d. summands $g(X_i) - \mathbb{E}[g(X)]$ yields, with probability at least $1 - e^{-t}$,

$$\frac{1}{k}\sum_{i=1}^{k}(g(X_i) - \mathbb{E}[g(X)]) \le \sqrt{\frac{2\mathbb{V}[g(X)]\,t}{k}} + \frac{t}{3k} \le \sqrt{\frac{2t}{k}\|h\|^2_{L^2(X)}} + \frac{t}{3k}.$$

The same derivation is valid even if we exchange $\|h\|^2_{L^2_k(X)}$ and $\|h\|^2_{L^2(X)}$. □

Proof of Lem. 8.

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda}{k}\sum_{i=1}^{k} h_i\epsilon_i\right)\right] \\
&= \prod_{i=1}^{k}\mathbb{E}\left[\exp\left(\frac{\lambda}{k}h_i\epsilon_i\right)\right] \\
&= \prod_{i=1}^{k}\exp\left(\frac{\lambda^2\sigma^2 h_i^2}{2k^2}\right) \\
&= \exp\left(\frac{\lambda^2\sigma^2}{2k}\cdot\frac{1}{k}\sum_{i=1}^{k}h_i^2\right).
\end{aligned}$$

By the Chernoff bound, we have

$$\mathbb{P}\left\{\frac{1}{k}\sum_{i=1}^{k} h_i\epsilon_i > t\right\} \le \inf_{\lambda>0}\exp\left(\frac{\lambda^2\sigma^2}{2k}\cdot\frac{1}{k}\sum_{i=1}^{k}h_i^2 - \lambda t\right) \le \exp\left(-\frac{t^2 k}{2\sigma^2}\left(\frac{1}{k}\sum_{i=1}^{k}h_i^2\right)^{-1}\right).$$

Choosing t appropriately yields the claim. □

Proof of Lem. 9. For any $\lambda > 0$ and $x \in \mathbb{R}$, the inequality $e^{\lambda|x|} \le e^{\lambda x} + e^{-\lambda x}$ holds, so the sub-gaussian assumption gives

$$\mathbb{E}\left[e^{\lambda|\epsilon_i|}\right] \le \mathbb{E}\left[e^{\lambda\epsilon_i}\right] + \mathbb{E}\left[e^{-\lambda\epsilon_i}\right] \le 2e^{\lambda^2\sigma^2/2}.$$

By independence,

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{k}\sum_{i=1}^{k}|\epsilon_i|\right)\right] \le 2^k\exp\left(\frac{\lambda^2\sigma^2}{2k}\right).$$

The Chernoff bound then gives, for any $s > 0$,

$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}|\epsilon_i| > s\right) \le \inf_{\lambda>0}\exp\left(-\lambda k s + k\ln 2 + \frac{k\lambda^2\sigma^2}{2}\right).$$

The infimum over $\lambda > 0$ is attained at $\lambda^* = s/\sigma^2$, yielding

$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}|\epsilon_i| > s\right) \le \exp\left(-\frac{ks^2}{2\sigma^2} + k\ln 2\right).$$

Setting $s = \sigma\left(\frac{2t}{k} + 2\ln 2\right)^{1/2}$ gives $\frac{ks^2}{2\sigma^2} = t + k\ln 2$, so the right-hand side equals e^{-t}. □

C Experiments

C.1 Tasks and Architectures

We meta-train on three settings: Omniglot (Lake et al., 2015) 5-way and 20-way 1-shot on a Conv4-64 backbone over 28×28 grayscale inputs, and Mini-ImageNet (Vinyals et al., 2016; Ravi & Larochelle, 2017) 5-way 1-shot on a Conv4-128 backbone over 84×84 RGB inputs, each with five queries per task. Omniglot features transfer to SVHN (Netzer et al., 2011), while Mini-ImageNet features transfer to CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). We set the pre-training sample size $m \in \{1{,}000, 2{,}000, 4{,}000, 8{,}000, 16{,}000\}$ for Omniglot and $m \in \{8{,}000, 16{,}000, 24{,}000, 32{,}000, 40{,}000\}$ for Mini-ImageNet, and the fine-tuning sample size $n \in \{100, 500, 1{,}000, 2{,}500, 5{,}000\}$.

C.2 Algorithms, Regularizers, and Optimization

We compare four meta-learners spanning the gradient-based and metric-based families: first- and second-order MAML (Finn et al.), Prototypical Networks (Snell et al., 2017), and R2-D2 (Bertinetto et al., 2019). Meta-training adds a regularization term $\lambda \cdot \mathcal{R}(\theta)$ to the meta-loss, with $\mathcal{R}$ the spectral norm $\sum_l \sigma_{\max}(W_l)$ or the entry-wise ℓ_1 norm $\sum_l \|W_l\|_1$, each globally normalized so that λ is comparable across architectures and penalties. The outer loop runs 3,000 steps of Adam (Kingma & Ba, 2015) (β_1=0.9, β_2=0.95) at lr $= 3 \times 10^{-4}$ with meta batch size 8. Fine-tuning trains the full backbone plus a fresh linear head for 30 epochs of Adam at 3×10^{-4}. The regularization coefficient λ is selected by grid search over $\{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. The mean values over three random seeds with standard deviation are reported.

C.3 Implementation and Compute

The pipeline is built on `JAX` (Bradbury et al., 2018) v0.10 with CUDA 12.9, `Equinox` (Kidger & Garcia, 2021), and `Optax` (DeepMind et al., 2020). Every trial occupies an NVIDIA H100 (80 GB) in single precision. The full sweep consumed about 1,000 H100-hours.

C.4 Additional Results

Fig. 3 presents the downstream test error rates on SVHN, CIFAR-10, and CIFAR-100 for CNN models pre-trained on Omniglot (5-way 1-shot and 20-way 1-shot) and Mini-ImageNet. It can be seen that regularizing model complexity improves downstream sample efficiency in various settings.

In addition to the results with the spectral norm regularization, Fig. 4 displays the results with the ℓ_1-norm regularization. We can observe the improvements by ℓ_1-norm regularization, although they are not as vivid as those by spectral-norm regularization.

D Comprehensive survey of related work

D.1 Meta-Learning Methodologies

Meta-learning is a framework that seeks to acquire a learning procedure capable of adapting to future unseen tasks using limited samples. Rather than focusing on generalization within a single task, it leverages experiences from a collection of past tasks drawn from a task distribution (Hospedales et al., 2021; Wang et al., 2020). A widely adopted taxonomy for these methods consists of a tripartite classification: (i) metric-based methods, which rely on distances and similarities; (ii) optimization-based methods, which incorporate gradient updates in an inner loop; and (iii) model-based methods, which construct the learner itself using mechanisms such as memory or hypernetworks.

Metric-based approaches. The metric-based family employs a framework for classifying queries based on proximity, attention, or comparison within an embedding space. As a representative example, Matching Networks proposed one-shot classification via attention

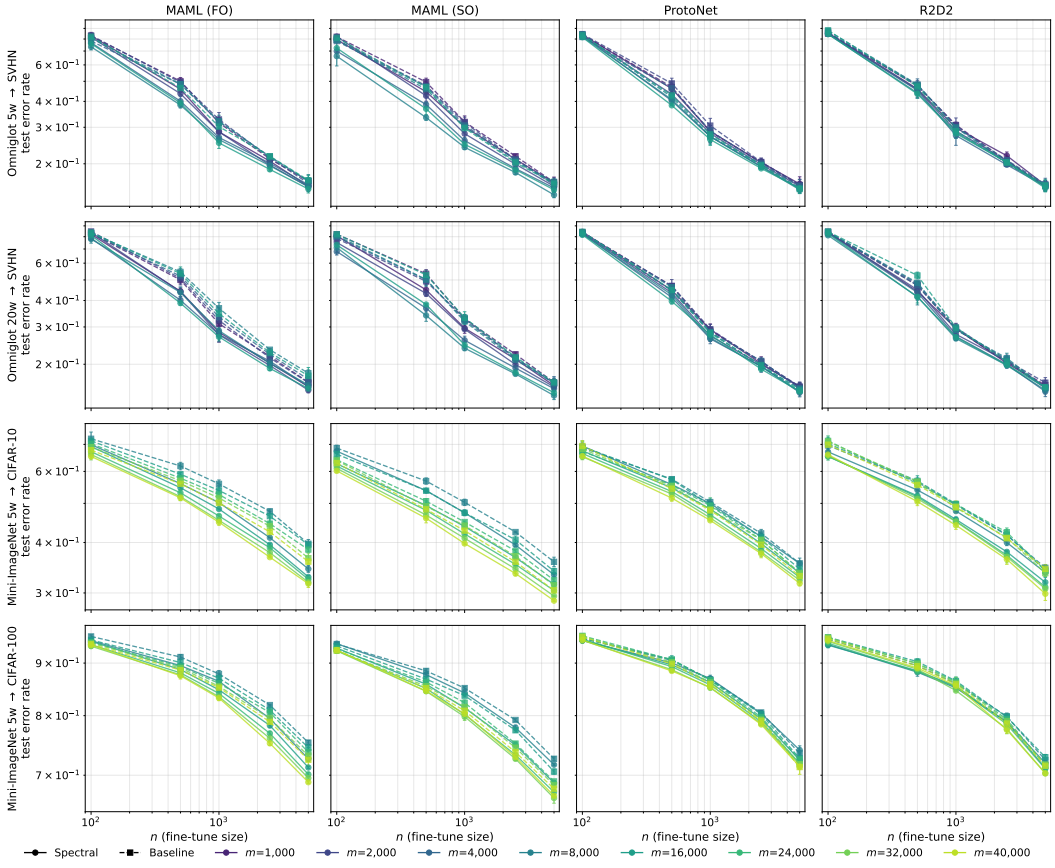

Figure 3: Downstream test error rate on SVHN, CIFAR-10, CIFAR-100 (log scale) vs. fine-tuning sample size (n, log scale) with different pre-training sample size (m) for four meta-learning algorithms with and without parameter spectral-norm regularization trained on Omniglot or Mini-ImageNet.

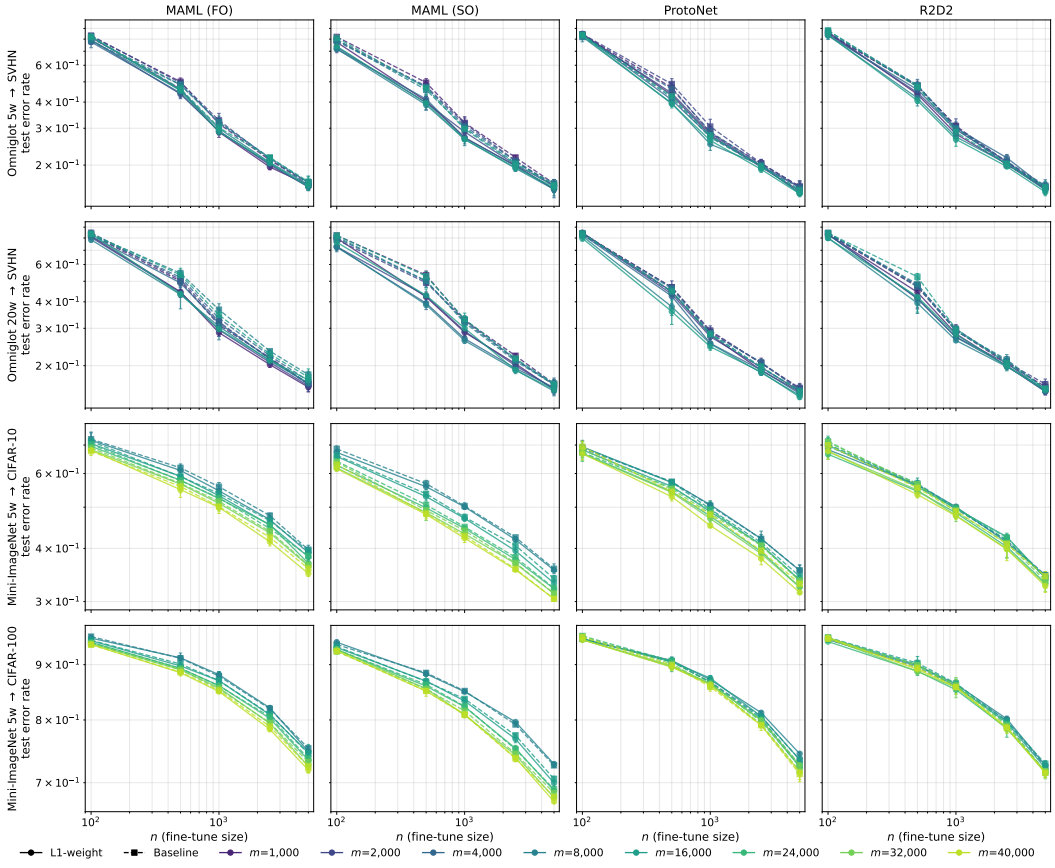

Figure 4: Downstream test error rate on SVHN, CIFAR-10, CIFAR-100 (log scale) vs. fine-tuning sample size (n, log scale) with different pre-training sample size (m) for four meta-learning algorithms with and without parameter ℓ_1-norm regularization trained on Omniglot or Mini-ImageNet.

between a support set and a query, formalizing the episodic training regime (Vinyals et al., 2016). Prototypical Networks introduced a concise classifier based on distances to class-specific "prototypes" (mean embeddings), providing a clear perspective on meta-representation learning (Snell et al., 2017). Relation Networks enable more expressive comparisons by learning the distance function itself using a neural network (Sung et al., 2018).

Optimization-based approaches. The optimization-based family views meta-learning as "learning an initial parameter set or update rule such that a few steps of optimization on an unseen task lead to high performance." MAML established a model-agnostic framework by adapting to task-specific parameters via K-step gradient descent in the inner loop and optimizing the post-adaptation performance in the outer loop (Finn et al.). First-order methods such as FO-MAML and Reptile are categorized as algorithms that avoid second-order derivative computations while shifting initial values in a direction that makes "simultaneous learning from the same starting point" easier across tasks (Nichol et al., 2018). Meta-SGD further parameterizes not only the initial values but also the update directions and learning rates, thereby learning a higher-capacity "way to learn" (Li et al., 2017). Another line of work replaces the iterative inner-loop adaptation with differentiable closed-form or rapidly convergent solvers. R2-D2 introduced a differentiable ridge-regression base learner that constructs task-specific classifiers on top of learned embeddings, allowing the meta-objective to be optimized by backpropagating through the solver itself (Bertinetto et al., 2019). This approach occupies an intermediate position between metric-based methods, which often rely on fixed nearest-neighbor or prototype rules after representation learning, and gradient-based methods such as MAML, which perform explicit iterative parameter adaptation. Additionally, iMAML, which computes meta-gradients using implicit gradients without explicitly unrolling the inner loop, is a representative example of scaling these methods by treating them as bilevel optimization problems (Rajeswaran et al., 2019).

Model-based approaches. The model-based family implements intra-task adaptation as part of the network's computation using external memory, hypernetworks, or architectures designed to learn the optimizer. Memory-Augmented Neural Networks (MANN) demonstrated rapid one/few-shot adaptation by using external memory to quickly write and read new information, providing a foundation for meta-learning via model design (Santoro et al., 2016). Furthermore, the classical lineage of "learning to learn" such as "Optimization as a Model" (which uses LSTMs to learn update rules), can be understood as a bridge between model-based and optimization-based approaches (Ravi & Larochelle, 2017; Hochreiter et al., 2001; Andrychowicz et al., 2016). SNAIL demonstrated high performance across multiple domains (supervised and reinforcement learning) as a general-purpose meta-learner combining temporal convolutions with attention (Mishra et al., 2017).

D.2 Meta-Representation Learning: Sharing Representations Across Tasks

Meta-learning is closely related to transfer learning in that it transfers experiences from numerous tasks to an unseen task (Pan & Yang, 2009). Representation learning naturally motivates the acquisition of shared representations in transfer and multi-task learning, based on the general principle that mapping inputs to useful latent representations facilitates learning (Bengio et al., 2013). Meta-representation learning is characterized by specializing these shared representations for few-shot adaptation within a task, aiming for both statistical and computational efficiency simultaneously.

Classically, Baxter's model of inductive bias learning clarified the concept of meta-generalization, namely learning a good hypothesis space by observing multiple tasks from a task environment, and served as the starting point for subsequent formalizations (Baxter, 2000). In more recent learning theory, Multi-Task Representation Learning (MTRL) has emerged, showing the benefits of learning low-dimensional representations such as dictionaries or feature maps from multiple tasks through generalization error bounds; this provides theoretical support for the acquisition of shared representations in meta-representation learning (Maurer et al., 2016; 2013).

As a theory dealing more explicitly with meta-representation learning, research has provided algorithms and lower bounds for achieving sample-efficient representation estimation and

transfer to unseen tasks in settings where a group of linear regression tasks shares a common low-dimensional linear representation (Tripuraneni et al., 2021). Additionally, some studies analyze the effects of overparameterization on the sample efficiency of meta-representation learning using linear regression sequences, beginning to explain the phenomena observed in deep meta-learning where few-shot adaptation is possible even with large-scale models (Sun et al., 2021).

D.3 Learning Theory of Meta-Learning

Learning theory for meta-learning must account for a dual-sampling structure: the extraction of tasks from a task distribution and the sampling of data points within each individual task (Baxter, 2000; Hospedales et al., 2021). Recent theoretical studies commonly employ a framework that decomposes excess risk into statistical estimation error, optimization error, and model approximation error, centered around the meta-generalization gap, the discrepancy between the expected risk on unseen tasks and the empirical meta-objective (Rezazadeh, 2022; Wang & Arora, 2024). In particular, significant progress has been made in precisely analyzing the effects of representation sharing across tasks and how the number of adaptation steps (the inner loop) influences the overall stability of the algorithm (Huang et al., 2022; Chen et al., 2022).

Algorithmic Stability. This measures the sensitivity of the output to the replacement of a single data point in the training set. By introducing the concept of "meta-stability" which accounts for the stability of both the inner and outer loops, this framework provides realistic bounds even for gradient-based methods involving non-convex optimization (Wang & Arora, 2024; Bousquet & Elisseeff, 2002).

PAC-Bayes Theory. By introducing a hierarchy of meta-priors and task-specific posteriors, this approach derives bounds dependent on both the task count and sample size (Pentina & Lampert, 2014; Amit & Meir, 2018; Rezazadeh, 2022).

Information-Theoretic Approach. This approach evaluates generalization error using the mutual information between the algorithm's output and the input data, thereby quantifying the dependency on the underlying data distribution (Chen et al., 2021).

Uniform Convergence. Although this framework utilizes traditional complexity measures, the resulting bounds tend to be loose in the context of deep learning and meta-learning. Consequently, data-dependent analyses have become the mainstream approach in recent years (Nagarajan & Kolter, 2019).

D.4 Deep Learning Theory

Theoretical understanding of deep learning has been developed from several complementary perspectives, including approximation theory, statistical generalization, optimization, representation learning, and scaling laws. A classical line of work studies the expressive power and statistical estimation properties of neural networks. Deep ReLU networks are known to approximate rich function classes with rates depending on smoothness, sparsity, compositionality, or intrinsic dimension (Yarotsky, 2017; Schmidt-Hieber). Particularly relevant to our work is the theory of adaptive approximation and estimation by deep networks. Suzuki (Suzuki, 2019) showed that deep ReLU networks achieve minimax optimal rates over Besov and mixed-smooth Besov spaces and can adapt to spatially inhomogeneous smoothness. Hayakawa and Suzuki (Hayakawa & Suzuki) further established minimax optimality and the superiority of deep neural network learning over sparse parameter spaces. These results clarify an important statistical mechanism behind deep learning: deep nonlinear architectures can exploit hidden structural regularities that are difficult for non-adaptive linear or kernel methods to capture. This adaptivity perspective has also been extended to modern architectures, including convolutional and ResNet-type networks, Transformers, and diffusion models (Oono & Suzuki, 2019; Takakura & Suzuki, 2023; Oko et al., 2023).

Another major line of work studies generalization in overparameterized neural networks. Since classical capacity bounds based on the raw number of parameters are too pessimistic for modern deep learning, refined analyses have been developed using norms, margins, PAC-Bayes

bounds, compression, and algorithm-dependent complexity measures. For example, Bartlett et al. (Bartlett et al., 2017) derived spectrally-normalized margin bounds, and Neyshabur et al. (Neyshabur et al., 2018) developed PAC-Bayesian spectrally-normalized bounds. At the same time, empirical and theoretical studies of interpolation, benign overfitting, and double descent have shown that the classical bias–variance trade-off does not fully explain modern neural network generalization (Zhang et al., 2017; Belkin et al., 2019; Nakkiran et al., 2021). These studies mainly concern single-task learning, whereas our work studies how representations learned from multiple source tasks affect the sample complexity of future tasks.

Optimization theory provides another perspective. The neural tangent kernel (NTK) theory shows that infinitely wide neural networks trained by gradient descent can behave like kernel methods (Jacot et al., 2018), and related overparameterization analyses establish global convergence of gradient-based methods under suitable assumptions (Du et al., 2019; Allen-Zhu et al., 2019). However, NTK analyses typically describe a lazy-training regime in which features remain nearly fixed during training (Chizat et al., 2019). This perspective alone is insufficient to explain representation learning, where the features themselves are learned. Mean-field analyses provide an alternative view in which the distribution of neurons evolves during training and feature learning can occur (Mei et al., 2018; Sirignano & Spiliopoulos, 2020; Woodworth et al., 2020).

Recent work has therefore focused on feature learning beyond fixed-kernel or lazy-training regimes. Ba et al. (Ba et al., 2022) showed that even a single gradient step on the first-layer weights of a two-layer network can improve the learned representation over random features and outperform broad classes of fixed-kernel methods. Suzuki et al. (Suzuki et al., 2023) analyzed feature learning via mean-field Langevin dynamics and showed that mean-field neural networks can achieve sample-complexity improvements over kernel methods for structured problems such as sparse parity learning. More recently, Nishikawa et al. (Nishikawa et al., 2025) showed that nonlinear Transformers can perform inference-time feature learning in in-context learning. These works are closely aligned with our motivation: the statistical advantage of deep learning comes not only from large model capacity, but also from the ability to learn task-relevant representations.

The success of large pre-trained models has also motivated theoretical studies of representation learning and scaling laws. Contrastive and self-supervised representation learning have been analyzed as mechanisms for extracting downstream-useful features from auxiliary or unlabeled data (Saunshi et al., 2019; HaoChen et al., 2021). Empirical scaling laws have shown that loss often follows predictable power-law behavior as data, model size, or compute increases (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022), and recent theoretical work has attempted to explain such laws through variance-limited regimes, data geometry, kernel spectra, and feature learning (Bahri et al., 2024; Bordelon et al., 2025). Scaling laws have also been studied in transfer and downstream settings (Hernandez et al., 2021; Lourie et al., 2025), where the relation between source data and target tasks becomes essential.

Our work is situated at the intersection of these theories and the theory of meta-learning. Classical and modern meta-learning theory shows that multiple related tasks can reduce the sample complexity of future tasks by learning a shared inductive bias or representation (Baxter, 2000; Maurer et al., 2016; Tripuraneni et al., 2021). In contrast to most general deep learning theory, which primarily studies single-task approximation, optimization, or generalization, we analyze a meta-representation learning algorithm and prove the achievability of a data scaling law. Thus, our result connects the adaptivity and feature-learning viewpoint of deep learning theory with the statistical theory of meta-learning, making explicit how the number of source tasks and the number of samples per task jointly determine downstream sample efficiency.