Transformers are Minimax Optimal Nonparametric In-Context Learners

Juno Kim[1,2]* **Tai Nakamaki**[1] **Taiji Suzuki**[1,2]
[1]University of Tokyo [2]Center for Advanced Intelligence Project, RIKEN
*junokim@g.ecc.u-tokyo.ac.jp

Abstract

In-context learning (ICL) of large language models has proven to be a surprisingly effective method of learning a new task from only a few demonstrative examples. In this paper, we study the efficacy of ICL from the viewpoint of statistical learning theory. We develop approximation and generalization error bounds for a transformer composed of a deep neural network and one linear attention layer, pretrained on nonparametric regression tasks sampled from general function spaces including the Besov space and piecewise γ-smooth class. We show that sufficiently trained transformers can achieve – and even improve upon – the minimax optimal estimation risk in context by encoding the most relevant basis representations during pretraining. Our analysis extends to high-dimensional or sequential data and distinguishes the *pretraining* and *in-context* generalization gaps. Furthermore, we establish information-theoretic lower bounds for meta-learners w.r.t. both the number of tasks and in-context examples. These findings shed light on the roles of task diversity and representation learning for ICL.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in understanding and generating natural language data. In particular, the phenomenon of *in-context learning* (ICL) has recently garnered widespread attention. ICL refers to the ability of pretrained LLMs to perform a new task by being provided with a few examples within the context of a prompt, without any parameter updates or fine-tuning. It has been empirically observed that few-shot prompting is especially effective in large-scale models (Brown et al., 2020) and requires only a couple of examples to consistently achieve high performance (García et al., 2023). In contrast, Raventos et al. (2023) demonstrate that sufficient pretraining task diversity is required for the emergence of ICL. However, we still lack a comprehensive understanding of the statistical foundations of ICL and few-shot prompting.

A vigorous line of research has been directed towards understanding ICL of single-layer linear attention models pretrained on the query prediction loss of linear regression tasks (Garg et al., 2022; Akyürek et al., 2023; Zhang et al., 2023; Ahn et al., 2023; Mahankali et al., 2023; Wu et al., 2024). It has been shown that the global minimizer of the L^2 pretraining loss implements one step of GD on a least-squares linear regression objective (Mahankali et al., 2023) and is nearly Bayes optimal (Wu et al., 2024). Moreover, risk bounds with respect to the context length (Zhang et al., 2023) and number of tasks (Wu et al., 2024) have been obtained.

Other works have examined ICL of more complex multi-layer transformers. Bai et al. (2023); von Oswald et al. (2023) give specific transformer constructions which simulate GD in context, however it is unclear how such meta-algorithms may be learned. Another approach is to study learning *with representations*, where tasks consist of a fixed nonlinear feature map composed with a varying linear function. Guo et al. (2023) empirically found that trained transformers exhibit a separation where lower layers transform the input and upper layers perform linear ICL. Recently, Kim and Suzuki

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(2024) analyzed a model consisting of a shallow neural network followed by a linear attention layer and proved that the MLP component learns to encode the true features during pretraining. However, they assumed the infinite task and sample size limit and did not study generalization capabilities.

Our contributions. In this paper, we analyze the optimality of ICL from the perspective of statistical learning theory. Our object of study is a transformer consisting of a deep neural network with N-dimensional output followed by one linear attention layer. The model is pretrained on n input-output samples from T nonparametric regression tasks, generated from a suitably decaying distribution on a general function space. Compared to previous works, we take a crucial step towards understanding practical multi-layer transformers by incorporating the representation learning capabilities of the DNN module. From a more abstract perspective, this work can also be situated as a nonlinear extension of meta-learning. Our contributions are highlighted below.

- We develop a general framework for upper bounding the in-context estimation error of the empirical risk minimizer in terms of the approximation error of the neural network and separate *in-context* and *pretraining* generalization gaps depending on n, T, respectively.
- In the Besov space setting, we show that **ICL achieves nearly minimax optimal risk** $n^{-\frac{2\alpha}{2\alpha+d}}$ when T is sufficiently large. Since LLMs are pretrained on vast amounts of data in practice, T can be taken to be nearly infinite, justifying the emergence of ICL at large scales. We extend the optimality guarantees to nearly dimension-free rates in the anisotropic Besov space and also to learning sequential data with deep transformers in the piecewise γ-smooth function class.
- We show that ICL can **improve upon the a priori optimal rate** when the task class basis resides in a coarser Besov space by learning to encode informative basis representations, emphasizing the importance of pretraining on diverse tasks.
- We also derive **information-theoretic lower bounds** for the minimax risk in both n, T, rigorously confirming that ICL is jointly optimal when T is large (Besov space setting), while any meta-learning method is jointly suboptimal when T is small (coarser space setting). This separation aligns with empirical observations of a *task diversity threshold* (Raventos et al., 2023).

The paper is structured as follows. In Section 2, the regression tasks and transformer model are defined in an abstract setting. In Section 3, we present the general framework for estimating the ICL approximation and generalization error. In Section 4, we specialize to the Besov-type and piecewise γ-smooth class settings and show that transformers can achieve or exceed the minimax optimal rate in context. In Section 5, we derive minimax lower bounds. All proofs are deferred to the appendix; moreover, we provide **numerical experiments** validating our results in Appendix E.

1.1 Other Related Works

Meta-learning. The theoretical setting of ICL is closely related to *meta-learning*, where the goal is to infer a shared representation ψ° with samples from a set of transformed tasks $\beta_i^\top \psi^\circ$. When ψ° is linear, fast rates have been established by Tripuraneni et al. (2020); Du et al. (2021), while the nonlinear case has been studied by Meunier et al. (2023) where ψ° is a feature projection into a reproducing kernel Hilbert space. Our results can be viewed as extending this body of work to function spaces of generalized smoothness with a specific deep transformer architecture.

Optimal rates for DNNs. Our analysis extends established optimality results for classes of DNNs in ordinary supervised regression settings to ICL. Suzuki (2019) has shown that deep feedforward networks with the ReLU activation can efficiently approximate functions in the Besov space and thus achieve the minimax optimal rate. This has been extended to the anisotropic Besov space (Suzuki and Nitanda, 2021), convolutional neural networks for infinite-dimensional input (Okumoto and Suzuki, 2022), and transformers for sequence-to-sequence functions (Takakura and Suzuki, 2023). We remark that a work in progress (Imaizumi, 2024) also studies the sample complexity of ICL of transformers in a Sobolev space setting.

Pretraining dynamics for ICL. While how ICL arises from optimization is not fully understood, there are encouraging developments in this direction. Zhang et al. (2023) has shown for one layer of linear attention that running GD on the population risk always converges to the global optimum. This was extended to incorporate a linear output layer by Zhang et al. (2024), and to softmax attention by Huang et al. (2023); Li et al. (2024); Chen et al. (2024). Kim and Suzuki (2024) considered a compound transformer equivalent to ours with a shallow MLP component and proved that the loss landscape becomes benign in the mean-field limit, deriving convergence guarantees for the corresponding gradient dynamics. These analyses indicate that the attention mechanism, while highly nonconvex, may possess structures favorable for gradient-based optimization.

2 Problem Setup

2.1 Nonparametric Regression

In this paper, we analyze the ability of a transformer to solve nonparametric regression problems in context when pretrained on examples from a *family* of regression tasks, which we describe below. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space (d is allowed to be infinite), $\mathcal{P}_{\mathcal{X}}$ a probability distribution on $\mathcal{X}$, and $(\psi_j^\circ)_{j=1}^\infty$ a fixed countable subset of $L^2(\mathcal{P}_{\mathcal{X}})$. A regression function $F_\beta^\circ : \mathcal{X} \to \mathbb{R}$ is randomly generated for each task by sampling the sequence of coefficients $\beta \in \mathbb{R}^\infty$ from a distribution $\mathcal{P}_\beta$ on $\mathscr{B}(\mathbb{R}^\infty)$; the class of tasks $\mathcal{F}^\circ$ is defined as

$$\mathcal{F}^\circ = \left\{ F_\beta^\circ = \textstyle\sum_{j=1}^\infty \beta_j \psi_j^\circ \,\middle|\, \beta \in \operatorname{supp} \mathcal{P}_\beta \right\},$$

endowed with the induced distribution. Each task prompt contains n example input-response pairs $\{(x_k, y_k)\}_{k=1}^n$. The covariates x_k are i.i.d. drawn from $\mathcal{P}_{\mathcal{X}}$ and the responses are generated as

$$y_k = F_\beta^\circ(x_k) + \xi_k, \quad 1 \le k \le n, \tag{1}$$

where the noise ξ_k is assumed to be i.i.d. with mean zero and $|\xi_k| \le \sigma$ almost surely.[1] In addition, we independently generate a query token $\tilde{x}$ and corresponding output $\tilde{y}$ in the same manner.

We proceed to state our assumptions for the regression model. Informally, we suppose a relaxed version of sparsity and orthonormality of ψ_j° and suitable decay rates for the basis expansion. These will be subsequently verified for specific function spaces of interest with their natural decay rate.

Assumption 1 (relaxed sparsity and orthonormality of basis functions)**.** *For $N \in \mathbb{N}$, there exist integers $\underline{N} < \bar{N} \lesssim N$ with $\bar{N} - \underline{N} + 1 = N$ such that $\psi_{\underline{N}}^\circ, \cdots, \psi_{\bar{N}}^\circ$ are independent and $\psi_1^\circ, \cdots, \psi_{\underline{N}-1}^\circ$ are all contained in the linear span of $\psi_{\underline{N}}^\circ, \cdots, \psi_{\bar{N}}^\circ$. Moreover, there exist $r, C_1, C_2, C_\infty > 0$ such that $\Sigma_{\Psi,N} := \left(\mathbb{E}_{x\sim\mathcal{P}_{\mathcal{X}}}[\psi_j^\circ(x)\psi_k^\circ(x)]\right)_{j,k=\underline{N}}^{\bar{N}}$ satisfies $C_1 \mathbf{I}_N \preceq \Sigma_{\Psi,N} \preceq C_2 \mathbf{I}_N$ and*

$$\left\| \textstyle\sum_{j=\underline{N}}^{\bar{N}} (\psi_j^\circ)^2 \right\|_{L^\infty(\mathcal{P}_{\mathcal{X}})}^{1/2} \le C_\infty N^r. \tag{2}$$

Denoting the $\bar{N}$-basis approximation of F_β° as $F_{\beta,\bar{N}}^\circ := \sum_{j=1}^{\bar{N}} \beta_j \psi_j^\circ$, by Assumption 1 there exist 'aggregated' coefficients $\bar{\beta}_{\underline{N}}, \cdots, \bar{\beta}_{\bar{N}}$ uniquely determined by β such that $F_{\beta,\bar{N}}^\circ = \sum_{j=\underline{N}}^{\bar{N}} \bar{\beta}_j \psi_j^\circ$. We define two types of coefficient covariance matrices

$$\Sigma_{\beta,\bar{N}} := (\mathbb{E}_\beta[\beta_j\beta_k])_{j,k=1}^{\bar{N}} \in \mathbb{R}^{\bar{N}\times\bar{N}} \quad \text{and} \quad \Sigma_{\bar{\beta},N} := \left(\mathbb{E}_\beta[\bar{\beta}_j\bar{\beta}_k]\right)_{j,k=\underline{N}}^{\bar{N}} \in \mathbb{R}^{N\times N}.$$

Assumption 2 (decay of β)**.** *For $s, B > 0$ it holds that $\|F_\beta^\circ\|_{L^\infty(\mathcal{P}_{\mathcal{X}})} \le B$ for all $F_\beta^\circ \in \mathcal{F}^\circ$ and*

$$\|F_\beta^\circ - F_{\beta,N}^\circ\|_{L^2(\mathcal{P}_{\mathcal{X}})}^2 \lesssim N^{-2s} \tag{3}$$

uniformly over $\beta \in \operatorname{supp} \mathcal{P}_\beta$. Furthermore, $\operatorname{Tr}(\Sigma_{\bar{\beta},N})$ is bounded for all N and

$$0 \prec \Sigma_{\beta,\bar{N}} \lesssim \operatorname{diag}\left[(j^{-2s-1}(\log j)^{-2})_{j=1}^{\bar{N}}\right]. \tag{4}$$

Remark 2.1. In the simple case where $(\psi_j^\circ)_{j=1}^\infty$ is a basis for $L^2(\mathcal{P}_{\mathcal{X}})$, we may set $\underline{N} = 1, \bar{N} = N$ so that the dependency condition of Assumption 1 is trivially satisfied, moreover, $\Sigma_{\bar{\beta},N} = \Sigma_{\beta,N}$ and boundedness of $\operatorname{Tr}(\Sigma_{\bar{\beta},N})$ automatically follows from (4). However, the assumptions in the stated form also allow for hierarchical bases with dependencies such as wavelet systems. We also note that (3) and (4) entail basically the same rate but are not equivalent: the *uniform* bound $|\beta_j|^2 \lesssim j^{-2s-1}(\log j)^{-2}$ along with Assumption 1 implies (3). The $(\log j)^{-2}$ term can be replaced with any $g(j)$ such that $\sum_{j=1}^\infty j^{-1}g(j)$ is convergent.

[1] This implies $\operatorname{Var} \xi_k \le \sigma^2$. We require boundedness since the values y_k form part of the prompt input, and we wish to utilize sup-norm covering number estimates for the attention map; this technicality can be removed with more careful analysis. However, for the information-theoretic lower bound we assume Gaussian noise.

2.2 In-Context Learning

We now describe our transformer model, which takes n context pairs $\mathbf{X} = (x_1, \cdots, x_n) \in \mathbb{R}^{d\times n}$, $\boldsymbol{y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ and a query token $\tilde{x}$ as input and returns a prediction for the corresponding output. The covariates are first passed through a nonlinear representation or feature mapping $\phi : \mathcal{X} \to \mathbb{R}^N$, which we assume belongs to a sufficiently powerful class of estimators $\mathcal{F}_N$. Specifically:

Assumption 3 (expressivity of $\mathcal{F}_N$)**.** *$\|\phi(x)\|_2 \le B'_N$ for some $B'_N > 0$ for all $x \in \mathcal{X}$, $\phi \in \mathcal{F}_N$. Moreover for some $\delta_N > 0$, there exist $\phi^*_{\underline{N}}, \cdots, \phi^*_{\bar{N}} \in \mathcal{F}_N$ satisfying*

$$\max_{\underline{N}\le j\le \bar{N}} \|\psi^\circ_j - \phi^*_j\|_{L^\infty(\mathcal{P}_\mathcal{X})} \le \delta_N.$$

By choosing $\mathcal{F}_N$ and δ_N to satisfy the above assumption, we will be able to utilize established approximation and generalization guarantees for families of deep neural networks in Section 4.

The extracted representations $\phi(\mathbf{X}) = (\phi(x_1), \cdots, \phi(x_n))$ are then mapped to a scalar output via a linear attention layer parametrized by a matrix $\Gamma \in \mathcal{S}_N$ for $\mathcal{S}_N \subset \mathbb{R}^{N\times N}$,

$$\check{f}_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}) := \frac{1}{n}\sum_{k=1}^n y_k \phi(x_k)^\top \Gamma^\top \phi(\tilde{x}) = \left\langle \frac{\Gamma\phi(\mathbf{X})\boldsymbol{y}}{n}, \phi(\tilde{x}) \right\rangle, \quad \text{where } \Theta = (\Gamma, \phi) \in \mathcal{S}_N \times \mathcal{F}_N\,.$$

Finally, the output is constrained to lie on $[-\bar{B}, \bar{B}]$ by applying $\operatorname{clip}_{\bar{B}}(u) := \max\{\min\{u, \bar{B}\}, -\bar{B}\}$, yielding $f_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}) := \operatorname{clip}_{\bar{B}}(\check{f}_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}))$. We set $\mathcal{S}_N = \{\Gamma \in \mathbb{R}^{N\times N} \mid 0 \preceq \Gamma \preceq C_3\mathbf{I}_N\}$ for some $C_3 > 0$ and fix $\bar{B} = B$ for simplicity.

The above setup is a restricted reparametrization of linear attention widely used in theoretical analyses (see e.g. Zhang et al., 2023; Wu et al., 2024, for more details), where the values only refer to $\boldsymbol{y}$ and the query and key matrices are consolidated into one matrix Γ. The form is equivalent to one step of GD with matrix step size and has been shown to be optimal for a single layer of linear attention for linear regression tasks (Ahn et al., 2023; Mahankali et al., 2023). The placement of the attention layer after the DNN module ϕ is justified by the observation that lower layers of trained transformers act as data representations on top of which upper layers perform ICL (Guo et al., 2023).

During pretraining, the model is presented with T prompts $\{(\mathbf{X}^{(t)}, \boldsymbol{y}^{(t)}, \tilde{x}^{(t)})\}_{t=1}^T$ where the tasks $F^\circ_{\beta^{(t)}} \in \mathcal{F}^\circ$, $\beta^{(t)} \sim \mathcal{P}_\beta$ and tokens $\mathbf{X}^{(t)} = (x_1^{(t)}, \cdots, x_n^{(t)})$, $\boldsymbol{y}^{(t)} = (y_1^{(t)}, \cdots, y_n^{(t)})^\top$, $\tilde{x}^{(t)}$ and $\tilde{y}^{(t)}$ are independently generated as described in Section 2.1, and is trained to minimize the empirical risk

$$\widehat{\Theta} = \operatorname*{arg\,min}_{\Theta\in\mathcal{S}_N\times\mathcal{F}_N} \widehat{R}(\Theta), \quad \widehat{R}(\Theta) = \frac{1}{T}\sum_{t=1}^T \left(\tilde{y}^{(t)} - f_\Theta(\mathbf{X}^{(t)}, \boldsymbol{y}^{(t)}, \tilde{x}^{(t)})\right)^2.$$

Our goal is to verify the efficiency of ICL as a learning algorithm and show that learning the optimal $\widehat{\Theta}$ allows the transformer to solve new random regression problems $y = F^\circ_\beta(x) + \xi$ for $F^\circ_\beta \in \mathcal{F}^\circ$ in context. To this end, we evaluate the convergence of the mean-squared risk or estimation error,

$$\bar{R}(\widehat{\Theta}) := \mathbb{E}_{(\mathbf{X}^{(t)}, \boldsymbol{y}^{(t)}, \tilde{x}^{(t)}, \tilde{y}^{(t)})_{t=1}^T}[R(\widehat{\Theta})], \quad R(\Theta) := \mathbb{E}_{\mathbf{X}, \boldsymbol{y}, \tilde{x}, \beta}\left[(F^\circ_\beta(\tilde{x}) - f_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}))^2\right].$$

Note that we do not study whether the transformer always converges to $\widehat{\Theta}$; the training dynamics of a DNN is already a very difficult problem. For the attention layer, see the discussion in Section 1.1.

3 Risk Bounds for In-Context Learning

In this section, we outline our framework for analyzing the in-context estimation error $\bar{R}(\widehat{\Theta})$. Some additional definitions are in order. The ϵ-covering number $\mathcal{N}(\mathcal{C}, \rho, \epsilon)$ of a metric space $\mathcal{C}$ equipped with a metric ρ for $\epsilon > 0$ is defined as the minimal number of balls in ρ with radius ϵ needed to cover $\mathcal{C}$ (van der Vaart and Wellner, 1996). The *ϵ-covering entropy* or *metric entropy* is given as $\mathcal{V}(\mathcal{F}, \rho, \epsilon) := \log \mathcal{N}(\mathcal{F}, \rho, \epsilon)$. The ϵ-packing number $\mathcal{M}(\epsilon, \mathcal{C}, \rho)$ is given as the maximal cardinality of a ϵ-separated set $\{c_1, \ldots, c_M\} \subseteq \mathcal{C}$ such that $\rho(c_i, c_j) \ge \delta$ for all $i \ne j$. The transformer model class is defined as $\mathcal{T}_N := \{f_\Theta \mid \Theta \in \mathcal{S}_N \times \mathcal{F}_N\}$.

To bound the overall risk, we first decompose into the approximation and generalization gaps.

Theorem 3.1 (Schmidt-Hieber (2020), Lemma 4, adapted)**.** *There exists a universal constant C such that for any $\epsilon > 0$ such that $\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon) \geq 1$,*

$$\bar{R}(\widehat{\Theta}) \leq 2 \inf_{\Theta \in \mathcal{S}_N \times \mathcal{F}_N} R(\Theta) + C\left(\frac{B^2+\sigma^2}{T}\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon) + (B+\sigma)\epsilon\right).$$

Proof. The convergence rate of the empirical risk minimizer is established for a fixed regression problem $y = f^\circ(z) + \xi$ in Schmidt-Hieber (2020) when ξ is Gaussian; we modify the proof to incorporate bounded noise in Appendix B.1. The ICL setup can be reduced to the ordinary case as follows. We consider the entire batch $(\beta, \mathbf{X}, \xi_{1:n}, \tilde{x})$ including the hidden coefficient β as a single datum z with output $\tilde{y}$. The true function is given as $f^\circ(z) = F^\circ_\beta(\tilde{x})$ and the model class is taken to be $\mathcal{T}_N$ implicitly concatenated with the generative process $(\beta, \mathbf{X}, \xi_{1:n}, \tilde{x}) \mapsto (\mathbf{X}, \boldsymbol{y}, \tilde{x})$. Then $R(\Theta)$, $\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon)$ and T agree with the ordinary L^2 risk, model class entropy and sample size. □

Here, the second term is the *pretraining* generalization error dependent on the number of tasks T; the *in-context* generalization error dependent on the prompt length n manifests as part of the first term. This separation allows us to compare the relative difficulty of the two types of learning.

Bounding approximation error. In order to bound the first term, we analyze the risk of the choice

$$\Theta^* = (\Gamma^*, \phi^*) := \left(\left(\Sigma_{\Psi,N} + \tfrac{1}{n}\Sigma^{-1}_{\bar{\beta},N}\right)^{-1}, \phi^*_{\underline{N}:\bar{N}}\right)$$

where ϕ^* is given as in Assumption 3 for a suitable δ_N to be determined. The definition of Γ^* approximately generalizes the global optimum $\Gamma = \left((1+\frac{1}{n})\Lambda + \frac{1}{n}\mathrm{tr}(\Lambda)\mathbf{I}_d\right)^{-1}$ for the Gaussian linear regression setup where $x \sim \mathcal{N}(0, \Lambda)$ (Zhang et al., 2023). Since $\Sigma_{\Psi,N} \succeq C_1\mathbf{I}_N$ we have $\Gamma^* \preceq C_1^{-1}\mathbf{I}_N$ and hence we may assume $\Gamma^* \in \mathcal{S}_N$ by replacing C_3 with $C_3 \vee C_1^{-1}$ if necessary.

Proposition 3.2. *Under Assumptions 1-3, it holds that*

$$\inf_{\Theta \in \mathcal{S}_N \times \mathcal{F}_N} R(\Theta) \leq R(\Theta^*) \lesssim \frac{N^{2r}}{n}\log N + \frac{N^{4r}}{n^2}\log^2 N + \frac{N}{n} + N^{-2s} + N^2\delta_N^4 + N^{2r+1}\delta_N^2.$$

The proof is presented throughout Appendix A. The overall scheme is to approximate $F^\circ_\beta(\tilde{x})$ by its truncation $F^\circ_{\beta,\bar{N}}(\tilde{x})$ and the finite basis $\psi^\circ_{\underline{N}:\bar{N}}$ by $\phi^*_{\underline{N}:\bar{N}}$, which incurs errors N^{-2s} and the terms pertaining to δ_N, respectively. The first three terms arise from the concentration of the n token representations $\psi^\circ_{\underline{N}:\bar{N}}(x_k)$. All hidden constants are at most polynomial in problem parameters.

Bounding generalization error. To estimate the metric entropy of $\mathcal{T}_N$, we first reduce to the metric entropy of the representation class $\mathcal{F}_N$. Here, $\|\cdot\|_{L^\infty}$ refers to the essential supremum over the support of $\mathcal{P}_\mathcal{X}$ and also over all N components for $\mathcal{F}_N$. The proof is given in Appendix B.2.

Lemma 3.3. *Under Assumptions 1-3, there exists $D > 0$ such that for all ϵ sufficiently small,*

$$\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon) \lesssim N^2 \log\frac{B'^2_N}{\epsilon} + \mathcal{V}\left(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \frac{\epsilon}{DB'_N\sqrt{N}}\right).$$

4 Minimax Optimality of In-Context Learning

4.1 Besov Space and DNNs

We now apply our theory to study the sample complexity of ICL when $\mathcal{F}_N$ consists of (clipped, see (6)) deep neural networks. These can be also seen as simplified transformers with attention layers and skip connections removed. To be precise, we define the set of DNNs with depth L, width W, sparsity S, norm bound M and ReLU activation $\eta(x) = x \vee 0$ (applied element-wise) as

$$\mathcal{F}_{\mathrm{DNN}}(L, W, S, M) = \left\{(\mathbf{W}^{(L)}\eta + b^{(L)}) \circ \cdots \circ (\mathbf{W}^{(1)}x + b^{(1)}) \,\middle|\, \mathbf{W}^{(1)} \in \mathbb{R}^{W\times d}, \mathbf{W}^{(\ell)} \in \mathbb{R}^{W\times W},\right.$$

$$\left.\mathbf{W}^{(L)} \in \mathbb{R}^W, b^{(\ell)} \in \mathbb{R}^W, b^{(L)} \in \mathbb{R}, \sum_{\ell=1}^L \|\mathbf{W}^{(\ell)}\|_0 + \|b^{(\ell)}\|_0 \leq S, \max_{1\leq\ell\leq L}\|\mathbf{W}^{(\ell)}\|_\infty \vee \|b^{(\ell)}\|_\infty \leq M\right\}.$$

The *Besov space* is a very general class of functions including the Hölder and Sobolev spaces which captures spatial inhomogeneity in smoothness, and provides a natural setting in which to study the expressive power of deep neural networks (Suzuki, 2019). Here, we fix $\mathcal{X} = [0,1]^d$ for simplicity.

Definition 4.1 (Besov space)**.** For $2 \leq p \leq \infty, 0 < q \leq \infty$, fractional smoothness $\alpha > 0$ and $r = \lfloor \alpha \rfloor + 1$, the rth modulus of $f \in L^p(\mathcal{X})$ is defined using the difference operator $\Delta_h^r, h \in \mathbb{R}^d$ as

$$w_{r,p}(f,t) := \sup_{\|h\|_2 \leq t} \|\Delta_h^r(f)\|_p, \quad \Delta_h^r(f)(x) = 1_{\{x,x+rh \in \mathcal{X}\}} \textstyle\sum_{j=0}^r \binom{r}{j} (-1)^{r-j} f(x+jh).$$

Also, the Besov (quasi-)norm is given as $\|\cdot\|_{B_{p,q}^\alpha} = \|\cdot\|_{L^p} + |\cdot|_{B_{p,q}^\alpha}$ where

$$|f|_{B_{p,q}^\alpha} := \begin{cases} \left(\int_0^\infty t^{-q\alpha} w_{r,p}(f,t)^q \frac{\mathrm{d}t}{t}\right)^{1/q} & q < \infty \\ \sup_{t>0} t^{-\alpha} w_{r,p}(f,t) & q = \infty \end{cases}$$

and the Besov space is defined as $B_{p,q}^\alpha(\mathcal{X}) = \{f \in L^p(\mathcal{X}) \mid \|f\|_{B_{p,q}^\alpha} < \infty\}$. We write $\mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))$ for the unit ball in $(B_{p,q}^\alpha(\mathcal{X}), \|\cdot\|_{B_{p,q}^\alpha})$.

We have that the Hölder space $C^\alpha(\mathcal{X}) = B_{\infty,\infty}^\alpha(\mathcal{X})$ for order $\alpha > 0, \alpha \notin \mathbb{N}$ and the Sobolev space $W_2^m(\mathcal{X}) = B_{2,2}^m(\mathcal{X})$ for $m \in \mathbb{N}$ as well as the embeddings $B_{p,1}^m(\mathcal{X}) \hookrightarrow W_p^m(\mathcal{X}) \hookrightarrow B_{p,\infty}^m(\mathcal{X})$; if $\alpha > d/p$, $B_{p,q}^\alpha(\mathcal{X})$ compactly embeds into the space of continuous functions on $\mathcal{X}$. See Triebel (1983); Giné and Nickl (2015) for more details. The difficulty of learning a regression function in the Besov is quantified by the minimax risk; the following rate is classical.

Proposition 4.2 (Donoho and Johnstone (1998))**.** *The minimax risk for an estimator $\widehat{f}_n$ with n i.i.d. samples $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ over $\mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))$ satisfies*

$$\inf_{\widehat{f}_n : \mathcal{D}_n \to \mathbb{R}} \sup_{f^\circ \in \mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))} \mathbb{E}_{\mathcal{D}_n}[\|f^\circ - \widehat{f}_n\|_{L^2(\mathcal{X})}^2] \asymp n^{-\frac{2\alpha}{2\alpha+d}}.$$

A natural basis system for $B_{p,q}^\alpha(\mathcal{X})$ is formed by the ***B-splines***, which can be seen as a type of wavelet decomposition or multiresolution analysis (DeVore and Popov, 1988). As B-splines are piecewise polynomials, they can be efficiently approximated by DNNs with at most log depth (Suzuki, 2019).

Definition 4.3 (B-spline wavelet basis)**.** The tensor product B-spline of order $m \in \mathbb{N}$ satisfying $m > \alpha + 1 - 1/p$, at resolution $k \in \mathbb{Z}_{\geq 0}^d$ and location $\ell \in I_k^d = \prod_{i=1}^d [-m : 2^{k_i}]$ is

$$\omega_{k,\ell}^d(x) = \prod_{i=1}^d \iota_m(2^{k_i} x_i - \ell_i), \quad \text{where} \quad \iota_m(x) = (\underbrace{\iota * \iota * \cdots * \iota}_{m+1})(x), \quad \iota(x) = 1_{\{x \in [0,1]\}}.$$

When $k_1 = \cdots = k_d$, we abuse notation and write $\omega_{k,\ell}^d$ for $k \in \mathbb{Z}_{\geq 0}$ in place of $\omega_{(k,\cdots,k),\ell}^d$.

4.2 Estimation Error Analysis

To apply our framework, we set the task class as the unit ball $\mathcal{F}^\circ = \mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))$ and take as basis $\{\psi_j^\circ \mid j \in \mathbb{N}\} = \{2^{kd/2}\omega_{k,\ell}^d \mid k \in \mathbb{Z}_{\geq 0}, \ell \in I_k^d\}$ the set of all B-spline wavelets ordered primarily by increasing k and scaled to counteract the dilation in x. Abusing notation, we also write $\beta_{k,\ell}$ to denote the coefficient in β corresponding to $2^{kd/2}\omega_{k,\ell}^d$. The set of B-splines at each resolution are independent, while those of lower resolution can always be decomposed into a linear sum of B-splines of higher resolution satisfying certain decay rates, which we prove in Proposition C.10.

From this setup, in Appendix C.1.1, we verify Assumptions 1 and 2 with $r = 1/2, s = \alpha/d$ under:

Assumption 4. *$\mathcal{F}^\circ = \mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))$, $\alpha > d/p$ and $\mathcal{P}_\mathcal{X}$ has positive Lebesgue density $\rho_\mathcal{X}$ bounded above and below on $\mathcal{X}$. Also, all coefficients $\beta_{k,\ell}$ are independent and*

$$\mathbb{E}_\beta[\beta_{k,\ell}] = 0, \quad 0 < \mathbb{E}_\beta[\beta_{k,\ell}^2] \lesssim 2^{-k(2\alpha+d)} k^{-2}, \quad \forall k \geq 0, \ \ell \in I_k^d. \tag{5}$$

We can check that we have not given ourselves an easier learning problem with (5): the assumed variance decay rate is tight (up to the logarithmic factor k^{-2}) in the sense that any $f \in \mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))$ can indeed be expanded into a sum of wavelets with the same coefficient decay when averaged over $\ell \in I_k^d$. See Lemma C.1 and the following discussion. We also obtain the following in-context approximation and entropy bounds in Appendix C.1.2.

Lemma 4.4. *For any $\delta_N > 0$, Assumption 3 is satisfied by taking*

$$\mathcal{F}_N = \{\Pi_{B'_N} \circ \phi \mid \phi = (\phi_j)_{j=1}^N, \phi_j \in \mathcal{F}_{\mathrm{DNN}}(L, W, S, M)\} \tag{6}$$

where $\Pi_{B'_N}$ is the projection in $\mathbb{R}^N$ to the centered ball of radius $B'_N = O(\sqrt{N})$ and each ϕ_j is a ReLU network such that $L = O(\log N + \log \delta_N^{-1})$ and $W, S, M = O(1)$. Also, the metric entropy of $\mathcal{F}_N$ is bounded as $\mathcal{V}(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \epsilon) \lesssim N \log \frac{N}{\delta_N \epsilon}$.

Hence Assumptions 1-3 all follow from Assumption 4, and we conclude in Appendix C.1.3:

Theorem 4.5 (minimax optimality of ICL in Besov space)**.** *Under Assumption 4, if $n \gtrsim N \log N$,*

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-\frac{2\alpha}{d}} \begin{pmatrix}\text{DNN}\\ \text{approximation}\\ \text{error}\end{pmatrix} + \frac{N \log N}{n} \begin{pmatrix}\text{in-context}\\ \text{generalization}\\ \text{error}\end{pmatrix} + \frac{N^2 \log N}{T} \begin{pmatrix}\text{pretraining}\\ \text{generalization}\\ \text{error}\end{pmatrix}.$$

Hence if $T \gtrsim n^{\frac{2\alpha+2d}{2\alpha+d}}$ and $N \asymp n^{\frac{d}{2\alpha+d}}$, in-context learning achieves the minimax optimal rate $n^{-\frac{2\alpha}{2\alpha+d}}$ up to a log factor.

The first term arises from the N-term truncation and oracle approximation error of the DNN module, and is equal to the N-term optimal error (Dũng, 2011a). The second and third term each correspond to the in-context and pretraining generalization gap. With regard to N, we see that $n = \widetilde{\Omega}(N)$ is enough to learn the basis expansion in context, while $T = \widetilde{\Omega}(N^2)$ is necessary to learn the attention layer. However if $T/N = o(n)$, the third term dominates and the overall complexity scales suboptimally as $T^{-\frac{\alpha}{\alpha+d}}$, illustrating the importance of sufficient pretraining. This also aligns with the task diversity threshold observed by Raventos et al. (2023). Since the amount of training data for LLMs is practically infinite in practice, our result justifies the effectiveness of ICL at large scales with only a small number of in-context samples.

A limitation of ICL. In the regime $1 \leq p < 2$, the approximation error is strictly worse without an *adaptive* representation scheme and the resulting rate is suboptimal (see Remark C.3). While DNNs can adapt to task smoothness in supervised settings (Suzuki, 2019), ICL and any other meta-learning methods are fundamentally constrained to non-adaptive representations since they cannot update at inference time, and hence are bounded below by the best linear approximation rate or Kolmogorov width, which is strictly worse than the minimax optimal rate when $p < 2$. Indeed, for any N-dimensional subspace $\mathcal{L}_N \subset B^\alpha_{p,q}(\mathcal{X})$ it holds that (Vybíral, 2008)

$$\inf_{\mathcal{L}_N} \sup_{f^\circ \in \mathbb{U}(B^\alpha_{p,q}(\mathcal{X}))} \inf_{\ell_n \in \mathcal{L}_N} \|f^\circ - \ell_n\|_{L^2(\mathcal{P}_\mathcal{X})} \gtrsim N^{-\alpha/d + (1/p - 1/2)_+}.$$

Remark 4.6. The $N^2 \log N$ term in the pretraining generalization gap is due to the covering bound of the attention matrix Γ, while the entropy of the DNN class is only $N \log N$. Hence the task diversity requirement may be lessened to the latter by considering low-rank structure or approximation of attention heads (Bhojanapalli et al., 2020; Chen et al., 2021).

4.3 Avoiding the Curse of Dimensionality

The above rate inevitably suffers from the curse of dimensionality as d appears in the exponent of the optimal rate. We also consider the *anisotropic Besov space* (Nikol'skii, 1975), a generalization allowing for different degrees of smoothness $(\alpha_1, \cdots, \alpha_d)$ in each coordinate. Then the optimal rate is nearly dimension-free in the sense that the rate only depends on d through the quantity $\widetilde{\alpha} := (\sum_i \alpha_i^{-1})^{-1}$, and becomes independent of dimension if only a few directions are important i.e. have small α_i. Rigorous definitions, statements and proofs are provided in Appendix C.2.

Extending Theorem 4.5, we show that ICL again attains near-optimal estimation error in the anisotropic Besov space, circumventing the curse of dimensionality and theoretically establishing the efficiacy of in-context learning in high-dimensional settings.

Theorem 4.7 (informal version of Theorem C.7)**.** *For the anisotropic Besov space of smoothness $(\alpha_1, \cdots, \alpha_d)$, assume variance decay (4) with $s = \widetilde{\alpha} > 1/p$. If $T \gtrsim nN$ and $N \asymp n^{\frac{1}{2\widetilde{\alpha}+1}}$, in-context learning achieves the minimax optimal rate $n^{-\frac{2\widetilde{\alpha}}{2\widetilde{\alpha}+1}}$ up to a log factor.*

4.4 Learning a Coarser Basis

Thus far, we have demonstrated the importance of sufficient pretraining to achieve optimal risk; as another application of our framework, we illustrate how pretraining can actively *mitigate* the complexity of in-context learning. Consider the case where $(\psi_j^\circ)_{j=1}^\infty$ is no longer the B-spline basis of $B_{p,q}^\alpha(\mathcal{X})$ but instead is chosen from some wider function space, say the unit ball of $B_{p,q}^\tau(\mathcal{X})$ for a smaller smoothness $\tau < \alpha$. Without knowledge of the basis, the sample complexity of learning any regression function F_β° is a priori lower bounded by the minimax rate $n^{-\frac{2\tau}{2\tau+d}}$ by Proposition 4.2. For ICL, this difficulty manifests as an increase in the metric entropy of the class $\mathcal{F}_N$ which must be powerful enough to approximate $\psi_{1:N}^\circ$ (Corollary C.12), giving rise to the modified risk bound:

Corollary 4.8 (ICL for coarser basis)**.** *Suppose $\alpha > \tau > d/p$, the basis $(\psi_j^\circ)_{j=1}^\infty \subset \mathbb{U}(B_{p,q}^\tau(\mathcal{X}))$ and Assumptions 1, 2 hold with $r = 1/2, s = \alpha/d$. Then if $n \gtrsim N \log N$,*

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-\frac{2\alpha}{d}} + \frac{N \log N}{n} + \frac{N^{1+\frac{\alpha}{\tau}+\frac{d}{\tau}} \log^3 N}{T}.$$

Hence if $T \gtrsim n^{1+\frac{d}{2\alpha+d}\frac{\alpha+d}{\tau}}$ and $N \asymp n^{\frac{d}{2\alpha+d}}$, the risk converges as $n^{-\frac{2\alpha}{2\alpha+d}} \log n$.

The pretraining generalization gap is now dominated by the higher complexity $N^{1+\frac{\alpha}{\tau}+\frac{d}{\tau}} \log^3 N$ of the DNN class and strictly worse compared to $N^2 \log N$ for Theorem 4.5. The required number of tasks also suffers and the exponent is no longer $\frac{2\alpha+2d}{2\alpha+d} \in (1,2)$ but scales as $O(d)$. Nevertheless, observe that the burden of complexity is entirely carried by T; with sufficient pretraining, the third term can be made arbitrarily small and the ICL risk again attains $n^{-\frac{2\alpha}{2\alpha+d}}$. Hence ICL improves upon the a priori lower bound $n^{-\frac{2\tau}{2\tau+d}}$ at inference time by encoding information on the coarser basis during pretraining. We remark that the result is also readily adapted to the anisotropic setting.

4.5 Sequential Input and Transformers

We now consider a more complex setting where the inputs $x \in [0,1]^{d\times\infty}$ are bidirectional *sequences* of tokens (e.g. entire documents) and ϕ is itself a transformer network.[2] In this infinite-dimensional setting, transformers can still circumvent the curse of dimensionality and in fact achieve near-optimal sample complexity due to their parameter sharing and feature extraction capabilities (Takakura and Suzuki, 2023). Our goal in this section is to extend this guarantee to ICL of trained transformers.

For sequential data, it is natural to suppose the smoothness w.r.t. each coordinate can vary depending on the input. For example, the position of important tokens in a sentence will change if irrelevant strings are inserted. To this end, we adopt the *piecewise γ-smooth function class* introduced by Takakura and Suzuki (2023), which allows for arbitrary bounded permutations among input tokens; see Appendix D.1 for definitions. Also borrowing from their setup, we consider multi-head sliding window self-attention layers with window size U, embedding dimension D, number of heads H with key, query, value matrices $K^{(h)}, Q^{(h)} \in \mathbb{R}^{D\times d}$, $V^{(h)} \in \mathbb{R}^{d\times d}$ and norm bound M defined as[3]

$$\mathcal{F}_{\text{Attn}}(U,D,H,M) = \Bigg\{ g : \mathbb{R}^{d\times\infty} \to \mathbb{R}^{d\times\infty} \;\Bigg|\; \max_{1\le h\le H} \|K^{(h)}\|_\infty \vee \|Q^{(h)}\|_\infty \vee \|V^{(h)}\|_\infty \le M,$$
$$g(x)_i = x_i + \sum_{h=1}^H V^{(h)} x_{i-U:i+U} \text{Softmax}\left((K^{(h)} x_{i-U:i+U})^\top Q^{(h)} x_i \right) \Bigg\}.$$

We also consider a linear embedding layer $\text{Enc}(x) = Ex + P$, $E \in \mathbb{R}^{D\times d}$ with absolute positional encoding $P \in \mathbb{R}^D$ of bounded norm. Then the class of depth J transformers is defined as

$$\mathcal{F}_{\text{TF}}(J,U,D,H,L,W,S,M) := \big\{ f_J \circ g_J \circ \cdots \circ f_1 \circ g_1 \circ \text{Enc} \mid \|E\| \le M,$$
$$f_i \in \mathcal{F}_{\text{DNN}}(L,W,S,M),\; g_i \in \mathcal{F}_{\text{Attn}}(U,D,H,M) \big\}.$$

Our result, proved in Appendix D.2, reads:

[2] We clarify that this is *not* equivalent to a multi-layer transformer setting where ϕ is the rest of the transformer. Instead, ϕ operates on individual tokens x_i separately, which may now themselves be sequences of unbounded dimension. The extracted per-token features are cross-referenced only at the final attention layer f_Θ.

[3] Here the ith column and (j,i)th component of $x \in \mathbb{R}^{d\times\infty}$ for $i \in \mathbb{Z}, j \in [d]$ are denoted by x_i and x_{ij}, respectively. These are not to be confused with sample indexes (1) as those will not be used in this section.

Theorem 4.9 (informal version of Theorem D.1)**.** *Suppose $\mathcal{F}^\circ$ consists of functions on $[0,1]^{d\times\infty}$ of bounded piecewise γ-smooth and L^∞-norm with smoothness $\alpha \in \mathbb{R}_{>0}^{d\times\infty}$, and let γ be mixed or anisotropic smoothness with $\alpha^\dagger = \max_{i,j}\alpha_{ij}$ or $(\sum_{i,j}\alpha_{ij}^{-1})^{-1}$, respectively. Under suitable regularity and decay assumptions, by taking $\mathcal{F}_N$ to be a class of clipped transformers it holds that*

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-2\alpha^\dagger} + \frac{N\log N}{n} + \frac{N^{2\vee(1+1/\alpha^\dagger)}\operatorname{polylog}(N)}{T}.$$

Hence if $T \gtrsim nN^{1\vee 1/\alpha^\dagger}$ and $N \asymp n^{\frac{1}{2\alpha^\dagger+1}}$, ICL achieves the rate $n^{-\frac{2\alpha^\dagger}{2\alpha^\dagger+1}}\operatorname{polylog}(n)$.

This matches the optimal rate in finite dimensions independently of the (possibly infinite) length of the input or context window. The dynamical feature extraction ability of attention layers in the $\mathcal{F}_{\mathrm{TF}}$ class is essential in dealing with input-dependent smoothness, further justifying the efficiacy of ICL of sequential data.

5 Minimax Lower Bounds

In this section, we provide lower bounds for the minimax rate in both n, T by extending the theory of Yang and Barron (1999), which can be leveraged to yield results stronger than optimality in merely n. The bound is purely information-theoretic and hence applies to not just ICL but any meta-learning scheme for the regression problem of Section 2.1 from the data $\mathcal{D}_{n,T} = \{(\mathbf{X}^{(t)}, \boldsymbol{y}^{(t)})\}_{t=1}^{T+1}$, where the index $T+1$ corresponds to the test task.

For this section we assume that the noise (1) is i.i.d. Gaussian, $\xi_k \sim \mathcal{N}(0,\sigma^2)$, instead of bounded; while the exact shape of the noise distribution is not important, having restricted support may convey additional information and affect the minimax rate. We also suppose for simplicity that the support of $\mathcal{P}_\beta$ is included in $\mathcal{B} := \{\beta \in \mathbb{R}^\infty \mid |\beta_j|^2 \lesssim j^{-2s-1}(\log j)^{-2},\ j \in \mathbb{N}\}$ and that the aggregated coefficients $\bar{\beta}_j$ for $\underline{N} \le j \le \bar{N}$ satisfy $\mathbb{E}_\beta[\bar{\beta}_j^2] \le \sigma_\beta^2$ for some σ_β dependent on N. The proof of the following statement is given in Appendix F.1.

Proposition 5.1. *For $\varepsilon_{n,1}, \varepsilon_{n,2}, \delta_n > 0$, let Q_1 and Q_2 be the $\varepsilon_{n,1}$- and $\varepsilon_{n,2}$-covering numbers of $\mathcal{F}_N$ and $\mathcal{B}$ respectively, and M be the δ_n-packing number of $\mathcal{F}^\circ$. Suppose that the following conditions are satisfied:*

$$\tfrac{1}{2\sigma^2}\left(n(T+1)\sigma_\beta^2\varepsilon_{n,1}^2 + C_2 n\varepsilon_{n,2}^2\right) \le \log Q_1 + \log Q_2 \le \tfrac{1}{8}\log M, \quad 4\log 2 \le \log M. \tag{7}$$

Then the minimax rate is lower bounded as

$$\inf_{\widehat{f}:\mathcal{D}_{n,T}\to\mathbb{R}} \sup_{f^\circ\in\mathcal{F}^\circ} \mathbb{E}_{\mathcal{D}_{n,T}}[\|\widehat{f} - f^\circ\|_{L^2(\mathcal{P}_\mathcal{X})}^2] \ge \tfrac{1}{4}\delta_n^2.$$

Finally, Proposition 5.1 is applied to obtain concrete lower bounds for the settings studied in Section 4 throughout Appendices F.2-F.4.

Corollary 5.2 (minimax lower bound)**.** *The minimax rates in the previous regression settings are lower bounded as follows.*

(i) *Besov space (Section 4.2):* $\inf_{\widehat{f}}\sup_{f^\circ} \mathbb{E}_{\mathcal{D}_{n,T}}[\|\widehat{f} - f^\circ\|^2] \gtrsim n^{-\frac{2\alpha}{2\alpha+d}}$,

(ii) *Coarser basis (Section 4.4):* $\inf_{\widehat{f}}\sup_{f^\circ} \mathbb{E}_{\mathcal{D}_{n,T}}[\|\widehat{f} - f^\circ\|^2] \gtrsim n^{-\frac{2\alpha}{2\alpha+d}} + (nT)^{-\frac{2\tau}{2\tau+d}}$,

(iii) *Sequential input (Section 4.5):* $\inf_{\widehat{f}}\sup_{f^\circ} \mathbb{E}_{\mathcal{D}_{n,T}}[\|\widehat{f} - f^\circ\|^2] \gtrsim n^{-\frac{2\alpha^\dagger}{2\alpha^\dagger+1}}$.

These results match the upper bounds for (i), (iii) and show that **ICL is provably jointly optimal** in n, T in the 'large T' regime. Moreover, we can check for the coarser basis setting that insufficient pretraining $T = O(1)$ indeed leads to the worse complexity $n^{-\frac{2\tau}{2\tau+d}}$, while the faster rate $n^{-\frac{2\alpha}{2\alpha+d}}$ is retrieved when $T \gtrsim n^{\frac{(\alpha-\tau)d}{(2\alpha+d)\tau}}$. This aligns with the discussion in Section 4.4, showing that ICL is **provably suboptimal** in the 'small T' regime.

Remark 5.3. The obtained upper and lower bounds in the coarser basis setting are not tight as T varies, hence it remains to be shown whether there exists a meta-learning algorithm that attains the lower bound (ii). The task diversity threshold for optimal learning suggested by the bounds are also different ($n^{1+\frac{d(\alpha+d)}{(2\alpha+d)}}$ v.s. $n^{\frac{(\alpha-\tau)d}{(2\alpha+d)\tau}}$); it would be interesting for future work to resolve this gap.

6 Conclusion

In this paper, we performed a learning-theoretic analysis of ICL of a transformer consisting of a DNN and a linear attention layer pretrained on nonparametric regression tasks. We developed a general framework for bounding the in-context estimation error of the empirical risk minimizer in terms of both the number of tasks and samples, and proved that ICL can achieve nearly minimax optimal rates in the Besov space, anisotropic Besov space and γ-smooth class. We also demonstrated that ICL can improve upon the a priori optimal rate by learning informative representations during pretraining. We supplemented our analyses with corresponding minimax lower bounds jointly in n, T and also performed numerical experiments validating our findings. Our work opens up interesting approaches of adapting classical learning theory to study emergent phenomena of foundation models.

Limitations. Our transformer model is limited to a single layer of linear self-attention and does not consider more complex in-context learning behavior which may arise in transformers with multiple attention layers. Moreover, the obtained upper and lower bounds are not tight in certain regimes, suggesting future research directions for meta-learning.

Acknowledgments

JK was partially supported by JST CREST (JPMJCR2015). TS was partially supported by JSPS KAKENHI (24K02905, 20H00576) and JST CREST (JPMJCR2115). We would like to express our gratitude to Masaaki Imaizumi for valuable and insightful discussions on the topic in relation to his work in progress (Imaizumi, 2024).

References

K. Ahn, X. Cheng, H. Daneshmand, and S. Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

Y. Bai, F. Chen, H. Wang, C. Xiong, and S. Mei. Transformers as statisticians: provable in-context learning with in-context algorithm selection. In *ICML Workshop on Efficient Systems for Foundation Models*, 2023.

S. Bhojanapalli, C. Yun, A. S. Rawat, S. J. Reddi, and S. Kumar. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning*, 2020.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

B. Chen, T. Dao, E. Winsor, Z. Song, A. Rudra, and C. Ré. Scatterbrain: unifying sparse and low-rank attention approximation. In *Advances in Neural Information Processing Systems*, 2021.

S. Chen, H. Sheen, T. Wang, and Z. Yang. Training dynamics of multi-head softmax attention for in-context learning: emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

R. A. DeVore and V. A. Popov. Interpolation of Besov spaces. *Transactions of the American Mathematical Society*, 305(1):397–414, 1988.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

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

D. Dũng. Optimal adaptive sampling recovery. *Advances in Computational Mathematics*, 34:1–41, 2011a.

D. Dũng. B-spline quasi-interpolant representations and sampling recovery of functions with mixed smoothness. *Journal of Complexity*, 27(6):541–567, 2011b.

X. García, Y. Bansal, C. Cherry, G. F. Foster, M. Krikun, F. Feng, M. Johnson, and O. Firat. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, 2023.

S. Garg, D. Tsipras, P. Liang, and G. Valiant. What can Transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

T. Guo, W. Hu, S. Mei, H. Wang, C. Xiong, S. Savarese, and Y. Bai. How do Transformers learn in-context beyond simple functions? A case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.

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

Y. Huang, Y. Cheng, and Y. Liang. In-context convergence of Transformers. *arXiv preprint arXiv:2310.05249*, 2023.

M. Imaizumi. Statistical analysis on in-context learning, 2024. Personal communication.

J. Kim and T. Suzuki. Transformers learn nonlinear features in context: nonconvex mean-field dynamics on the attention landscape. In *International Conference on Machine Learning*, 2024.

H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen. Training nonlinear Transformers for efficient in-context learning: a theoretical learning and generalization analysis. *arXiv preprint arXiv:2402.15607*, 2024.

A. Mahankali, T. B. Hashimoto, and T. Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

D. Meunier, Z. Li, A. Gretton, and S. Kpotufe. Nonlinear meta-learning can guarantee faster rates. *arXiv preprint arXiv:2307.10870*, 2023.

S. M. Nikol'skii. *Approximation of functions of several variables and imbedding theorems*, volume 205 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, 1975.

Y. Nishimura and T. 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*, 2024. URL https://openreview.net/forum?id=EW8ZExRZkJ.

S. Okumoto and T. Suzuki. Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. In *International Conference on Learning Representations*, 2022.

A. Raventos, M. Paul, F. Chen, and S. Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, 2023.

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

T. 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.

T. Suzuki and A. Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, 2021.

S. J. Szarek. Nets of Grassmann manifold and orthogonal group. *Proceedings of Banach Space Workshop*, pages 169–186, 1981.

S. Takakura and T. Suzuki. Approximation and estimation ability of Transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, 2023.

H. Triebel. *Theory of function spaces*. Monographs in mathematics. Birkhäuser Verlag, 1983.

H. Triebel. Entropy numbers in function spaces with mixed integrability. *Revista Matematica Complutense*, 24:169–188, 2011.

N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, 2020.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, May 2015. ISSN 1935-8237.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, New York, 1996.

J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

J. Vybíral. *Function spaces with dominating mixed smoothness*, volume 30 of *Lectures in Mathematics*. European Mathematical Society, 2006.

J. Vybíral. Widths of embeddings in function spaces. *Journal of Complexity*, 24(4):545–570, 2008.

J. Wu, D. Zou, Z. Chen, V. Braverman, Q. Gu, and P. L. Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, 2024.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2016.

R. Zhang, S. Frei, and P. L. Bartlett. Trained Transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

R. Zhang, J. Wu, and P. L. Bartlett. In-context learning of a linear Transformer block: benefits of the MLP component and one-step GD initialization. *arXiv preprint arXiv:2402.14951*, 2024.

Table of Contents

Appendix

A Proof of Proposition 3.2

A.1 Decomposing Approximation Error

Recall that

$$\Theta^* = (\Gamma^*, \phi^*) = \left(\left(\Sigma_{\Psi,N} + \frac{1}{n}\Sigma_{\bar{\beta},N}^{-1}\right)^{-1}, \phi^*_{\underline{N}:\bar{N}}\right).$$

We introduce some additional notation in the following fashion. For brevity, we write $\bar{N}:\infty$ instead of $(\bar{N}+1):\infty$ as an exception.

$$\begin{aligned}
&\Phi^* = (\phi^*_{\underline{N}:\bar{N}}(x_1), \cdots, \phi^*_{\underline{N}:\bar{N}}(x_n)) \in \mathbb{R}^{N\times n}, \quad \xi = (\xi_1, \dots, \xi_n)^\top \in \mathbb{R}^n, \\
&\Psi^\circ = (\psi^\circ(x_1), \dots, \psi^\circ(x_n)) \in \mathbb{R}^{\infty\times n}, \quad \Psi^\circ_{\underline{N}:\bar{N}} = (\psi^\circ_{\underline{N}:\bar{N}}(x_1), \cdots, \psi^\circ_{\underline{N}:\bar{N}}(x_n)) \in \mathbb{R}^{N\times n}, \\
&\Psi^\circ_{\underline{N}:\infty} = (\psi^\circ_{\underline{N}:\infty}(x_1), \cdots, \psi^\circ_{\underline{N}:\infty}(x_n)) \in \mathbb{R}^{\infty\times n}.
\end{aligned}$$

Since clipping $\check{f}_\Theta$ does not make its difference with $F^\circ_\beta \in [-B, B]$ larger, it holds that

$$\begin{aligned}
R(\Theta^*) &= \mathbb{E}\left[\left(F^\circ_\beta(\tilde{x}) - f_{\Theta^*}(\mathbf{X}, \boldsymbol{y}, \tilde{x})\right)^2\right] \\
&\leq \mathbb{E}\left[\left(F^\circ_\beta(\tilde{x}) - \check{f}_{\Theta^*}(\mathbf{X}, \boldsymbol{y}, \tilde{x})\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(F^\circ_\beta(\tilde{x}) - F^\circ_{\beta,\bar{N}}(\tilde{x})\right)^2\right] + 2\mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x}) - \check{f}_{\Theta^*}(\mathbf{X}, \boldsymbol{y}, \tilde{x})\right)^2\right] \\
&\lesssim N^{-2s} + \mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x}) - \phi^*(\tilde{x})^\top \frac{\Gamma^*\Phi^*\boldsymbol{y}}{n}\right)^2\right]
\end{aligned}$$

due to Assumption 2 and $\bar{N} \asymp N$. Expanding the attention output as

$$\begin{aligned}
\phi^*(\tilde{x})^\top \frac{\Gamma^*\Phi^*\boldsymbol{y}}{n} &= (\phi^*(\tilde{x}) - \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}) + \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top \Gamma^* \frac{(\Phi^* - \Psi^\circ_{\underline{N}:\bar{N}} + \Psi^\circ_{\underline{N}:\bar{N}})\boldsymbol{y}}{n} \\
&= (\phi^*(\tilde{x}) - \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top \Gamma^* \frac{(\Phi^* - \Psi^\circ_{\underline{N}:\bar{N}})\boldsymbol{y}}{n} \\
&\quad + (\phi^*(\tilde{x}) - \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}\boldsymbol{y}}{n} \\
&\quad + \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{(\Phi^* - \Psi^\circ_{\underline{N}:\bar{N}})\boldsymbol{y}}{n} \\
&\quad + \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}\boldsymbol{y}}{n},
\end{aligned}$$

and the final term further as

$$\begin{aligned}
\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}\boldsymbol{y}}{n} &= \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}(\Psi^{\circ\top}\beta + \xi)}{n} \\
= \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}\bar{\beta}_{\underline{N}:\bar{N}}}{n} &+ \psi^\circ_{1:N}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}(\Psi^{\circ\top}_{\bar{N}:\infty}\beta_{\bar{N}:\infty} + \xi)}{n},
\end{aligned}$$

we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x}) - \phi^*(\tilde{x})^\top \frac{\Gamma^*\,\Phi^*\boldsymbol{y}}{n}\right)^2\right] \\
&\lesssim \mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x}) - \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}\bar{\beta}_{\underline{N}:\bar{N}}}{n}\right)^2\right] \qquad (8)
\end{aligned}$$

$$+\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}(\Psi^{\circ\top}_{\bar{N}:\infty}\beta_{\bar{N}:\infty}+\xi)}{n}\right)^{2}\right] \tag{9}$$

$$+\mathbb{E}\left[\left((\phi^{*}(\tilde{x})-\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x}))^{\top}\Gamma^{*}\frac{(\Phi^{*}-\Psi^{\circ}_{\underline{N}:\bar{N}})\boldsymbol{y}}{n}\right)^{2}\right] \tag{10}$$

$$+\mathbb{E}\left[\left((\phi^{*}(\tilde{x})-\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x}))^{\top}\Gamma^{*}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\boldsymbol{y}}{n}\right)^{2}\right] \tag{11}$$

$$+\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\frac{(\Phi^{*}-\Psi^{\circ}_{\underline{N}:\bar{N}})\boldsymbol{y}}{n}\right)^{2}\right]. \tag{12}$$

We proceed to control each term separately, from which the statement of Proposition 3.2 will follow.

A.2 Bounding Term (8)

Since $F^{\circ}_{\beta,\bar{N}}(\tilde{x})=\psi^{\circ}_{1:\bar{N}}(\tilde{x})^{\top}\beta_{1:\bar{N}}=\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\bar{\beta}_{\underline{N}:\bar{N}}$, we can introduce a $(\Gamma^{*})^{-1}$ factor to bound

$$\begin{aligned}
&\mathbb{E}\left[\left(F^{\circ}_{\beta,\bar{N}}(\tilde{x})-\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}\bar{\beta}_{\underline{N}:\bar{N}}}{n}\right)^{2}\right]\\
&=\mathbb{E}\left[\left(F^{\circ}_{\beta,\bar{N}}(\tilde{x})-\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\left(\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-(\Gamma^{*})^{-1}+(\Gamma^{*})^{-1}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^{2}\right]\\
&=\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\left(\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-\Sigma_{\Psi,N}-\frac{1}{n}\Sigma^{-1}_{\bar{\beta},N}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^{2}\right]\\
&\leq 2\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\left(\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-\Sigma_{\Psi,N}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^{2}\right] &(13)\\
&\quad+2\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\frac{\Sigma^{-1}_{\bar{\beta},N}}{n}\bar{\beta}_{\underline{N}:\bar{N}}\right)^{2}\right]. &(14)
\end{aligned}$$

Denote the operator norm (with respect to L^2 norm of vectors) by $\|\cdot\|_{\mathrm{op}}$. For (13), noting that $\Sigma_{\Psi,N}$ is positive definite,

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\left(\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-\Sigma_{\Psi,N}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^{2}\right]\\
&=\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\Gamma^{*}\Sigma^{1/2}_{\Psi,N}\left(\Sigma^{-1/2}_{\Psi,N}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma^{-1/2}_{\Psi,N}-\mathbf{I}_{N}\right)\Sigma^{1/2}_{\Psi,N}\bar{\beta}_{\underline{N}:\bar{N}}\right)^{2}\right]\\
&=\mathbb{E}_{\mathbf{X},\beta}\left[\bar{\beta}^{\top}_{\underline{N}:\bar{N}}\Sigma^{1/2}_{\Psi,N}\left(\Sigma^{-1/2}_{\Psi,N}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma^{-1/2}_{\Psi,N}-\mathbf{I}_{N}\right)\Sigma^{1/2}_{\Psi,N}\Gamma^{*}\mathbb{E}_{\tilde{x}}\left[\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^{\top}\right]\right.\\
&\qquad\left.\times\Gamma^{*}\Sigma^{1/2}_{\Psi,N}\left(\Sigma^{-1/2}_{\Psi,N}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma^{-1/2}_{\Psi,N}-\mathbf{I}_{N}\right)\Sigma^{1/2}_{\Psi,N}\bar{\beta}_{\underline{N}:\bar{N}}\right]\\
&=\mathbb{E}_{\mathbf{X}}\left[\mathrm{Tr}\left[\left(\Sigma^{-1/2}_{\Psi,N}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma^{-1/2}_{\Psi,N}-\mathbf{I}_{N}\right)\Sigma^{1/2}_{\Psi,N}\mathbb{E}_{\beta}\left[\bar{\beta}_{\underline{N}:\bar{N}}\bar{\beta}^{\top}_{\underline{N}:\bar{N}}\right]\Sigma^{1/2}_{\Psi,N}\right.\right.\\
&\qquad\left.\left.\times\left(\Sigma^{-1/2}_{\Psi,N}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma^{-1/2}_{\Psi,N}-\mathbf{I}_{N}\right)\Sigma^{1/2}_{\Psi,N}\Gamma^{*}\Sigma_{\Psi,N}\Gamma^{*}\Sigma^{1/2}_{\Psi,N}\right]\right]
\end{aligned}$$

$$\lesssim \mathbb{E}_{\mathbf{X}}\left[\operatorname{Tr}\left[\left(\Sigma_{\Psi,N}^{-1/2}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma_{\Psi,N}^{-1/2}-\mathbf{I}_N\right)^2\Sigma_{\Psi,N}^{1/2}\Sigma_{\bar{\beta},N}\Sigma_{\Psi,N}^{1/2}\right]\right]$$

due to the independence of $\mathbf{X}$, $\tilde{x}$ and β. For the last inequality, we have used the fact that both the $\Sigma_{\Psi,N}^{1/2}\Gamma^*\Sigma_{\Psi,N}\Gamma^*\Sigma_{\Psi,N}^{1/2}$ term and the matrix multiplied to it are positive semi-definite, and the former is bounded above as $\precsim \mathbf{I}_N$ by Assumption 1.

Furthermore, we utilize the following result proved in Appendix A.5:

Lemma A.1. $\mathbb{E}_{\mathbf{X}}\left[\left\|\Sigma_{\Psi,N}^{-1/2}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma_{\Psi,N}^{-1/2}-\mathbf{I}_N\right\|_{\mathrm{op}}^2\right]\lesssim \frac{N^{2r}}{n}\log N+\frac{N^{4r}}{n^2}\log^2 N.$

It follows that (13) is bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\left(\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-\Sigma_{\Psi,N}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^2\right]\\
&\lesssim \mathbb{E}_{\mathbf{X}}\left[\left\|\Sigma_{\Psi,N}^{-1/2}\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}\Sigma_{\Psi,N}^{-1/2}-\mathbf{I}_N\right\|_{\mathrm{op}}^2\right]\operatorname{Tr}\left[\Sigma_{\Psi,N}^{1/2}\Sigma_{\bar{\beta},N}\Sigma_{\Psi,N}^{1/2}\right]\\
&\lesssim\left(\frac{N^{2r}}{n}\log N+\frac{N^{4r}}{n^2}\log^2 N\right)\|\Sigma_{\Psi,N}\|_{\mathrm{op}}\operatorname{Tr}(\Sigma_{\bar{\beta},N})\\
&\lesssim\frac{N^{2r}}{n}\log N+\frac{N^{4r}}{n^2}\log^2 N
\end{aligned}$$

since $\operatorname{Tr}(\Sigma_{\bar{\beta},N})$ is bounded by Assumption 2.

Moreover for (14), we compute

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\bar{\beta}_{\underline{N}:\bar{N}}\right)^2\right]\\
&=\mathbb{E}\left[\bar{\beta}^\top_{\underline{N}:\bar{N}}\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\Gamma^*\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\bar{\beta}_{\underline{N}:\bar{N}}\right]\\
&=\mathbb{E}\left[\operatorname{Tr}\left[\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\Gamma^*\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\bar{\beta}_{\underline{N}:\bar{N}}\bar{\beta}^\top_{\underline{N}:\bar{N}}\right]\right]\\
&=\operatorname{Tr}\left[\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\Gamma^*\mathbb{E}_{\tilde{x}}\left[\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^\top\right]\Gamma^*\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\mathbb{E}_{\beta}\left[\bar{\beta}_{\underline{N}:\bar{N}}\bar{\beta}^\top_{\underline{N}:\bar{N}}\right]\right]\\
&=\operatorname{Tr}\left[\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\underbrace{\Gamma^*\Sigma_{\Psi,N}\Gamma^*}_{\text{positive definite}}\frac{\Sigma_{\bar{\beta},N}^{-1}}{n}\Sigma_{\bar{\beta},N}\right]\\
&\leq\frac{1}{n}\operatorname{Tr}\left[\left(\Sigma_{\Psi,N}+\frac{1}{n}\Sigma_{\bar{\beta},N}^{-1}\right)\Gamma^*\Sigma_{\Psi,N}\Gamma^*\right]\\
&=\frac{1}{n}\operatorname{Tr}\left[\Sigma_{\Psi,N}\left(\Sigma_{\Psi,N}+\frac{1}{n}\Sigma_{\bar{\beta},N}^{-1}\right)^{-1}\right]\leq\frac{N}{n}.
\end{aligned}$$

A.3 Bounding Term (9)

Since the sequence of covariates $x_1,\cdots,x_n$ and noise $\xi_1,\cdots,\xi_n$ are each i.i.d.,

$$\mathbb{E}\left[\left(\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Psi^{\circ}_{\underline{N}:\bar{N}}(\Psi^{\circ\top}_{\bar{N}:\infty}\beta_{\bar{N}:\infty}+\xi)}{n}\right)^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\psi^\circ_{\underline{N}:\bar{N}}(x_i)(\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x_i)+\xi_i)}{n}\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x_i)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x_i) + \sum_{i=1}^{n} \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x_i)\xi_i\right)^2\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right)^2\right] \qquad (15) \\
&\quad + \frac{n-1}{n}\mathbb{E}\left[\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x')\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x')\right] \\
&\qquad (16) \\
&\quad + \frac{\sigma^2}{n}\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\right)^2\right] \qquad (17)
\end{aligned}
$$

for independent samples $x, x', \tilde{x} \sim \mathcal{P}_{\mathcal{X}}$. We now bound the three terms separately below.

For (15), we have that

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\psi^\circ_{\underline{N}:\bar{N}}(x)^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\psi^\circ_{\bar{N}:\infty}(x)^\top\beta_{\bar{N}:\infty}\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right] \\
&= \frac{1}{n}\mathbb{E}_x\left[\psi^\circ_{\underline{N}:\bar{N}}(x)^\top \Gamma^* \Sigma_{\Psi,N}\Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\psi^\circ_{\bar{N}:\infty}(x)^\top\mathbb{E}_\beta\left[\beta_{\bar{N}:\infty}\beta^\top_{\bar{N}:\infty}\right]\psi^\circ_{\bar{N}:\infty}(x)\right] \\
&\lesssim \frac{1}{n}\mathbb{E}_x\left[\|\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2\psi^\circ_{\bar{N}:\infty}(x)^\top\mathbb{E}_\beta\left[\beta_{\bar{N}:\infty}\beta^\top_{\bar{N}:\infty}\right]\psi^\circ_{\bar{N}:\infty}(x)\right] \\
&\lesssim \frac{1}{n}\sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2 \cdot \lim_{M\to\infty}\operatorname{Tr}\left(\mathbb{E}_\beta\left[\beta_{\bar{N}:M}\beta^\top_{\bar{N}:M}\right]\mathbb{E}_x\left[\psi^\circ_{\bar{N}:M}(x)\psi^\circ_{\bar{N}:M}(x)^\top\right]\right) \\
&\lesssim \frac{1}{n}\sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2 \cdot \lim_{M\to\infty}\operatorname{Tr}\left(\mathbb{E}_\beta\left[\beta_{\bar{N}:M}\beta^\top_{\bar{N}:M}\right]\right)
\end{aligned}
$$

since $\Sigma_{\Psi,M} \preceq C_2\mathbf{I}_N$ as $M \to \infty$ by Assumption 2. As $\sup_x\|\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2 \lesssim N^{-2r}$ by (2) and the diagonal of $\mathbb{E}_\beta[\beta_{\bar{N}:M}\beta^\top_{\bar{N}:M}]$ decays faster than $\bar{N}^{-2s}M^{-1}(\log M)^{-2}$ when $M > \bar{N}$ due to (4), it follows that

$$
\frac{1}{n}\sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2 \cdot \lim_{M\to\infty}\operatorname{Tr}\left(\mathbb{E}_\beta\left[\beta_{\bar{N}:M}\beta^\top_{\bar{N}:M}\right]\right) \lesssim \frac{1}{n}N^{2r}N^{-2s}.
$$

Similarly, for (16), we have that

$$
\begin{aligned}
&\frac{n-1}{n}\mathbb{E}\left[\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x')\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x')\right] \\
&\leq \mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right)^\top\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x')\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x')\right)\right] \\
&= \mathbb{E}_{x,x',\beta}\left[\psi^\circ_{\bar{N}:\infty}(x)^\top\beta_{\bar{N}:\infty}\psi^\circ_{\underline{N}:\bar{N}}(x)^\top\Gamma^*\Sigma_{\Psi,N}\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x')\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x')\right] \\
&= \mathbb{E}_\beta\left[\mathbb{E}_x\left[\psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right]^\top\Gamma^*\Sigma_{\Psi,N}\Gamma^*\mathbb{E}_x\left[\psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right]\right] \\
&\lesssim \mathbb{E}_\beta\left[\left\|\mathbb{E}_x\left[\psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right]\right\|^2\right] \\
&\lesssim \mathbb{E}_\beta\left[\psi^\circ_{\bar{N}:\infty}(x)^\top\beta_{\bar{N}:\infty}\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right] \\
&\lesssim N^{-2s}.
\end{aligned}
$$

And for (17), we have that

$$
\frac{\sigma^2}{n}\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \psi^\circ_{\underline{N}:\bar{N}}(x)\right)^2\right]
$$

$$
\begin{aligned}
&= \frac{\sigma^2}{n}\operatorname{Tr}\left[\Gamma^*\mathbb{E}_{\tilde{x}}\left[\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\right]\Gamma^*\mathbb{E}_x\left[\psi^\circ_{\underline{N}:\bar{N}}(x)\psi^\circ_{\underline{N}:\bar{N}}(x)^\top\right]\right]\\
&= \frac{\sigma^2}{n}\operatorname{Tr}\left[\underbrace{\Gamma^*\Sigma_{\Psi,N}\Gamma^*}_{\text{positive definite}}\Sigma_{\Psi,N}\right]\\
&\le \frac{\sigma^2}{n}\operatorname{Tr}\left[\Gamma^*\Sigma_{\Psi,N}\Gamma^*\left(\Sigma_{\Psi,N}+\frac{1}{n}\Sigma^{-1}_{\bar{\beta},N}\right)\right]\\
&= \frac{\sigma^2}{n}\operatorname{Tr}\left[\Gamma^*\Sigma_{\Psi,N}\right] \le \frac{\sigma^2 N}{n}.
\end{aligned}
$$

Therefore, we obtain the following bound:

$$
\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Psi^\circ_{\underline{N}:\bar{N}}(\Psi^{\circ\top}_{\bar{N}:\infty}\beta_{\bar{N}:\infty}+\xi)}{n}\right)^2\right] \lesssim \frac{N^{2r}N^{-2s}}{n}+N^{-2s}+\frac{\sigma^2 N}{n}.
$$

A.4 Bounding Terms (10)-(12)

For (10), we use the Cauchy-Schwarz inequality and Assumption 3 to bound

$$
\begin{aligned}
&\mathbb{E}\left[\left((\phi^*(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top\Gamma^*\frac{(\Phi^*-\Psi^\circ_{\underline{N}:\bar{N}})\boldsymbol{y}}{n}\right)^2\right]\\
&\le \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left((\phi^*(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top\Gamma^*(\phi^*(x_i)-\psi^\circ_{\underline{N}:\bar{N}}(x_i))y_i\right)^2\right]\\
&\le \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\|\phi^*(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})\|^2\|\Gamma^*(\phi^*(x_i)-\psi^\circ_{\underline{N}:\bar{N}}(x_i))\|^2y_i^2\right]\\
&\le \|\Gamma^*\|^2_{\mathrm{op}}\left(\sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\phi^*(x)-\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2\right)^2\mathbb{E}\left[y^2\right]\\
&\le \|\Gamma^*\|^2_{\mathrm{op}}\left(N\max_{\underline{N}\le j\le\bar{N}}\|\phi^*_j-\psi^\circ_j\|^2_{L^\infty(\mathcal{P}_{\mathcal{X}})}\right)^2\mathbb{E}\left[y^2\right]\\
&\lesssim N^2\delta_N^4(B^2+\sigma^2).
\end{aligned}
$$

Similarly for (11) and (12), we have

$$
\begin{aligned}
&\mathbb{E}\left[\left((\phi^*(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top\Gamma^*\frac{\Psi^\circ_{\underline{N}:\bar{N}}\boldsymbol{y}}{n}\right)^2\right]\\
&\le \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left((\phi^*(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x_i)y_i\right)^2\right]\\
&\le \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\|\phi^*(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})\|^2\|\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x_i)\|^2y_i^2\right]\\
&\le \sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\phi^*(x)-\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2\|\Gamma^*\|^2_{\mathrm{op}}\sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\psi^\circ_{\underline{N}:\bar{N}}(x)\|^2\mathbb{E}\left[y^2\right]\\
&\lesssim N\delta_N^2\cdot N^{2r}(B^2+\sigma^2)=N^{2r+1}\delta_N^2(B^2+\sigma^2)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{(\Phi^*-\Psi^\circ_{\underline{N}:\bar{N}})\boldsymbol{y}}{n}\right)^2\right]\\
&\le \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*(\phi^*(x_i)-\psi^\circ_{\underline{N}:\bar{N}}(x_i))y_i\right)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\psi^{\circ}_{\underline{N}:\bar{N}}(\tilde{x})\|^2\|\Gamma^*(\phi^*(x_i)-\psi^{\circ}_{\underline{N}:\bar{N}}(x_i))\|^2 y_i^2\right]\\
&\le \sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|\psi^{\circ}_{\underline{N}:\bar{N}}(x)\|^2\|\Gamma^*\|_{\mathrm{op}}^2\sup_{x\in\operatorname{supp}\mathcal{P}_{\mathcal{X}}}\|(\phi^*(x)-\psi^{\circ}_{\underline{N}:\bar{N}}(x))\|^2\mathbb{E}\left[y^2\right]\\
&\lesssim N^{2r+1}\delta_N^2(B^2+\sigma^2).
\end{aligned}
$$

This concludes the proof of Proposition 3.2.

A.5 Proof of Lemma A.1

We will make use of the following concentration bound and its corollary, proved at the end of this subsection.

Theorem A.2 (matrix Bernstein inequality)**.** *Let $\mathbf{S}_1,\cdots,\mathbf{S}_n\in\mathbb{R}^{N\times N}$ be independent random matrices such that $\mathbb{E}[\mathbf{S}_i]=0$ and $\|\mathbf{S}_i\|_{\mathrm{op}}\le L$ almost surely for all i. Define $\mathbf{Z}=\sum_{i=1}^n\mathbf{S}_n$ and the* matrix variance statistic *$v(\mathbf{Z})$ as*

$$
v(\mathbf{Z})=\max\left\{\|\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]\|_{\mathrm{op}},\|\mathbb{E}[\mathbf{Z}^\top\mathbf{Z}]\|_{\mathrm{op}}\right\}=\max\left\{\left\|\textstyle\sum_{i=1}^n\mathbb{E}[\mathbf{S}_i\mathbf{S}_i^\top]\right\|_{\mathrm{op}},\left\|\textstyle\sum_{i=1}^n\mathbb{E}[\mathbf{S}_i^\top\mathbf{S}_i]\right\|_{\mathrm{op}}\right\}.
$$

Then it holds for all $t>0$ that

$$
\Pr\left(\|\mathbf{Z}\|_{\mathrm{op}}\ge t\right)\le 2N\exp\left(-\frac{t^2}{2v(\mathbf{Z})+\frac{2}{3}Lt}\right).
$$

Proof. See Tropp (2015), Theorem 6.1.1. □

Corollary A.3. *For matrices $\mathbf{S}_1,\cdots,\mathbf{S}_n$ satisfying the conditions of Theorem A.2, it holds that*

$$
\mathbb{E}\left[\frac{1}{n^2}\|\mathbf{Z}\|_{\mathrm{op}}^2\right]\le\frac{4v(\mathbf{Z})}{n^2}(1+\log 2N)+\frac{16L^2}{9n^2}(2+2\log 2N+(\log 2N)^2).
$$

We will apply Corollary A.3 to the matrices

$$
\mathbf{S}_i:=\Sigma_{\Psi,N}^{-1/2}\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\Sigma_{\Psi,N}^{-1/2}-\mathbf{I}_N.
$$

It is straightforward to verify that

$$
\mathbb{E}[\mathbf{S}_i]=\Sigma_{\Psi,N}^{-1/2}\Sigma_{\Psi,N}\Sigma_{\Psi,N}^{-1/2}-\mathbf{I}_N=0
$$

and

$$
\|\mathbf{S}_i\|_{\mathrm{op}}\lesssim\|\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\|_{\mathrm{op}}+1=\sum_{j=\underline{N}}^{\bar{N}}\psi_j^{\circ}(x_i)^2+1\lesssim N^{2r}
$$

almost surely by Assumption 1.

Next, we evaluate the matrix variance statistic. Since each $\mathbf{S}_i$ is symmetric,

$$
\begin{aligned}
v(\mathbf{Z})&=\left\|\sum_{i=1}^n\mathbb{E}[\mathbf{S}_i\mathbf{S}_i^\top]\right\|_{\mathrm{op}}\\
&=\left\|\sum_{i=1}^n\left(\mathbb{E}_{\mathbf{X}}\left[\Sigma_{\Psi,N}^{-1/2}\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\Sigma_{\Psi,N}^{-1}\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\Sigma_{\Psi,N}^{-1/2}\right]\right.\right.\\
&\qquad\left.\left.-2\mathbb{E}_{\mathbf{X}}\left[\Sigma_{\Psi,N}^{-1/2}\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\Sigma_{\Psi,N}^{-1/2}\right]+\mathbf{I}_N\right)\right\|_{\mathrm{op}}\\
&=\left\|\sum_{i=1}^n\mathbb{E}_{\mathbf{X}}\left[\Sigma_{\Psi,N}^{-1/2}\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\Sigma_{\Psi,N}^{-1}\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)\psi^{\circ}_{\underline{N}:\bar{N}}(x_i)^\top\Sigma_{\Psi,N}^{-1/2}\right]-n\mathbf{I}_N\right\|_{\mathrm{op}}\\
&\le n+n\left\|\mathbb{E}_x\left[\Sigma_{\Psi,N}^{-1/2}\psi^{\circ}_{\underline{N}:\bar{N}}(x)\psi^{\circ}_{\underline{N}:\bar{N}}(x)^\top\Sigma_{\Psi,N}^{-1}\psi^{\circ}_{\underline{N}:\bar{N}}(x)\psi^{\circ}_{\underline{N}:\bar{N}}(x)^\top\Sigma_{\Psi,N}^{-1/2}\right]\right\|_{\mathrm{op}}
\end{aligned}
$$

for a single sample $x \sim \mathcal{P}_{\mathcal{X}}$. The second term is further bounded as

$$\begin{aligned}
&\left\| \mathbb{E}_x \left[\Sigma_{\Psi,N}^{-1/2} \psi^\circ_{\underline{N}:\bar{N}}(x) \psi^\circ_{\underline{N}:\bar{N}}(x)^\top \Sigma_{\Psi,N}^{-1} \psi^\circ_{\underline{N}:\bar{N}}(x) \psi^\circ_{\underline{N}:\bar{N}}(x)^\top \Sigma_{\Psi,N}^{-1/2} \right] \right\|_{\mathrm{op}} \\
&\leq \left\| \mathbb{E}_x \left[\Sigma_{\Psi,N}^{-1/2} \psi^\circ_{\underline{N}:\bar{N}}(x) \psi^\circ_{\underline{N}:\bar{N}}(x)^\top \Sigma_{\Psi,N}^{-1/2} \right] \right\|_{\mathrm{op}} \cdot \left\| \psi^{\circ\top}_{\underline{N}:\bar{N}} \Sigma_{\Psi,N}^{-1} \psi^\circ_{\underline{N}:\bar{N}} \right\|_{L^\infty(\mathcal{P}_{\mathcal{X}})} \\
&\lesssim \|\mathbf{I}_N\|_{\mathrm{op}} \cdot \left\| \sum_{j=\underline{N}}^{\bar{N}} (\psi_j^\circ)^2 \right\|_{L^\infty(\mathcal{P}_{\mathcal{X}})} \\
&\lesssim N^{2r},
\end{aligned}$$

again by Assumption 1.

Hence we may apply Corollary A.3 with $v(\mathbf{Z}) \lesssim nN^{2r}$, $L \lesssim N^{2r}$ to conclude:

$$\mathbb{E}_{\mathbf{X}} \left[\left\| \Sigma_{\Psi,N}^{-1/2} \frac{\Psi^\circ_{\underline{N}:\bar{N}} \Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n} \Sigma_{\Psi,N}^{-1/2} - \mathbf{I}_N \right\|_{\mathrm{op}}^2 \right] = \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{S}_i \right\|_{\mathrm{op}}^2 \right] \lesssim \frac{N^{2r}}{n} \log N + \frac{N^{4r}}{n^2} \log^2 N.$$

Proof of Corollary A.3. From Theorem A.2 and with the change of variables $\lambda = t^2/n^2$, we have

$$\Pr\left(\frac{1}{n^2} \|\mathbf{Z}\|_{\mathrm{op}}^2 \geq \lambda \right) \leq 2N \exp\left(-\frac{n^2\lambda}{2v(\mathbf{Z}) + \frac{2}{3} Ln\sqrt{\lambda}} \right).$$

Since the probability is also bounded above by 1, it follows that

$$\begin{aligned}
\Pr\left(\frac{1}{n^2} \|\mathbf{Z}\|_{\mathrm{op}}^2 \geq \lambda \right) &\leq 1 \wedge 2N \exp\left(-\frac{n^2\lambda}{2(2v(\mathbf{Z}) \vee \frac{2}{3} Ln\sqrt{\lambda})} \right) \\
&\leq 1 \wedge 2N \exp\left(-\frac{n^2\lambda}{4v(\mathbf{Z})} \right) + 1 \wedge 2N \exp\left(-\frac{3n\sqrt{\lambda}}{4L} \right),
\end{aligned}$$

and the expectation can be controlled as

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n^2} \|\mathbf{Z}\|_{\mathrm{op}}^2 \right] &= \int_0^\infty \Pr\left(\frac{1}{n^2} \|\mathbf{Z}\|_{\mathrm{op}}^2 \geq \lambda \right) \mathrm{d}\lambda \\
&\leq \int_0^\infty 1 \wedge 2N \exp\left(-\frac{n^2\lambda}{4v(\mathbf{Z})} \right) \mathrm{d}\lambda + \int_0^\infty 1 \wedge 2N \exp\left(-\frac{3n\sqrt{\lambda}}{4L} \right) \mathrm{d}\lambda.
\end{aligned}$$

For the first integral, we truncate at $\lambda_1 := \frac{4v(\mathbf{Z})}{n^2} \log 2N$ so that

$$\begin{aligned}
\int_0^\infty 1 \wedge 2N \exp\left(-\frac{n^2\lambda}{4v(\mathbf{Z})} \right) \mathrm{d}\lambda &= \lambda_1 + \int_{\lambda_1}^\infty 2N \exp\left(-\frac{n^2\lambda}{4v(\mathbf{Z})} \right) \mathrm{d}\lambda \\
&= \frac{4v(\mathbf{Z})}{n^2} \log 2N + \frac{4v(\mathbf{Z})}{n^2}.
\end{aligned}$$

For the second integral, we truncate at $\lambda_2 := \left(\frac{4L}{3n} \log 2N \right)^2$ so that

$$\begin{aligned}
&\int_0^\infty 1 \wedge 2N \exp\left(-\frac{3n\sqrt{\lambda}}{4L} \right) \mathrm{d}\lambda \\
&= \lambda_2 + \int_{\lambda_2}^\infty 2N \exp\left(-\frac{3n\sqrt{\lambda}}{4L} \right) \mathrm{d}\lambda \\
&= \lambda_2 - \frac{16LN}{3n} \left(\sqrt{\lambda} + \frac{4L}{3n} \right) \exp\left(-\frac{3n\sqrt{\lambda}}{4L} \right) \Bigg|_{\lambda=\lambda_2}^\infty \\
&= \left(\frac{4L}{3n} \log 2N \right)^2 + \frac{8L}{3n} \left(\frac{4L}{3n} \log 2N + \frac{4L}{3n} \right).
\end{aligned}$$

Adding up, we conclude that

$$\mathbb{E}\left[\frac{1}{n^2} \|\mathbf{Z}\|_{\mathrm{op}}^2 \right] \leq \frac{4v(\mathbf{Z})}{n^2} (1 + \log 2N) + \frac{16L^2}{9n^2} (2 + 2\log 2N + (\log 2N)^2)$$

as was to be shown. □

B Proofs on Metric Entropy

B.1 Modified Proof of Theorem 3.1

For a full proof of the original statement, we refer the reader to Section B.1 of Hayakawa and Suzuki (2020), which corrects some technical flaws in the original proof. Here we only outline the necessary modification to incorporate the bounded noise setting.

Denote an ϵ-cover of the model space by $f_1, \cdots, f_M$. The only step which relies on the normality of noise $\xi_{1:n}$ is a concentration result for the random variables

$$\varepsilon_j := \frac{\sum_{i=1}^n \xi_i(f_j(x_i) - f^\circ(x_i))}{\left[(\sum_{i=1}^n (f_j(x_i) - f^\circ(x_i))^2\right]^{1/2}}, \quad 1 \leq j \leq M,$$

where it is shown via the normality of ε_j that

$$\mathbb{E}\left[\max_{1\leq j\leq M} \varepsilon_j^2\right] \leq 4\sigma^2(\log M + 1).$$

We will instead rely on Hoeffding's inequality. By writing $\varepsilon_j = w_j^\top \xi_{1:n} = \sum_{i=1}^n w_{j,i}\xi_i$ where

$$w_{j,i} = \frac{f_j(x_i) - f^\circ(x_i)}{\left[(\sum_{i=1}^n (f_j(x_i) - f^\circ(x_i))^2\right]^{1/2}},$$

since $|w_{j,i}\xi_i| \leq \sigma|w_{j,i}|$ a.s. it follows that

$$\Pr\left(|\varepsilon_j| \geq u\right) \leq 2\exp\left(-\frac{2u^2}{\sum_{i=1}^n (2\sigma|w_{j,i}|)^2}\right) = 2\exp\left(-\frac{u^2}{2\sigma^2}\right).$$

for all $u > 0$. Then the squared-exponential moment of each ε_j is bounded as

$$\begin{aligned}
\mathbb{E}\left[\exp(t\varepsilon_j^2)\right] &= 1 + \int_1^\infty \Pr\left(\exp(t\varepsilon_j^2) \geq \lambda\right) \mathrm{d}\lambda \\
&\leq 1 + \int_1^\infty 2\exp\left(-\frac{\log\lambda}{2\sigma^2 t}\right) \mathrm{d}\lambda \\
&\leq 1 + 2\int_1^\infty \lambda^{-\frac{1}{2\sigma^2 t}} \mathrm{d}\lambda \leq 3
\end{aligned}$$

by setting $t = \frac{1}{4\sigma^2}$. Hence via Jensen's inequality we obtain

$$\exp\left(t\mathbb{E}\left[\max_{1\leq j\leq M} \varepsilon_j^2\right]\right) \leq \mathbb{E}\left[\max_{1\leq j\leq M} \exp(t\varepsilon_j^2)\right] \leq \sum_{j=1}^M \mathbb{E}\left[\exp(t\varepsilon_j^2)\right] \leq 3M$$

and thus

$$\mathbb{E}\left[\max_{1\leq j\leq M} \varepsilon_j^2\right] \leq 4\sigma^2 \log 3M,$$

retrieving the original result up to a constant. □

B.2 Proof of Lemma 3.3

Let us take two functions $f_{\Theta_1}, f_{\Theta_2} \in \mathcal{T}_N$ for $\Theta_i = (\Gamma_i, \phi_i)$, $i = 1, 2$ separated as

$$\|\Gamma_1 - \Gamma_2\|_{\text{op}} \leq \delta_1, \quad \max_{1\leq j\leq N} \|\phi_{1,j} - \phi_{2,j}\|_{L^\infty(\mathcal{P}_\mathcal{X})} \leq \delta_2.$$

Then it holds that

$$\begin{aligned}
&|f_{\Theta_1}(\mathbf{X}, \boldsymbol{y}, \tilde{x}) - f_{\Theta_2}(\mathbf{X}, \boldsymbol{y}, \tilde{x})| \\
&\leq |\check{f}_{\Theta_1}(\mathbf{X}, \boldsymbol{y}, \tilde{x}) - \check{f}_{\Theta_2}(\mathbf{X}, \boldsymbol{y}, \tilde{x})| \\
&= \left|\left\langle \frac{\Gamma_1\phi_1(\mathbf{X})\boldsymbol{y}}{n}, \phi_1(\tilde{x})\right\rangle - \left\langle \frac{\Gamma_2\phi_2(\mathbf{X})\boldsymbol{y}}{n}, \phi_2(\tilde{x})\right\rangle\right|
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\Big|\phi_1(\tilde{x})^\top \Gamma_1 \phi_1(\mathbf{X})\boldsymbol{y} - \phi_2(\tilde{x})^\top \Gamma_1 \phi_1(\mathbf{X})\boldsymbol{y} + \phi_2(\tilde{x})^\top \Gamma_1 \phi_1(\mathbf{X})\boldsymbol{y} \\
&\qquad - \phi_2(\tilde{x})^\top \Gamma_2 \phi_1(\mathbf{X})\boldsymbol{y} + \phi_2(\tilde{x})^\top \Gamma_2 \phi_1(\mathbf{X})\boldsymbol{y} - \phi_2(\tilde{x})^\top \Gamma_2 \phi_2(\mathbf{X})\boldsymbol{y}\Big| \\
&\le \frac{1}{n}\|\phi_1(\tilde{x}) - \phi_2(\tilde{x})\|\|\Gamma_1\|_{\mathrm{op}}\|\phi_1(\mathbf{X})\boldsymbol{y}\| + \frac{1}{n}\|\phi_2(\tilde{x})\|\|\Gamma_1 - \Gamma_2\|_{\mathrm{op}}\|\phi_1(\mathbf{X})\boldsymbol{y}\| \\
&\qquad + \frac{1}{n}\|\phi_2(\tilde{x})\|\|\Gamma_2\|_{\mathrm{op}}\|(\phi_1(\mathbf{X}) - \phi_2(\mathbf{X}))\boldsymbol{y}\| \\
&\le \left(\frac{\sqrt{N}\delta_2 C_2}{n} + \frac{B_N'\delta_1}{n}\right)\sum_{i=1}^n \|\phi_1(x_i)\||y_i| + \frac{B_N' C_2}{n}\sum_{i=1}^n \|\phi_1(x_i) - \phi_2(x_i)\||y_i| \\
&\le (\sqrt{N}\delta_2 C_2 + B_N'\delta_1)B_N'(B+\sigma) + B_N' C_2\sqrt{N}\delta_2(B+\sigma) \\
&= B_N'^2(B+\sigma)\delta_1 + 2B_N'(B+\sigma)C_2\sqrt{N}\delta_2.
\end{aligned}
$$

Hence to construct an ϵ-cover $G_{\mathcal{T}}$ of $\mathcal{T}_N$, it suffices to exhibit a δ_1-cover $G_{\mathcal{S}}$ of $\mathcal{S}_N$ and a δ_2-cover $G_{\mathcal{F}}$ of $\mathcal{F}_N$ for

$$
\delta_1 = \frac{\epsilon}{2B_N'^2(B+\sigma)}, \qquad \delta_2 = \frac{\epsilon}{4B_N'(B+\sigma)C_2\sqrt{N}}
$$

and set $G_{\mathcal{T}} = G_{\mathcal{S}} \times G_{\mathcal{F}}$. For the metric entropy, this implies

$$
\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon) \le \mathcal{V}(\mathcal{S}_N, \|\cdot\|_{\mathrm{op}}, \delta_1) + \mathcal{V}(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \delta_2).
$$

We further bound the metric entropy of $\mathcal{S}_N$ with the following result, proved in Appendix B.3.

Lemma B.1. *For $\delta \le \frac{1}{2}$ it holds that $\mathcal{V}(\mathcal{S}_N, \|\cdot\|_{\mathrm{op}}, \delta) \lesssim N^2 \log \frac{1}{\delta}$.*

Substituting into the above, we conclude that

$$
\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon) \lesssim N^2 \log \frac{B_N'^2}{\epsilon} + \mathcal{V}\left(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \frac{\epsilon}{4B_N'(B+\sigma)C_2\sqrt{N}}\right).
$$

The choice of δ_2 is not important as long as the metric entropy of $\mathcal{F}_N$ is at most polynomial in δ_2. □

B.3 Proof of Lemma B.1

Let $\Gamma_1, \Gamma_2 \in \mathcal{S}_N$ and consider their diagonalizations

$$
\Gamma_i = \mathbf{U}_i \Lambda_i \mathbf{U}_i^\top, \quad \Lambda_i = \mathrm{diag}(\lambda_{i,1}, \cdots, \lambda_{i,N}), \quad i = 0, 1,
$$

where $\mathbf{U}_i \in \mathcal{O}_N$, the orthogonal group in dimension N, and $0 \le \lambda_{i,j} \le C_3$ for each $1 \le j \le N$. Assuming

$$
\|\mathbf{U}_1 - \mathbf{U}_2\|_{\mathrm{op}} \le \frac{\delta}{4C_3}, \quad |\lambda_{1,j} - \lambda_{2,j}| \le \frac{\delta}{2} \quad \forall j \le N,
$$

it follows that

$$
\begin{aligned}
&\|\Gamma_1 - \Gamma_2\|_{\mathrm{op}} \\
&= \|\mathbf{U}_1\Lambda_1\mathbf{U}_1^\top - \mathbf{U}_1\Lambda_1\mathbf{U}_2^\top + \mathbf{U}_1\Lambda_1\mathbf{U}_2^\top - \mathbf{U}_2\Lambda_1\mathbf{U}_2^\top + \mathbf{U}_2\Lambda_1\mathbf{U}_2^\top - \mathbf{U}_2\Lambda_2\mathbf{U}_2^\top\|_{\mathrm{op}} \\
&\le 2\|\Lambda_1\|_{L^\infty}\|\mathbf{U}_1 - \mathbf{U}_2\|_{\mathrm{op}} + \|\Lambda_1 - \Lambda_2\|_{L^\infty} \\
&\le 2C_3 \cdot \frac{\delta}{4C_3} + \frac{\delta}{2} = \delta.
\end{aligned}
$$

Moreover, the covering number of $\mathcal{O}_N$ in operator norm is given by the following result.

Theorem B.2 (Szarek (1981), Proposition 6)**.** *There exist universal constants $c_1, c_2 > 0$ such that for all $N \in \mathbb{N}$ and $\delta \in (0, 2]$,*

$$
\left(\frac{c_1}{\delta}\right)^{\frac{N(N-1)}{2}} \le \mathcal{N}(\mathcal{O}_N, \|\cdot\|_{\mathrm{op}}, \delta) \le \left(\frac{c_2}{\delta}\right)^{\frac{N(N-1)}{2}}.
$$

Hence we obtain that

$$\begin{aligned}\mathcal{V}(\mathcal{S}_N, \|\cdot\|_{\mathrm{op}}, \delta) &\leq \mathcal{V}\left(\mathcal{O}_N, \|\cdot\|_{\mathrm{op}}, \frac{\delta}{4C_3}\right) + \mathcal{V}\left([0, C_3]^N, \|\cdot\|_{L^\infty}, \frac{\delta}{2}\right) \\ &\leq \frac{N(N-1)}{2} \log \frac{c_2}{\delta} + N \log \frac{2C_3}{\delta} \\ &\lesssim N^2 \log \frac{1}{\delta}.\end{aligned}$$

Finally, we remark that if elements of the domain $\mathcal{S}_N$ are not constrained to be symmetric, we can alternatively consider the singular value decomposition and separately bound entropy of the two rotation components, giving the same result up to constants. □

C Details on Besov-type Spaces

C.1 Besov Space

C.1.1 Verification of Assumptions

We first give some background on wavelet decomposition. The decay rate $s = \alpha/d$ is intrinsic to the Besov space as shown by the following result, which allows us to translate between functions $f \in B^\alpha_{p,q}(\mathcal{X})$ and their B-spline coefficient sequences.

Lemma C.1 (DeVore and Popov (1988), Corollary 5.3)**.** *If $\alpha > d/p$ and $m > \alpha + 1 - 1/p$, a function $f \in L^p(\mathcal{X})$ is in $B^\alpha_{p,q}(\mathcal{X})$ if and only if f can be represented as*

$$f = \sum_{k=0}^\infty \sum_{\ell \in I_k^d} \tilde{\beta}_{k,\ell} \omega_{k,\ell}^d$$

such that the coefficients satisfy

$$\|\tilde{\beta}\|_{b^\alpha_{p,q}} := \left[\sum_{k=0}^\infty \left[2^{k(\alpha - d/p)} \left(\sum_{\ell \in I_k^d} |\tilde{\beta}_{k,\ell}|^p\right)^{1/p}\right]^q\right]^{1/q} < \infty,$$

with appropriate modifications if $p = \infty$ or $q = \infty$. Moreover, the two norms $\|\tilde{\beta}\|_{b^\alpha_{p,q}}$ and $\|f\|_{B^\alpha_{p,q}}$ are equivalent.

In particular, this implies that for any $f \in \mathbb{U}(B^\alpha_{p,q}(\mathcal{X}))$ the p-norm average of $\tilde{\beta}_{k,\ell}$ for $\ell \in I_k^d$ at resolution k is bounded as

$$\left(\frac{1}{|I_k^d|} \sum_{\ell \in I_k^d} |\tilde{\beta}_{k,\ell}|^p\right)^{1/p} \lesssim (2^{-kd})^{1/p} \cdot 2^{k(d/p - \alpha)} \|f\|_{B^\alpha_{p,q}} \leq 2^{-k\alpha},$$

and the coefficients $\beta_{k,\ell} = 2^{-kd/2} \tilde{\beta}_{k,\ell}$ w.r.t. the scaled basis $(2^{kd/2}\omega_{k,\ell}^d)_{k,\ell}$ satisfy

$$\left(\frac{1}{|I_k^d|} \sum_{\ell \in I_k^d} |\beta_{k,\ell}|^p\right)^{1/p} \lesssim (2^{kd})^{-\alpha/d - 1/2}. \tag{18}$$

Thus it is natural in a probabilistic sense to assume $\mathbb{E}[\beta_{k,\ell}^p]^{1/p} \lesssim (2^{kd})^{-\alpha/d-1/2}$. This will be the case if for instance we sample $(\beta_{k,\ell})_{\ell \in I_k^d}$ uniformly from the p-norm ball (18). This matches Assumption 4 up to the logarithmic factor and hence the rate is nearly tight in variance, even though (18) only applies to the average over locations rather than each coefficient explicitly. See also the discussion in Lemma 2 of Suzuki (2019).

Assumption 1. We take m to be even for simplicity. The wavelet system $(\omega^d_{K,\ell})_{\ell\in I^d_K}$ at each resolution K is linearly independent; for any $g\in L^2(\mathcal{X})$ that can be expressed as

$$g=\sum_{\ell\in I^d_K}\beta_{K,\ell}2^{Kd/2}\omega^d_{K,\ell}$$

we have the quasi-norm equivalence (Dũng, 2011b, 2.15)

$$\|g\|_2\asymp 2^{-Kd/2}\Bigg(\sum_{\ell\in I^d_K}2^{Kd}\beta^2_{K,\ell}\Bigg)^{1/2}=\|(\beta_{K,\ell})_{\ell\in I^d_K}\|_2$$

which implies that the covariance matrix $\mathbb{E}_{x\sim\mathrm{Unif}([0,1]^d)}[\psi^\circ_{\underline{N}:\bar{N}}(x)\psi^\circ_{\underline{N}:\bar{N}}(x)^\top]$ is bounded above and below. Since we assume $\mathcal{P}_{\mathcal{X}}$ has Lebesgue density bounded above and below, it follows that $\Sigma_{\Psi,N}$ is uniformly bounded above and below for all $K\geq 0$.

In contrast, any B-spline at a lower resolution $k<K$ can be exactly expressed as a linear combination of elements of $(\omega^d_{K,\ell})_{\ell\in I^d_K}$ by repeatedly applying the following relation.

Lemma C.2 (refinement equation)**.** *For even m and $r=(r_1,\cdots,r_d)^\top$, $\mathbf{1}=(1,\cdots,1)^\top\in\mathbb{R}^d$ it holds that*

$$\omega^d_{k,\ell}=\sum_{r_1,\cdots,r_d=0}^{m}2^{(-m+1)d}\prod_{i=1}^{d}\binom{m}{r_i}\cdot\omega^d_{k+1,2\ell+r-\frac{m}{2}\mathbf{1}}. \tag{19}$$

Proof. The relation for one-dimensional wavelets is given in equation (2.21) of Dũng (2011b),

$$\iota_m(x)=2^{-m+1}\sum_{r=0}^{m}\binom{m}{r}\iota_m\left(2x-r+\frac{m}{2}\right),$$

from which it follows that

$$\begin{aligned}\omega^d_{k,\ell}(x)&=\prod_{i=1}^{d}\iota_m(2^kx_i-\ell_i)\\&=\sum_{r_1,\cdots,r_d=0}^{m}2^{(-m+1)d}\prod_{i=1}^{d}\binom{m}{r_i}\iota_m\left(2^{k+1}x_i-2\ell_i-r_i+\frac{m}{2}\right)\\&=\sum_{r_1,\cdots,r_d=0}^{m}2^{(-m+1)d}\prod_{i=1}^{d}\binom{m}{r_i}\cdot\omega^d_{k+1,2\ell+r-\frac{m}{2}\mathbf{1}}(x)\end{aligned}$$

as was to be shown. □

Therefore we select all B-splines at a fixed resolution K to approximate the target tasks,

$$N=\left|I^d_K\right|=(m+1+2^K)^d\asymp 2^{Kd}$$

and

$$\underline{N}=\sum_{k=0}^{K-1}\left|I^d_k\right|+1\asymp N,\quad \bar{N}=\sum_{k=0}^{K}\left|I^d_k\right|\asymp N.$$

It is straightforward to see that $0\leq\omega^d_{k,\ell}(x)\leq 1$ for all $x\in\mathcal{X}$ and moreover the B-splines (extended to all $\ell\in\mathbb{Z}^d$) form a partition of unity of $\mathbb{R}^d$ at all resolutions:

$$\sum_{\ell\in\mathbb{Z}^d}\omega^d_{k,\ell}(x)\equiv 1,\quad\forall x\in\mathbb{R}^d,\quad\forall k\geq 0.$$

Then for all $x\in\mathcal{X}$ we have the bound

$$\sum_{j=\underline{N}}^{\bar{N}}\psi^\circ_j(x)^2=\sum_{\ell\in I^d_K}2^{Kd}\omega^d_{K,\ell}(x)^2\leq 2^{Kd}\sum_{\ell\in\mathbb{Z}^d}\omega^d_{K,\ell}(x)=2^{Kd}\lesssim N,$$

and hence (2) holds with $r=1/2$.

Assumption 2. For any $\beta \in \operatorname{supp} \mathcal{P}_\beta$ we have that

$$\begin{aligned}
\|F_\beta^\circ\|_{L^\infty(\mathcal{P}_\mathcal{X})} &\leq \sum_{k=0}^\infty \Bigg\| \sum_{\ell \in I_k^d} \beta_{k,\ell} \cdot 2^{kd/2} \omega_{k,\ell}^d \Bigg\|_{L^\infty(\mathcal{P}_\mathcal{X})} \\
&\leq \sum_{k=0}^\infty 2^{kd/2} \max_{\ell \in I_k^d} |\beta_{k,\ell}| \cdot \Bigg\| \sum_{\ell \in I_k^d} \omega_{k,\ell}^d \Bigg\|_{L^\infty(\mathcal{X})} \\
&\leq \sum_{k=0}^\infty 2^{kd(1/2+1/p)} \left(\frac{1}{|I_k^d|} \sum_{\ell \in I_k^d} |\beta_{k,\ell}|^p \right)^{1/p} \Bigg\| \sum_{\ell \in I_k^d} \omega_{k,\ell}^d \Bigg\|_{L^\infty(\mathcal{X})} \\
&\lesssim \sum_{k=0}^\infty 2^{kd(1/2+1/p)} \cdot (2^{kd})^{-\alpha/d-1/2} \|F_\beta^\circ\|_{B_{p,q}^\alpha} \\
&\lesssim (1-2^{d/p-\alpha})^{-1} =: B.
\end{aligned}$$

Furthermore, the convergence rate of the truncated approximation $F_{\beta,\bar{N}}^\circ$ is determined by the decay rate of β in Lemma C.1 as follows (it does not matter whether we bound $F_{\beta,\bar{N}}^\circ$ or $F_{\beta,N}^\circ$ since $\bar{N} \asymp N$). We consider a truncation of all resolutions lower than K so that $\bar{N}, N \asymp 2^{Kd}$. Then it holds that

$$\begin{aligned}
\|F_\beta^\circ - F_{\beta,\bar{N}}^\circ\|_{L^2(\mathcal{P}_\mathcal{X})}^2 &= \int_\mathcal{X} \left(\sum_{j=\bar{N}+1}^\infty \beta_j \psi_j^\circ(x) \right)^2 \mathcal{P}_\mathcal{X}(\mathrm{d}x) = \sum_{j,k=\bar{N}+1}^\infty \beta_j \beta_k \mathbb{E}_x[\psi_j^\circ(x) \psi_k^\circ(x)] \\
&\leq \lim_{M\to\infty} C_2 \|\beta_{\bar{N}:M}\|^2 = C_2 \sum_{k=K}^\infty \sum_{\ell \in I_k^d} \beta_{k,\ell}^2 \\
&\lesssim \sum_{k=K}^\infty |I_k^d|^{1-2/p} \left(\sum_{\ell \in I_k^d} |\beta_{k,\ell}|^p \right)^{2/p} \lesssim \sum_{k=K}^\infty 2^{kd} \cdot (2^{kd})^{-2\alpha/d-1} \\
&= \frac{2^{-2\alpha K}}{1-2^{-2\alpha}} \asymp N^{-2\alpha/d},
\end{aligned}$$

where for the last two inequalities we have used the inequality $\|z\|_2 \leq D^{1/2-1/p} \|z\|_p$ for $z \in \mathbb{R}^D$ and $p \geq 2$ in conjunction with (18). Thus our choice of $s = \alpha/d$ is justified. Under this choice, (5) directly implies (4) as

$$\mathbb{E}_\beta[\beta_{K,\ell}^2] \lesssim 2^{-Kd(2s+1)} K^{-2} \asymp \bar{N}^{-2s-1} (\log \bar{N})^{-2}$$

holds for the basis elements at each resolution K, that is for those numbered between $\underline{N}$ and $\bar{N}$.

Remark C.3. When $1 \leq p < 2$, the truncation up to $\bar{N}$ does not suffice to achieve the $N^{-2\alpha/d}$ approximation rate, and basis elements must be judiciously selected from a wider resolution range. More concretely, a size N subset of all wavelets up to resolution $K' = \lceil K(1+\nu^{-1}) \rceil$ where $\nu = p\alpha/2d - 1/2 > 0$ must be used (Suzuki and Nitanda, 2021, Lemma 2). Hence the exponent is a factor of $1+\nu^{-1}$ worse w.r.t. $N' \asymp 2^{K'd}$, leading to the inevitable suboptimal rate.

To show boundedness of $\mathrm{Tr}(\Sigma_{\bar{\beta},N})$, we analyze the composition of the aggregated coefficients $\bar{\beta}$ using the following result.

Corollary C.4. *For any $0 \leq k < k'$ there exists constants $\gamma_{k,k',\ell,\ell'} \geq 0$ for $\ell \in I_k^d$, $\ell' \in I_{k'}^d$ such that*

$$\sum_{\ell \in I_k^d} \beta_{k,\ell} 2^{kd/2} \omega_{k,\ell}^d = \sum_{\ell' \in I_{k'}^d} \bar{\beta}_{k',\ell'} 2^{k'd/2} \omega_{k',\ell'}^d, \quad \bar{\beta}_{k',\ell'} = \sum_{\ell \in I_k^d} \gamma_{k,k',\ell,\ell'} \beta_{k,\ell}$$

holds for all $(\beta_{k,\ell})_{\ell \in I_k^d}$. Moreover, it holds that

$$\sum_{\ell \in I_k^d} \gamma_{k,k',\ell,\ell'} \leq 2^{(k-k')d/2}, \quad \sum_{\ell' \in I_{k'}^d} \gamma_{k,k',\ell,\ell'} \leq 2^{(k'-k)d/2}.$$

The statement follows directly from the more general Proposition C.10, stated and proved in Appendix C.3 below, by restricting to wavelets with uniform resolution across dimensions. Using Corollary C.4, we can refine each lower resolution component of F_β° to resolution K:

$$F_{\beta,\bar{N}}^\circ = \sum_{j=1}^{\bar{N}} \beta_j \psi_j^\circ = \sum_{k=0}^{K} \sum_{\ell \in I_k^d} \beta_{k,\ell} 2^{kd/2} \omega_{k,\ell}^d = \sum_{k=0}^{K} \sum_{\ell' \in I_K^d} \sum_{\ell \in I_k^d} \gamma_{k,K,\ell,\ell'} \beta_{k,\ell} 2^{Kd/2} \omega_{K,\ell'}^d.$$

Thus each aggregated coefficient, indexed here by $\ell \in I_K^d$, can be expressed as

$$\bar{\beta}_{K,\ell} = \sum_{k=0}^{K} \sum_{\ell \in I_k^d} \gamma_{k,K,\ell,\ell'} \beta_{k,\ell}.$$

Hence it follows that

$$\begin{aligned}
\mathbb{E}_\beta[\bar{\beta}_{K,\ell}^2] = \sum_{k=0}^{K} \sum_{\ell \in I_k^d} \gamma_{k,K,\ell,\ell'}^2 \mathbb{E}_\beta[\beta_{k,\ell}^2] &\lesssim \sum_{k=0}^{K} \left(\sum_{\ell \in I_k^d} \gamma_{k,K,\ell,\ell'} \right)^2 2^{-k(2\alpha+d)} k^{-2} \\
&\leq \sum_{k=0}^{K} 2^{(k-K)d} \cdot 2^{-k(2\alpha+d)} k^{-2} \\
&\lesssim 2^{-Kd} \cdot \sum_{k=0}^{K} 2^{-2k\alpha} k^{-2} \asymp N^{-1},
\end{aligned}$$

from which we conclude that $\operatorname{Tr}(\Sigma_{\bar{\beta},N}) = \sum_{\ell \in I_K^d} \mathbb{E}_\beta[\bar{\beta}_{K,\ell}^2] \lesssim N \cdot N^{-1}$ is uniformly bounded.

Finally for the verification of Assumption 3, see Appendix C.1.2.

C.1.2 Proof of Lemma 4.4

We use the following result to approximate each wavelet $\omega_{K,\ell}^d$ at resolution K with the class (6). The proof, in turn, relies on the construction by Yarotsky (2016) of DNNs which efficiently approximates the multiplication operation.

Lemma C.5 (Suzuki (2019), Lemma 1)**.** *For all $\delta > 0$, there exists a ReLU neural network $\tilde{\omega} \in \mathcal{F}_{\mathrm{DNN}}(L, W, S, M)$ with*

$$\begin{aligned}
&L = 3 + 2 \left\lceil \log_2 \left(3^{d \vee m} (1 + dm^{-1/2} (2e)^{m+1}) \delta^{-1} \right) + 5 \right\rceil \lceil \log_2 (d \vee m) \rceil, \\
&W = W_0 = 6dm(m+2) + 2d, \qquad S = LW^2, \qquad M = 2(m+1)^m
\end{aligned}$$

satisfying $\operatorname{supp} \tilde{\omega} \subseteq [0, m+1]^d$ *and* $\|\omega_{0,0}^d - \tilde{\omega}\|_{L^\infty(\mathcal{X})} \leq \delta$.

Here, δ is also dependent on N.

Now consider N identical copies of $\tilde{\omega}$ in parallel, where each module is preceded by the scaling $(x_i)_{i=1}^d \mapsto (2^K x_i - \ell_i)_{i=1}^d$ for $\ell \in I_K^d$ and whose output is scaled by $2^{Kd/2}$. In particular, these operations can be implemented by $K \lesssim \log N$ consecutive additional layers with norm bounded by a constant. Hence each module $\phi_{\underline{N}}^*, \cdots, \phi_{\bar{N}}^*$ approximates the basis $2^{Kd/2} \omega_{K,\ell}^d$ with $2^{Kd/2} \delta \lesssim \sqrt{N} \delta$ accuracy, and substituting $\delta_N = \sqrt{N} \delta$ gives that

$$\|\psi_j^\circ - \phi_j^*\|_{L^\infty(\mathcal{P}_\mathcal{X})} \leq \delta_N, \quad \underline{N} \leq j \leq \bar{N},$$

with $L \lesssim \log \delta^{-1} + \log N \lesssim \log \delta_N^{-1} + \log N$. Note that the sparsity S is only multiplied by a factor of N since different modules do not share any connections. Moreover the target basis has 2-norm bounded as

$$\|\psi_{\underline{N}:\bar{N}}^\circ(x)\|_2 \lesssim \left(\sum_{\ell \in I_K^d} 2^{Kd} \omega_{K,\ell}(x)^2 \right)^{1/2} \leq \left(2^{Kd} \sum_{j \in \mathbb{Z}^d} \omega_{K,\ell}(x) \right)^{1/2} \asymp \sqrt{N},$$

where we have again used the sparsity of $\omega_{k,\ell}^d$ at each resolution. Hence we may clip the magnitude of the vector output ϕ by B_N' and the approximation guarantee remains unchanged.

To bound the covering number of $\mathcal{F}_N$, we directly apply the following result.

Lemma C.6 (Suzuki (2019), Lemma 3)**.** *The covering number of $\mathcal{F}_{\rm DNN}$ is bounded as*

$$\mathcal{N}(\mathcal{F}_{\rm DNN}(L,W,S,M),\|\cdot\|_{L^\infty},\epsilon)\leq\left(\frac{L(M\vee 1)^{L-1}(W+1)^{2L}}{\epsilon}\right)^S.$$

Since clipping the magnitude of the outputs does not increase the covering number, we conclude:

$$\begin{aligned}\mathcal{V}(\mathcal{F}_N,\|\cdot\|_{L^\infty},\epsilon)&\leq N\cdot\mathcal{V}(\mathcal{F}_{\rm DNN}(L,W,S,M),\|\cdot\|_{L^\infty},\epsilon)\\&\leq SN\log L+SLN\log M+2SLN\log(W+1)+SN\log\frac{1}{\epsilon}\\&\lesssim N\log\frac{N}{\delta_N}+N\log\frac{1}{\epsilon}.\end{aligned}$$

□

C.1.3 Proof of Theorem 4.5

By Lemma 3.3 and Lemma 4.4, the metric entropy of $\mathcal{T}_N$ is bounded as

$$\mathcal{V}(\mathcal{T}_N,\|\cdot\|_{L^\infty},\epsilon)\lesssim N^2\log\frac{N}{\epsilon}+N\log\frac{N^2}{\delta_N\epsilon}.$$

Combining with Theorem 3.1 and Proposition 3.2 with $r=1/2$ and $s=\alpha/d$ gives

$$\bar{R}(\widehat{\Theta})\lesssim\frac{N}{n}\log N+\frac{N^2}{n^2}\log^2 N+N^{-2\alpha/d}+N^2\delta_N^2+\frac{1}{T}\left(N^2\log\frac{N}{\epsilon}+N\log\frac{N^2}{\delta_N\epsilon}\right)+\epsilon.$$

Substituting $\delta_N\asymp N^{-1-\alpha/d}$ and $\epsilon\asymp N^{-2\alpha/d}$ yields the desired bound. □

C.2 Anisotropic Besov Space

C.2.1 Definitions and Results

For $1\leq p,q\leq\infty$, directional smoothness $\alpha=(\alpha_1,\cdots,\alpha_d)\in\mathbb{R}^d_{>0}$ and $r=\max_{i\leq d}\lfloor\alpha_i\rfloor+1$, we define $\|\cdot\|_{B^\alpha_{p,q}}=\|\cdot\|_{L^p}+|\cdot|_{B^\alpha_{p,q}}$ where

$$|f|_{B^\alpha_{p,q}}:=\begin{cases}\left(\sum_{k=0}^\infty\left[2^k w_{r,p}(f,(2^{-k/\alpha_1},\cdots,2^{-k/\alpha_d}))\right]^q\right)^{1/q} & q<\infty\\ \sup_{k\geq 0}2^k w_{r,p}(f,(2^{-k/\alpha_1},\cdots,2^{-k/\alpha_d})) & q=\infty.\end{cases}$$

The anisotropic Besov space is defined as

$$B^\alpha_{p,q}(\mathcal{X})=\{f\in L^p(\mathcal{X})\mid\|f\|_{B^\alpha_{p,q}}<\infty\}.$$

The definition reduces to the usual Besov space if $\alpha_1=\cdots=\alpha_d$; see Vybíral (2006); Triebel (2011) for details. We also write $\overline{\alpha}=\max_i\alpha_i$, $\underline{\alpha}=\min_i\alpha_i$ and the harmonic mean smoothness as

$$\widetilde{\alpha}:=\Big(\sum_{i=1}^d\alpha_i^{-1}\Big)^{-1}.$$

For the anisotropic Besov space, we need to redefine the wavelet basis so that the sensitivity to resolution $k\in\mathbb{Z}_{\geq 0}$ differs for each component depending on α. Define the quantities

$$\|k\|_{\underline{\alpha}/\alpha}:=\sum_{i=1}^d\lfloor k\underline{\alpha}/\alpha_i\rfloor,\quad I_k^{d,\alpha}:=\prod_{i=1}^d\{-m,-m+1,\cdots,2^{\lfloor k\underline{\alpha}/\alpha_i\rfloor}\}\subset\mathbb{Z}^d.$$

We then set for each $k\geq 0$ and $\ell\in I_k^{d,\alpha}$

$$\omega^{d,\alpha}_{k,\ell}(x):=\omega^d_{(\lfloor k\underline{\alpha}/\alpha_1\rfloor,\cdots,\lfloor k\underline{\alpha}/\alpha_d\rfloor),\ell}(x)=\prod_{i=1}^d\iota_m(2^{\lfloor k\underline{\alpha}/\alpha_i\rfloor}x_i-\ell_i),$$

and take the scaled basis

$$\{\psi_j^\circ \mid j\in\mathbb{N}\} = \{2^{\|k\|_{\underline{\alpha}/\alpha}/2}\omega^{d,\alpha}_{k,\ell} \mid k\in\mathbb{Z}_{\geq 0}, \ell\in I^{d,\alpha}_k\}$$

with the natural hierarchy induced by k.

The minimax optimal rate for the anisotropic Besov space is equal to $n^{-\frac{2\widetilde{\alpha}}{2\widetilde{\alpha}+1}}$ (Suzuki and Nitanda, 2021, Theorem 4). Our result for in-context learning is as follows.

Theorem C.7 (minimax optimality of ICL in anisotropic Besov space)**.** *Let $\alpha\in\mathbb{R}^d_{>0}$ with $\widetilde{\alpha}>1/p$ and $\mathcal{F}^\circ = \mathbb{U}(B^\alpha_{p,q}(\mathcal{X}))$. Suppose that $\mathcal{P}_\mathcal{X}$ has positive Lebesgue density $\rho_\mathcal{X}$ bounded above and below on $\mathcal{X}$. Also suppose that all coefficients are independent and*

$$\mathbb{E}_\beta[\beta_{k,\ell}] = 0, \quad \mathbb{E}_\beta[\beta^2_{k,\ell}] \lesssim 2^{-k\underline{\alpha}(2+1/\widetilde{\alpha})}k^{-2}, \quad \forall k\geq 0,\ \ell\in I^{d,\alpha}_k. \tag{20}$$

Then for $n\gtrsim N\log N$ we have

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-2\widetilde{\alpha}} + \frac{N\log N}{n} + \frac{N^2\log N}{T}.$$

Hence if $T\gtrsim nN$ and $N\asymp n^{\frac{1}{2\widetilde{\alpha}+1}}$, in-context learning achieves the rate $n^{-\frac{2\widetilde{\alpha}}{2\widetilde{\alpha}+1}}\log n$ which is minimax optimal up to a polylog factor.

C.2.2 Proof of Theorem C.7

The overall approach is similar to Appendix C.1. The decay rate of functions in the anisotropic Besov space is characterized by the following result which extends Lemma C.1.

Lemma C.8 (Suzuki and Nitanda (2021), Lemma 2)**.** *If $\widetilde{\alpha}>1/p$ and $m>\overline{\alpha}+1-1/p$, a function $f\in L^p(\mathcal{X})$ is in $MB^\alpha_{p,q}(\mathcal{X})$ if and only if f can be represented as*

$$f = \sum_{k\in\mathbb{Z}^d_{\geq 0}}\sum_{\ell\in I^{d,\alpha}_k}\tilde{\beta}_{k,\ell}\omega^{d,\alpha}_{k,\ell}(x)$$

such that the coefficients satisfy

$$\|\tilde{\beta}\|_{b^\alpha_{p,q}} := \left[\sum_{k=0}^\infty\left[2^{k\underline{\alpha}-\|k\|_{\underline{\alpha}/\alpha}/p}\left(\sum_{\ell\in I^{d,\alpha}_k}|\tilde{\beta}_{k,\ell}|^p\right)^{1/p}\right]^q\right]^{1/q} \lesssim \|f\|_{B^\alpha_{p,q}}.$$

Moreover, the two norms $\|\tilde{\beta}\|_{b^\alpha_{p,q}}$ and $\|f\|_{B^\alpha_{p,q}}$ are equivalent.

We again select all B-splines $(\omega^{d,\alpha}_{K,\ell})_{\ell\in I^d_K}$ at each resolution K to approximate the target functions. By repeatedly applying the refinement equation for one-dimensional wavelets as many times as needed to each dimension separately, we may express any B-spline at a lower resolution $k<K$ as a linear combination of $(\omega^{d,\alpha}_{K,\ell})_{\ell\in I^d_K}$ similarly to Lemma C.2. See Proposition C.10 for details. We thus have

$$N = |I^{d,\alpha}_K| = \prod_{i=1}^d(m+1+2^{\lfloor K\underline{\alpha}/\alpha_i\rfloor}) \asymp 2^{\|K\|_{\underline{\alpha}/\alpha}}.$$

Since $\|k\|_{\underline{\alpha}/\alpha} = k\underline{\alpha}/\widetilde{\alpha} + O_k(1)$ always holds, it also follows that

$$\bar{N} = \sum_{k=0}^K|I^{d,\alpha}_k| + 1 \asymp \sum_{k=0}^K 2^{\|k\|_{\underline{\alpha}/\alpha}} \asymp \sum_{k=0}^K(2^{\underline{\alpha}/\widetilde{\alpha}})^k \asymp 2^{K\underline{\alpha}/\widetilde{\alpha}} \asymp N$$

and similarly $\underline{N}\asymp N$. Therefore,

$$\sum_{j=\underline{N}}^{\bar{N}}\psi_j^\circ(x)^2 \leq 2^{\|K\|_{\underline{\alpha}/\alpha}}\sum_{\ell\in I^{d,\alpha}_k}\omega^{d,\alpha}_{K,\ell}(x)^2 \leq 2^{\|K\|_{\underline{\alpha}/\alpha}} \asymp N$$

and the scaled coefficients decay in average as

$$\begin{aligned}\left(\frac{1}{|I_k^{d,\alpha}|}\sum_{\ell\in I_k^{d,\alpha}}|\beta_{k,\ell}|^p\right)^{1/p} &\lesssim \left(\prod_{i=1}^{d} 2^{\lfloor k\underline{\alpha}/\alpha_i\rfloor}\right)^{-1/p} 2^{-\|k\|_{\underline{\alpha}/\alpha}/2}\left(\sum_{\ell\in I_k^{d,\alpha}}|\tilde{\beta}_{k,\ell}|^p\right)^{1/p}\\ &\lesssim 2^{-\|k\|_{\underline{\alpha}/\alpha}/2-k\underline{\alpha}}\|f\|_{B^{\alpha}_{p,q}}\\ &\asymp N^{-(\widetilde{\alpha}+1/2)}\|f\|_{B^{\alpha}_{p,q}}.\end{aligned}$$

For Assumption 2, we can check that

$$\begin{aligned}\|F^\circ_\beta\|_{L^\infty(\mathcal{P}_\mathcal{X})} &\le \sum_{k=0}^{\infty}\left\|\sum_{\ell\in I_k^{d,\alpha}}\beta_{k,\ell}\cdot 2^{\|k\|_{\underline{\alpha}/\alpha}/2}\omega^{d,\alpha}_{k,\ell}\right\|_{L^\infty(\mathcal{P}_\mathcal{X})}\\ &\le \sum_{k=0}^{\infty}2^{(1/2+1/p)\|k\|_{\underline{\alpha}/\alpha}}\left(\frac{1}{|I_k^{d,\alpha}|}\sum_{\ell\in I_k^{d,\alpha}}|\beta_{k,\ell}|^p\right)^{1/p}\\ &\lesssim \sum_{k=0}^{\infty}2^{(1/2+1/p)\|k\|_{\underline{\alpha}/\alpha}}\cdot 2^{-\|k\|_{\underline{\alpha}/\alpha}/2-k\underline{\alpha}}\\ &\lesssim \left(1-2^{\underline{\alpha}/\widetilde{\alpha}(1/p-\widetilde{\alpha})}\right)^{-1} =: B\end{aligned}$$

and for a resolution cutoff $K>0$, $\bar{N}\asymp 2^{\|K\|_{\underline{\alpha}/\alpha}}$ the truncation error satisfies

$$\begin{aligned}\|F^\circ_\beta - F^\circ_{\beta,\bar{N}}\|^2_{L^2(\mathcal{P}_\mathcal{X})} &\lesssim \sum_{k=K+1}^{\infty}\sum_{\ell\in I_k^{d,\alpha}}\beta^2_{k,\ell} \lesssim \sum_{k=K+1}^{\infty}|I_k^{d,\alpha}|^{1-2/p}\left(\sum_{\ell\in I_k^{d,\alpha}}|\beta_{k,\ell}|^p\right)^{2/p}\\ &\lesssim \sum_{k=K+1}^{\infty}2^{\|k\|_{\underline{\alpha}/\alpha}}\cdot 2^{-\|k\|_{\underline{\alpha}/\alpha}-2k\underline{\alpha}} \asymp 2^{-2K\underline{\alpha}} \asymp N^{-2\widetilde{\alpha}}.\end{aligned}$$

Thus we may set $r=1/2$, $s=\widetilde{\alpha}$ and take the variance decay rate (20) as

$$\mathbb{E}_\beta[\beta^2_{k,\ell}] \lesssim 2^{-\|k\|_{\underline{\alpha}/\alpha}(2\widetilde{\alpha}+1)}k^{-2} \asymp 2^{-k\underline{\alpha}(2+1/\widetilde{\alpha}))}k^{-2}.$$

For boundedness of $\mathrm{Tr}(\Sigma_{\bar{\beta},N})$, we use the following result which is also obtained from Proposition C.10 by considering resolution vectors $(\lfloor k\underline{\alpha}/\alpha_1\rfloor,\cdots,\lfloor k\underline{\alpha}/\alpha_d\rfloor)$ and $(\lfloor k'\underline{\alpha}/\alpha_1\rfloor,\cdots,\lfloor k'\underline{\alpha}/\alpha_d\rfloor)$.

Corollary C.9. *For any $0\le k<k'$ there exists constants $\gamma_{k,k',\ell,\ell'}\ge 0$ for $\ell\in I_k^{d,\alpha}$, $\ell'\in I_{k'}^{d,\alpha}$ such that*

$$\sum_{\ell\in I_k^{d,\alpha}}\beta_{k,\ell}2^{\|k\|_{\underline{\alpha}/\alpha}/2}\omega^{d,\alpha}_{k,\ell} = \sum_{\ell'\in I_{k'}^{d,\alpha}}\bar{\beta}_{k',\ell'}2^{\|k'\|_{\underline{\alpha}/\alpha}/2}\omega^{d,\alpha}_{k',\ell'},\quad \bar{\beta}_{k',\ell'} = \sum_{\ell\in I_k^{d,\alpha}}\gamma_{k,k',\ell,\ell'}\beta_{k,\ell}$$

holds for all $(\beta_{k,\ell})_{\ell\in I_k^{d,\alpha}}$. Moreover, it holds that

$$\sum_{\ell\in I_k^{d,\alpha}}\gamma_{k,k',\ell,\ell'} \le 2^{(\|k\|_{\underline{\alpha}/\alpha}-\|k'\|_{\underline{\alpha}/\alpha})/2},\qquad \sum_{\ell'\in I_{k'}^{d,\alpha}}\gamma_{k,k',\ell,\ell'} \le 2^{(\|k'\|_{\underline{\alpha}/\alpha}-\|k\|_{\underline{\alpha}/\alpha})/2}.$$

We apply Corollary C.9 to refine all components of $F^\circ_{\beta,\bar{N}}$ to resolution K:

$$F^\circ_{\beta,\bar{N}} = \sum_{k=0}^{K}\sum_{\ell\in I_k^{d,\alpha}}\beta_{k,\ell}2^{\|k\|_{\underline{\alpha}/\alpha}/2}\omega^{d,\alpha}_{k,\ell} = \sum_{k=0}^{K}\sum_{\ell'\in I_K^{d,\alpha}}\sum_{\ell\in I_k^{d,\alpha}}\gamma_{k,K,\ell,\ell'}\beta_{k,\ell}2^{\|K\|_{\underline{\alpha}/\alpha}/2}\omega^{d,\alpha}_{K,\ell'}.$$

Hence it follows that

$$\mathbb{E}_\beta[\bar{\beta}^2_{K,\ell}] = \sum_{k=0}^{K}\sum_{\ell\in I_k^{d,\alpha}}\gamma^2_{k,K,\ell,\ell'}\mathbb{E}_\beta[\beta^2_{k,\ell}] \lesssim \sum_{k=0}^{K}\left(\sum_{\ell\in I_k^{d,\alpha}}\gamma_{k,K,\ell,\ell'}\right)^2 2^{-k\underline{\alpha}(2+1/\widetilde{\alpha})}k^{-2}$$

$$\le \sum_{k=0}^{K} 2^{\|k\|_{\underline{\alpha}/\alpha} - \|K\|_{\underline{\alpha}/\alpha}} \cdot 2^{-k\underline{\alpha}(2+1/\widetilde{\alpha})} k^{-2}$$

$$\lesssim 2^{-\|K\|_{\underline{\alpha}/\alpha}} \cdot \sum_{k=0}^{K} 2^{-2k\underline{\alpha}} k^{-2} \asymp N^{-1},$$

and we again conclude that $\mathrm{Tr}(\Sigma_{\bar{\beta},N})$ is uniformly bounded.

The rest of the proof proceeds similarly to the ordinary Besov space. □

C.3 Wavelet Refinement

In this subsection, we present and prove an auxiliary result concerning the refinement of B-spline wavelets and the recurrence relations satisfied by their coefficient sequences.

Proposition C.10. *For any $k, k' \in \mathbb{Z}_{\ge 0}^d$ such that $k' - k \in \mathbb{Z}_{\ge 0}^d$ there exists constants $\gamma_{k,k',\ell,\ell'} \ge 0$ for $\ell \in I_k^d$, $\ell' \in I_{k'}^d$ such that*

$$\sum_{\ell \in I_k^d} \beta_{k,\ell} 2^{\|k\|_1/2} \omega_{k,\ell}^d = \sum_{\ell' \in I_{k'}^d} \bar{\beta}_{k',\ell'} 2^{\|k'\|_1/2} \omega_{k',\ell'}^d, \quad \bar{\beta}_{k',\ell'} = \sum_{\ell \in I_k^d} \gamma_{k,k',\ell,\ell'} \beta_{k,\ell} \tag{21}$$

holds for all $(\beta_{k,\ell})_{\ell \in I_k^d}$. Moreover, it holds that

$$\sum_{\ell \in I_k^d} \gamma_{k,k',\ell,\ell'} \le 2^{-\|k'-k\|_1/2}, \qquad \sum_{\ell' \in I_{k'}^d} \gamma_{k,k',\ell,\ell'} \le 2^{\|k'-k\|_1/2}.$$

Proof. We proceed by induction on $\|k' - k\|_1$. When $k' = k + e_j$ for some $1 \le j \le d$, we can refine each $\omega_{k,\ell}^d$ using equation (2.21) of Dũng (2011b) as

$$\begin{aligned} \omega_{k,\ell}^d(x) &= \prod_{i=1}^{d} \iota_m(2^{k_i} x_i - \ell_i) \\ &= 2^{-m+1} \prod_{i \ne j} \iota_m(2^{k_i} x_i - \ell_i) \sum_{r=0}^{m} \binom{m}{r} \iota_m\left(2^{k_j+1} x_j - 2\ell_j - r + \frac{m}{2}\right) \\ &= 2^{-m+1} \sum_{r=0}^{m} \binom{m}{r} \omega_{k+e_j,\ell+(\ell_j+r-\frac{m}{2})e_j}^d(x). \end{aligned}$$

Since $\ell + (\ell_j + r - \frac{m}{2})e_j$ matches a given location vector $\ell' \in I_{k+e_j}^d$ if and only if $\ell_i = \ell_i'$ $(i \ne j)$ and $\ell_j' = 2\ell_j + r - \frac{m}{2}$, comparing coefficients in (21) yields

$$\gamma_{k,k+e_j,\ell,\ell'} = 2^{-m+1/2} \mathbf{1}_{\{\ell_i = \ell_i' \ (i \ne j)\}} \binom{m}{\ell_j' - 2\ell_j + \frac{m}{2}}.$$

Here, $\mathbf{1}_A$ denotes the indicator function for condition A. It follows that $\gamma_{k,k+e_j,\ell,\ell'} \ge 0$ and

$$\sum_{\ell \in I_k^d} \gamma_{k,k+e_j,\ell,\ell'} \le \sum_{\ell_j \in \mathbb{Z}} 2^{-m+1/2} \binom{m}{\ell_j' - 2\ell_j + \frac{m}{2}} \le 2^{-1/2},$$

$$\sum_{\ell' \in I_{k+e_j}^d} \gamma_{k,k+e_j,\ell,\ell'} \le \sum_{\ell_j' \in \mathbb{Z}} 2^{-m+1/2} \binom{m}{\ell_j' - 2\ell_j + \frac{m}{2}} \le 2^{1/2},$$

by considering parities.

Now suppose the claim holds for a fixed difference $\|k' - k\|_1$. Applying the above derivation to further refine resolution k' to $k'' = k' + e_j$ for arbitrary j gives

$$\sum_{\ell \in I_k^d} \beta_{k,\ell} 2^{\|k\|_1/2} \omega_{k,\ell}^d = \sum_{\ell' \in I_{k'}^d} \bar{\beta}_{k',\ell'} 2^{\|k'\|_1/2} \omega_{k',\ell'}^d = \sum_{\ell'' \in I_{k'+1}^d} \bar{\bar{\beta}}_{k'+1,\ell''} 2^{(\|k'\|_1+1)/2} \omega_{k'+e_j,\ell''}^d$$

where

$$\bar{\bar{\beta}}_{k'+e_j,\ell''} = \sum_{\ell'\in I^d_{k'}} 2^{-m+1/2}\mathbf{1}_{\{\ell'_i=\ell''_i\ (i\neq j)\}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}}\bar{\beta}_{k',\ell'}$$
$$= \sum_{\ell\in I^d_k}\sum_{\ell'\in I^d_{k'}} 2^{-m+1/2}\mathbf{1}_{\{\ell'_i=\ell''_i\ (i\neq j)\}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}}\gamma_{k,k',\ell,\ell'}\beta_{k,\ell}.$$

Hence we obtain the recurrence relation

$$\gamma_{k,k'+e_j,\ell,\ell''} = \sum_{\ell'\in I^d_{k'}} 2^{-m+1/2}\mathbf{1}_{\{\ell'_i=\ell''_i\ (i\neq j)\}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}}\gamma_{k,k',\ell,\ell'},$$

from which we verify that $\gamma_{k,k'+e_j,\ell,\ell''}\geq 0$ and

$$\sum_{\ell\in I^d_k}\gamma_{k,k'+e_j,\ell,\ell''} = \sum_{\ell'\in I^d_{k'}} 2^{-m+1/2}\mathbf{1}_{\{\ell'_i=\ell''_i\ (i\neq j)\}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}}\sum_{\ell\in I^d_k}\gamma_{k,k',\ell,\ell'}$$
$$\leq 2^{-m+1/2-\|k'-k\|_1/2}\sum_{\ell'_j\in\mathbb{Z}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}} = 2^{-(\|k'-k\|_1+1)/2},$$

and furthermore

$$\sum_{\ell''\in I^d_{k'+e_j}}\gamma_{k,k'+e_j,\ell,\ell''} = \sum_{\ell'\in I^d_{k'}}\sum_{\ell''\in I^d_{k'+e_j}} 2^{-m+1/2}\mathbf{1}_{\{\ell'_i=\ell''_i\ (i\neq j)\}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}}\gamma_{k,k',\ell,\ell'}$$
$$\leq 2^{-m+1/2}\sum_{\ell'\in I^d_{k'}}\sum_{\ell''_j\in\mathbb{Z}}\binom{m}{\ell''_j-2\ell'_j+\frac{m}{2}}\gamma_{k,k',\ell,\ell'}$$
$$\leq 2^{1/2}\sum_{\ell'\in I^d_{k'}}\gamma_{k,k',\ell,\ell'}\leq 2^{(\|k'-k\|_1+1)/2}.$$

This concludes the proof. □

C.4 Proof of Corollary 4.8

In order to approximate arbitrary $\psi_j^\circ\in\mathbb{U}(B^\tau_{p,q}(\mathcal{X}))$, we need the following construction instead of Lemma C.5. Note that N' corresponds to the number of B-splines used to approximate the target function and can be freely chosen to match the desired error, which however affects the covering number of $\mathcal{F}_N$.

Lemma C.11 (Suzuki (2019), Proposition 1)**.** *Set $m\in\mathbb{N}$, $m>\tau+2-1/p$ and $\nu=(p\tau-d)/2d$. For all $N'\in\mathbb{N}$ sufficiently large and $\epsilon=N'^{-\tau/d}(\log N')^{-1}$, for any $f^\circ\in\mathbb{U}(B^\tau_{p,q}(\mathcal{X}))$ there exists a ReLU network $\tilde{f}\in\mathcal{F}_{\mathrm{DNN}}(L,W,S,M)$ with*

$$L = 3+2\left\lceil\log_2\left(3^{d\vee m}(1+dm^{-1/2}(2e)^{m+1})\epsilon^{-1}\right)+5\right\rceil\lceil\log_2(d\vee m)\rceil,$$
$$W=N'W_0,\qquad S=((L-1)W_0^2+1)N',\qquad M=O(N'^{1/\nu+1/d})$$

satisfying $\|f^\circ-\tilde{f}\|_{L^\infty(\mathcal{X})}\leq N'^{-\tau/d}$.

Also note that from Assumption 1 it follows that $\|\psi_j\|_{L^\infty(\mathcal{P}_\mathcal{X})}\leq C_\infty N^{1/2}$. Setting $N'\asymp\delta_N^{-d/\tau}$ and applying the covering number bound in Lemma C.6, after some algebra we obtain the following counterpart to Lemma 4.4.

Corollary C.12. *For any $\delta_N>0$, Assumption 3 is satisfied by taking*

$$\mathcal{F}_N=\{\Pi_{B'_N}\circ\phi\mid\phi=(\phi_j)_{j=1}^N,\phi_j\in\mathcal{F}_{\mathrm{DNN}}(L,W,S,M)\}$$

where $B'_N=C_\infty N^{1/2}$ and

$$L=O(\log\delta_N^{-1}),\quad W=O(\delta_N^{-d/\tau}),\quad S=O(\delta_N^{-d/\tau}\log\delta_N^{-1}),\quad \log M=O(\log\delta_N).$$

Also, the metric entropy of $\mathcal{F}_N$ is bounded as

$$\mathcal{V}(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \epsilon) \lesssim N\delta_N^{-d/\tau} \log\frac{1}{\delta_N}\left(\log\frac{1}{\epsilon} + \log^2\frac{1}{\delta_N}\right).$$

Then by combining with Lemma 3.3 and Proposition 3.2 via Theorem 3.1, it follows that

$$\mathcal{V}(\mathcal{T}_N, \|\cdot\|_{L^\infty}, \epsilon) \lesssim N^2 \log\frac{N}{\epsilon} + N\delta_N^{-d/\tau} \log\frac{1}{\delta_N}\left(\log\frac{N}{\epsilon} + \log^2\frac{1}{\delta_N}\right).$$

and

$$\begin{aligned}\bar{R}(\widehat{\Theta}) \lesssim \frac{N}{n}\log N + \frac{N^2}{n^2}\log^2 N + N^{-2\alpha/d} + N^2\delta_N^2 \\ + \frac{N^2}{T}\log\frac{N}{\epsilon} + \frac{N}{T}\delta_N^{-d/\tau}\log\frac{1}{\delta_N}\left(\log\frac{N}{\epsilon} + \log^2\frac{1}{\delta_N}\right) + \epsilon.\end{aligned}$$

Substituting $\delta_N \asymp N^{-1-\alpha/d}$ and $\epsilon \asymp N^{-2\alpha/d}$ concludes the desired bound. □

D Details on Sequential Input

D.1 Definitions and Results

γ-smooth class. We first define the γ-smooth function class introduced by Okumoto and Suzuki (2022). Let $r \in \mathbb{Z}_0^{d\times\infty}$ and $s \in \bar{\mathbb{N}}_0^{d\times\infty}$, where $\bar{\mathbb{N}} = \mathbb{N}\cup\{0\}$ and the subscript 0 indicates restriction to the subset of elements with a finite number of nonzero components. Consider the orthonormal basis $(\psi_r)_r$ of $L^2([0,1]^{d\times\infty})$ given as

$$\psi_r(x) = \prod_{i\in\mathbb{Z}}\prod_{j=1}^{d}\psi_{r_{ij}}(x_{ij}), \quad \psi_{r_{ij}}(x_{ij}) = \begin{cases}\sqrt{2}\cos(2\pi r_{ij}x_{ij}) & r_{ij} < 0\\ 1 & r_{ij} = 0\\ \sqrt{2}\sin(2\pi r_{ij}x_{ij}) & r_{ij} > 0\end{cases}.$$

The frequency s component $\delta_s(f)$ of $f \in L^2([0,1]^{d\times\infty})$ is defined as

$$\delta_s(f) := \sum_{\lfloor 2^{s_{ij}-1}\rfloor \le |r_{ij}| < 2^{s_{ij}}} \langle f, \psi_r\rangle\psi_r.$$

For a monotonically nondecreasing function $\gamma : \bar{\mathbb{N}}_0^{d\times\infty} \to \mathbb{R}$ and $p \geq 2, q \geq 1$, the γ-smooth norm and function class are defined as

$$\|f\|_{\mathcal{F}^\gamma_{p,q}(\mathcal{P}_\mathcal{X})} := \left(\sum_{s\in\bar{\mathbb{N}}_0^{d\times\infty}} 2^{q\gamma(s)}\|\delta_s(f)\|^q_{p,\mathcal{P}_\mathcal{X}}\right)^{1/q}$$

and

$$\mathcal{F}^\gamma_{p,q}(\mathcal{P}_\mathcal{X}) := \left\{f \in L^2([0,1]^{d\times\infty}) \mid \|f\|_{\mathcal{F}^\gamma_{p,q}(\mathcal{P}_\mathcal{X})} < \infty\right\}.$$

The γ-smooth class over finite-dimensional input space $[0,1]^{d\times m}$ is similarly defined.

In particular, we consider two specific cases of γ for the component-wise smoothness parameter $\alpha \in \mathbb{R}_{>0}^{d\times\infty}$, for which we also define the corresponding *degrees of smoothness* $\alpha^\dagger \in \mathbb{R}_{>0}$. Denote by $(\tilde{\alpha}_j)_{j=1}^\infty$ all components of α sorted by ascending magnitude.

- Mixed smoothness: $\gamma(s) = \langle\alpha, s\rangle$, $\alpha^\dagger = \tilde{\alpha}_1 = \max_{i,j}\alpha_{ij}$.
- Anisotropic smoothness: $\gamma(s) = \max_{i,j}\alpha_{ij}s_{ij}$, $\alpha^\dagger = (\sum_{i,j}\alpha_{ij}^{-1})^{-1}$.

Furthermore, the weak l^η-norm of α is defined as $\|\alpha\|_{wl^\eta} := \sup_j j^\eta\tilde{\alpha}_j^{-1}$ for $\eta > 0$.

Piecewise γ-smooth class. The piecewise γ-smooth class is an extension of the γ-smooth class allowing for arbitrary bounded permutations of the tokens of an input (Takakura and Suzuki, 2023). For a threshold $V \in \mathbb{N}$ and an index set Λ, let $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ be a disjoint partition of $\operatorname{supp}\mathcal{P}_\mathcal{X}$ and $\{\pi_\lambda\}_{\lambda\in\Lambda}$ a set of bijections from $[2V+1]$ to $[-V:V]$. Further define the permutation operator $\Pi : \operatorname{supp}\mathcal{P}_\mathcal{X} \to \mathbb{R}^{d\times(2V+1)}$ as

$$\Pi(x) = (x_{\pi_\lambda(1)}, \cdots, x_{\pi_\lambda(2V+1)}), \quad \text{if } x \in \Omega_\lambda.$$

Then the piecewise γ-smooth function class is defined as

$$\mathscr{P}^\gamma_{p,q}(\mathcal{P}_\mathcal{X}) := \left\{ g = f \circ \Pi \mid f \in \mathcal{F}^\gamma_{p,q}(\mathcal{P}_\mathcal{X}), \|g\|_{\mathscr{P}^\gamma_{p,q}(\mathcal{P}_\mathcal{X})} < \infty \right\},$$

where

$$\|g\|_{\mathscr{P}^\gamma_{p,q}(\mathcal{P}_\mathcal{X})} := \left(\sum_{s\in\bar{\mathbb{N}}_0^{d\times[-V:V]}} 2^{q\gamma(s)} \|\delta_s(f)\circ\Pi\|^q_{p,\mathcal{P}_\mathcal{X}} \right)^{1/q}.$$

Next, we state the set of assumptions inherited from Takakura and Suzuki (2023). In particular, the *importance function* makes precise a notion of relative importance between tokens which is preserved by permutations.

Assumption 5 (smoothness and importance function)**.** $1 < q \le 2$ *and:*

1. *The smoothness parameter α satisfies $\|\alpha\|_{wl^\eta} \le 1$ and $\alpha_{ij} = \Omega(|i|^\eta)$ for some $\eta > 1$. For mixed smoothness, we also require $\tilde{\alpha}_1 < \tilde{\alpha}_2$.*
2. *There exists a shift-equivariant map $\mu : \operatorname{supp}\mathcal{P}_\mathcal{X} \to \mathbb{R}^\infty$ such that $\mu_0 \in \mathbb{U}(\mathcal{F}^\gamma_{\infty,q})$, $\|\mu_0\| \le 1$ and $\Omega_\lambda = \{x \in \operatorname{supp}\mathcal{P}_\mathcal{X} \mid \mu(x)_{\pi_\lambda(1)} > \cdots > \mu(x)_{\pi_\lambda(2V+1)}\}$ for all $\lambda \in \Lambda$. μ is moreover well-separated, that is $\mu(x)_{\pi_\lambda(v)} - \mu(x)_{\pi_\lambda(v+1)} \ge C_\mu v^{-\varrho}$ for $C_\mu, \varrho > 0$.*

We focus on parameter ranges $1 < q \le 2$ and $\eta > 1$ strictly for simplicity of presentation, but the cases $q = 1, q > 2$ and $\eta > 0$ can be handled with some more analysis. Note that $\eta > 1$ ensures $\alpha^\dagger > 0$ for anisotropic smoothness.

Additionally, the assumption pertaining to our ICL setup is stated as follows.

Assumption 6. *For $r \in \mathbb{Z}_0^{d\times\infty}$ the coefficients β_r corresponding to ψ_r are independent and satisfy for $s \in \bar{\mathbb{N}}_0^{d\times\infty}$ such that the frequency component $\delta_s(f)$ contains the element ψ_r,*

$$\mathbb{E}_\beta[\beta_r] = 0, \quad \mathbb{E}_\beta[\beta_r^2] \lesssim 2^{-(2+1/\alpha^\dagger)\gamma(s)}\gamma(s)^{-2}. \tag{22}$$

Also $\Sigma_{\Psi,N} \asymp \mathbf{I}_N$ holds, for example $\mathcal{P}_\mathcal{X}$ is bounded above and below with respect to the product measure $\lambda^{d\times\infty}$ on $\mathscr{B}([0,1]^{d\times\infty})$ of the uniform measure λ on $\mathscr{B}([0,1])$.

We then obtain the following result for ICL with transformers:

Theorem D.1 (minimax optimality of ICL for sequential input)**.** *Let $\mathcal{F}^\circ = \{f \in \mathbb{U}(\mathscr{P}^\gamma_{p,q}(\mathcal{P}_\mathcal{X})) \mid \|f\|_{L^\infty(\mathcal{P}_\mathcal{X})} \le B\}$ for some $B > 0$ where γ corresponds to mixed or anisotropic smoothness. Suppose Assumptions 5 and 6 hold. Then for $n \gtrsim N\log N$ we have*

$$\bar{R}(\widehat{\Theta}) \lesssim N^{-2\alpha^\dagger} + \frac{N\log N}{n} + \frac{N^{2\vee(1+1/\alpha^\dagger)}\operatorname{polylog}(N)}{T}.$$

Hence if $T \gtrsim nN^{1\vee 1/\alpha^\dagger}$ and $N \asymp n^{\frac{1}{2\alpha^\dagger+1}}$, ICL achieves the rate $n^{-\frac{2\alpha^\dagger}{2\alpha^\dagger+1}}\operatorname{polylog}(n)$.

D.2 Proof of Theorem D.1

Since the system $(\psi_r)_r$ is orthonormal, we may take $\underline{N} = 1, \bar{N} = N$ following Remark 2.1. We mainly utilize the following approximation and covering number bounds.

Theorem D.2 (Takakura and Suzuki (2023), Theorem 4.5)**.** *For a function $F^\circ \in \mathbb{U}(\mathscr{P}^\gamma_{p,q}(\mathcal{P}_\mathcal{X}))$, $\|F^\circ\|_{L^\infty(\mathcal{P}_\mathcal{X})} \le B$ and any $K > 0$, there exists a transformer $\widehat{F} \in \mathcal{F}_{\mathrm{TF}}(J,U,D,H,L,W,S,M)$ such that*

$$\|\widehat{F}_0 - F^\circ\|_{L^2(\mathcal{P}_\mathcal{X})} \lesssim 2^{-K},$$

where

$$J \lesssim K^{1/\eta}, \quad \log U \lesssim \log K \vee \log V, \quad D \lesssim K^{2(1+\varrho)/\eta} \log V, \quad H \lesssim (\log K)^{1/\eta},$$
$$L \lesssim K^2, \quad W \lesssim 2^{K/\alpha^\dagger} K^{1/\eta}, \quad S \lesssim 2^{K/\alpha^\dagger} K^{2+2/\eta}, \quad \log M \lesssim K \vee \log\log V.$$

Theorem D.3 (Takakura and Suzuki (2023), Theorem 5.3)**.** *For $\epsilon > 0$ and $B \geq 1$ it holds that*

$$\log \mathcal{N}(\mathcal{F}_{\text{TF}}(J,U,D,H,L,W,S,M), \|\cdot\|_{L^\infty}, \epsilon) \lesssim J^3 L(S + HD^2) \log\left(\frac{DHLWM}{\epsilon}\right).$$

To analyze the decay rate in the γ-smooth class, we approximate a function $f \in L^2([0,1]^{d\times\infty})$ by the partial sum of its frequency components up to 'resolution' K, measured via the γ function:

$$R_K(f) := \sum_{\gamma(s)<K} \delta_s(f).$$

The basis functions ψ_r are thus ordered primarily ordered by increasing $\gamma(s)$.

Lemma D.4 (Okumoto and Suzuki (2022), Lemma 17)**.** *For $1 \leq q \leq 2$ it holds that*

$$\|f - R_K(f)\|_{L^2(\mathcal{P}_\mathcal{X})} \lesssim 2^{-K} \|f\|_{\mathcal{F}^\gamma_{p,q}(\mathcal{P}_\mathcal{X})}.$$

Note that if $\gamma(s) < K$ then $s_{ij} < K/a_{ij} \lesssim K/|i|^\eta$ for all i,j for both types of smoothness and so $\|s\|_0 \lesssim dK^{1/\eta}$. In addition, the number of basis functions ψ_r used in the sum for $\delta_s(f)$ is exactly $2^{\|s\|_1} = \prod_{i,j} 2^{s_{ij}}$. Theorem D.3 of Takakura and Suzuki (2023) shows that the number of basis elements used in the sum $R_K(f)$ satisfies

$$N \asymp \sum_{\gamma(s)<K} 2^{\|s\|_1} \lesssim 2^{K/\alpha^\dagger}$$

for both mixed and anisotropic smoothness. Hence the N-term approximation error decays as $N^{-\alpha^\dagger}$ so that the choice $s = \alpha^\dagger$ leading to the assumed variance bound $N^{-2\alpha^\dagger-1} \asymp 2^{-(2+1/\alpha^\dagger)K}$ in (22) is justified. Moreover for large K,

$$\sum_{\gamma(s)<K} \sum_{\lfloor 2^{s_{ij}-1} \rfloor \leq |r_{ij}| < 2^{s_{ij}}} \|\psi_r\|^2_{L^\infty(\mathcal{P}_\mathcal{X})} \leq \sum_{\gamma(s)<K} 2^{\|s\|_1} (\sqrt{2}^{\|s\|_0})^2 \lesssim 2^{K/\alpha^\dagger + O(K^{1/\eta})}$$

since $\|r\|_0 = \|s\|_0$, so that (2) of Assumption 1 is satisfied with $r = 1/2$. The second part of Assumption 1 holds since $(\psi_r)_{r\in\mathbb{Z}_0^{d\times\infty}}$ is orthonormal w.r.t. $\lambda^{d\times\infty}$. Furthermore, the discussion thus far immediately extends to the piecewise γ-smooth class for any partition $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ by composing with the permutation operator Π.

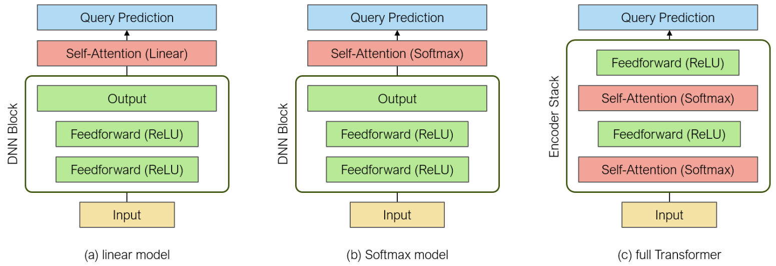
We proceed to use Theorem D.2 to approximate each basis function $\psi_r \circ \Pi$ up to resolution K. Moreover, we can see from the proof of Lemma 17 of Okumoto and Suzuki (2022) that we do not need to account for the sup-norm scaling of ψ_r and thus it suffices to find the parameter $K' \in \mathbb{N}$ such that the approximation error $2^{-K'} \asymp \delta_N$. Hence combining Theorems D.2 and D.3, we conclude that

$$\begin{aligned}\mathcal{V}(\mathcal{F}_{\text{TF}}(J,U,D,H,L,W,S,M), \|\cdot\|_{L^\infty}, \epsilon) &\lesssim K'^{3/\eta} K'^2 \cdot 2^{K'/\alpha^\dagger} K'^{2+2/\eta} \cdot K' \log\frac{1}{\epsilon} \\ &\lesssim \left(\frac{1}{\delta_N}\right)^{1/\alpha^\dagger} \operatorname{polylog}\left(N, \frac{1}{\delta_N}\right) \log\frac{1}{\epsilon}\end{aligned}$$

is sufficient to satisfy Assumption 3. Therefore, we can now apply our framework with $B'_N \asymp N$ to obtain the bound

$$\begin{aligned}\bar{R}(\widehat{\Theta}) \lesssim{}& \frac{N}{n}\log N + \frac{N^2}{n^2}\log^2 N + N^{-2\alpha^\dagger} + N^2\delta_N^2 \\ &+ \frac{N^2}{T}\log\frac{N}{\epsilon} + \frac{1}{T}\delta_N^{-1/\alpha^\dagger} \operatorname{polylog}\left(N, \frac{1}{\delta_N}\right)\log\frac{1}{\epsilon} + \epsilon.\end{aligned}$$

Substituting $\delta_N \asymp N^{-1-\alpha^\dagger}$ and $\epsilon \asymp N^{-2\alpha^\dagger}$ concludes the theorem. □



(a) linear model (b) Softmax model (c) full Transformer

Figure 1: Architecture of the compared models. Each model contains two MLP components, all attention layers are single-head and LayerNorm is not included. (a),(b) implement the simplified reparametrization for attention, while all layers in (c) utilize the full embeddings. The input dimension is 8 and all hidden layer and DNN output widths are 32. The query prediction is read off the last entry of the output at the query position.

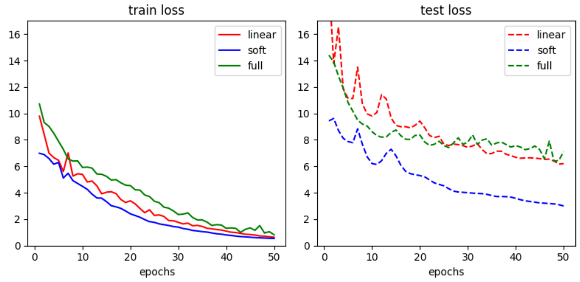


Figure 2: Training and test curves for the ICL pretraining objective. We use the Adam optimizer with a learning rate of 0.02 for all layers. For the task class we take $\alpha = 1$, $p = q = \infty$, $T = n = 512$ and generate samples from random combinations of order 2 wavelets.

E Numerical Experiments

In this section, we connect our theoretical contributions to practical transformers by conducting experiments verifying our results as well as justifying the simplified model setup and the empirical risk minimization assumption. We implement and compare the following toy models: (a) the simplified architecture studied in our paper; (b) the same model with linear attention replaced by softmax; and (c) a full transformer with 2 stacked encoder layers. The number of feedforward layers, widths of hidden layers, learning rate, etc. are set to equal for a fair comparison, see Figure 1 for details.

Figure 2 shows training (solid) and test loss curves (dashed) during pretraining. All 3 architectures exhibit similar behavior and converge to near zero training loss, justifying the use of our simplified model and also supporting the assumption that the empirical risk is minimized. Moreover, Figure 3 shows the converged losses over a wide range of N, n, T values. We verify that increasing N, n

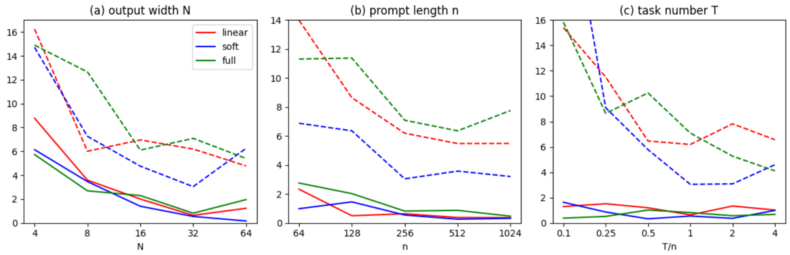


Figure 3: Training and test losses of the three models after 50 epochs while varying (a) DNN width N; (b) number of in-context samples n; (c) number of tasks T. For (a), the widths of all hidden layers also vary with N. We take the median over 5 runs for robustness.

leads to decreasing train and test error, corresponding to the approximation error of Theorem 3.1. We also observe that increasing T tends to improve the pretraining generalization gap up to a threshold, confirming our theoretical analysis of task diversity. Again, this behavior is consistent across the 3 architectures. We note that in the overparametrized regime when the number of total parameters $\gtrsim nT$, the trained model is likely not the empirical risk minimizer, which may also contribute to the large error for large N or small n, T.

F Proofs of Minimax Lower Bounds

F.1 Proof of Proposition 5.1

In this section, we develop our framework for obtaining minimax lower bounds in the ICL setup by adapting the information-theoretic approach of Yang and Barron (1999).

Let $\{(\psi^{(j)}, \beta_{T+1}^{(j)})\}_{j=1}^M$ be a δ_n-packing of the class $\mathcal{F}^\circ$ with respect to the $L^2(\mathcal{X})$-norm such that

$$\|\beta_{T+1}^{(j)\top}\psi^{(j)} - \beta_{T+1}^{(j')\top}\psi^{(j')}\|_{L^2(\mathcal{X})}^2 \geq \delta_n^2, \quad 1 \leq j < j' \leq M,$$

where M is the corresponding packing number. Then we have the following proposition as an application of Fano's inequality (Yang and Barron, 1999).

Proposition F.1. *Let Θ be a random variable uniformly distributed over $\{(\psi^{(j)}, \beta_{T+1}^{(j)})\}_{j=1}^M$. Then, it holds that*

$$\inf_{\widehat{f}_n:\mathcal{D}_{n,T}\to\mathbb{R}} \sup_{f^\circ\in\mathcal{F}^\circ} \mathbb{E}_{\mathcal{D}_{n,T}}[\|\widehat{f} - f^\circ\|_{L^2(\mathcal{P}_\mathcal{X})}^2] \geq \frac{\delta_n^2}{2}\left(1 - \frac{\mathbb{E}_\mathbf{X}[I_{\mathbf{X}^{(1:T+1)}}(\Theta, \boldsymbol{y}^{(1:T+1)})] + \log 2}{\log M}\right),$$

where $I_{\mathbf{X}^{(1:T+1)}}(\Theta, \boldsymbol{y}^{(1:T+1)})$ is the mutual information between $\Theta, \boldsymbol{y}^{(1:T+1)}$ for given $\mathbf{X}^{(1:T+1)}$.

The mutual information $I_{\mathbf{X}^{(1:T+1)}}(\Theta, \boldsymbol{y}^{(1:T+1)})$ is formulated more concretely as

$$\sum_{\theta\in\operatorname{supp}\Theta} w(\theta) \int p(\boldsymbol{y}^{(1:T+1)}|\theta, \mathbf{X}^{(1:T+1)}) \log\left(\frac{p(\boldsymbol{y}^{(1:T+1)}|\theta, \mathbf{X}^{(1:T+1)})}{p_w(\boldsymbol{y}^{(1:T+1)}|\mathbf{X}^{(1:T+1)})}\right) \mathrm{d}\boldsymbol{y}^{(1:T+1)},$$

where $p(\boldsymbol{y}|\theta, \mathbf{X})$ is the probability density of $\boldsymbol{y}$ conditioned on $\theta, \mathbf{X}$ and p_w is the marginal distribution of $\boldsymbol{y}^{(1:T+1)}$ where $w(\cdot) \equiv \frac{1}{M}$ is the probability mass function of Θ (i.e., $p_w(\cdot|\mathbf{X}^{(1:T+1)}) = \sum_{\theta\in\operatorname{supp}\Theta} w(\theta)p(\cdot|\theta, \mathbf{X}^{(1:T+1)})$). We let $P_{\boldsymbol{y}^{(t)}|\theta}$ (and $P_{\boldsymbol{y}^{(1:t)}|\theta}$) be the distribution of $\boldsymbol{y}^{(t)}$ conditioned on $\theta, \mathbf{X}^{(t)}$ (and $\mathbf{X}^{(1:t)}$) respectively, and let

$$\bar{P}_{\boldsymbol{y}^{(1:T+1)}} = \frac{1}{M}\sum_{j=1}^M P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}}$$

be the marginal distribution of $\boldsymbol{y}^{(1:T+1)}$ conditioned on $\mathbf{X}^{(1:T+1)}$.

Next, we define the set $\{\tilde{\psi}^{(j)}\}_{j=1}^{Q_1}$ to be a $\varepsilon_{n,1}$-covering of $\mathcal{F}_N$ w.r.t. the norm $d(\psi, \psi') := \sqrt{\mathbb{E}_x[\|\psi(x) - \psi'(x)\|^2]}$ with $\varepsilon_{n,1}$-covering number Q_1, and $\{\tilde{\beta}^{(j)}\}_{j=1}^{Q_2}$ to be a $\varepsilon_{n,2}$-covering of $\mathcal{B}$ w.r.t. the L^2 norm with $\varepsilon_{n,2}$-covering number Q_2. By taking all combinations of $(\tilde{\psi}^{(j)}, \tilde{\beta}^{(j')})$ for $1 \leq j \leq Q_1$ and $1 \leq j' \leq Q_2$, we obtain the covering $\{\tilde{\theta}^{(j)}\}_{j=1}^Q$ with respect to the quantity $\varepsilon_n^2 = \sigma_\beta^2\varepsilon_{n,1}^2 + C_2\varepsilon_{n,2}^2$ where $Q = Q_1Q_2$ and each $\tilde{\theta}^{(j)}$ is given by $\tilde{\theta}^{(j)} = (\tilde{\psi}^{(j_1)}, \tilde{\beta}^{(j_2)})$ for some indices j_1 and j_2.

Then as in the discussion of Yang and Barron (1999), the mutual information is bounded by

$$\begin{aligned} I_{\mathbf{X}^{(1:T+1)}}(\Theta, \boldsymbol{y}^{(1:T+1)}) &= \frac{1}{M}\sum_{j=1}^M D(P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}} \| \bar{P}_{\boldsymbol{y}^{(1:T+1)}}) \\ &\leq \frac{1}{M}\sum_{j=1}^M D(P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}} \| \tilde{P}_{\boldsymbol{y}^{(1:T+1)}}), \end{aligned}$$

where $D(\cdot\|\cdot)$ is the Kullback-Leibler divergence and $\tilde{P}_{\boldsymbol{y}^{(1:T+1)}} = \frac{1}{Q}\sum_{j=1}^{Q} P_{\boldsymbol{y}^{(1:T+1)}|\tilde{\theta}^{(j)}}$ because $\bar{P}_{\boldsymbol{y}^{(1:T+1)}}$ minimizes the right hand side. If we let

$$\kappa(j) := \operatorname*{arg\,min}_{1\le k\le Q} D(P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}}\|P_{\boldsymbol{y}^{(1:T+1)}|\tilde{\theta}^{(k)}}),$$

then each summand of the right-hand side is further bounded by

$$\log Q + D(P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}}\|P_{\boldsymbol{y}^{(1:T+1)}|\tilde{\theta}^{\kappa(j)}}).$$

Moreover, for $\theta = (\psi, \beta^{(T+1)})$ it holds that

$$\begin{aligned}
&p(\boldsymbol{y}^{(1:T+1)}|\theta, \mathbf{X}^{(1:T+1)}) \\
&= \prod_{t=1}^{T} p(\boldsymbol{y}^{(t)}|\psi, \mathbf{X}^{(t)}) \cdot p(\boldsymbol{y}^{(T+1)}|\psi, \beta^{(T+1)}, \mathbf{X}^{(T+1)}) \\
&= \prod_{t=1}^{T} \int p(\boldsymbol{y}^{(t)}|\psi, \beta^{(t)}, \mathbf{X}^{(t)}) p_\beta(\beta^{(t)})\mathrm{d}\beta^{(t)} \cdot p(\boldsymbol{y}^{(T+1)}|\psi, \beta^{(T+1)}, \mathbf{X}^{(T+1)}).
\end{aligned}$$

Then the KL-divergence can be bounded as

$$\begin{aligned}
&D(P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}}\|P_{\boldsymbol{y}^{(1:T+1)}|\tilde{\theta}^{\kappa(j)}}) \\
&= \sum_{t=1}^{T} D\left(P_{\boldsymbol{y}^{(t)}|\psi^{(j)}}\|P_{\boldsymbol{y}^{(t)}|\tilde{\psi}^{(\kappa(j))}}\right) + D\left(P_{\boldsymbol{y}^{(T+1)}|\psi^{(j)},\beta^{(j)}_{T+1}}\|P_{\boldsymbol{y}^{(T+1)}|\tilde{\psi}^{(\kappa(j))},\tilde{\beta}^{(\kappa(j))}_{T+1}}\right) \\
&\le \sum_{t=1}^{T} \int D\left(P_{\boldsymbol{y}^{(t)}|\psi^{(j)},\beta^{(t)}}\|P_{\boldsymbol{y}^{(t)}|\tilde{\psi}^{(\kappa(j))},\beta^{(t)}}\right) p_\beta(\beta^{(t)})\mathrm{d}\beta^{(t)} \\
&\quad + D\left(P_{\boldsymbol{y}^{(T+1)}|\psi^{(j)},\beta^{(j)}_{T+1}}\|P_{\boldsymbol{y}^{(T+1)}|\tilde{\psi}^{(\kappa(j))},\tilde{\beta}^{(\kappa(j))}_{T+1}}\right),
\end{aligned}$$

where the joint convexity of KL-divergence was used for the last inequality. Since the observation noise is assumed to be normally distributed, the integrand KL-divergence can be bounded as

$$D\left(P_{\boldsymbol{y}^{(t)}|\psi^{(j)},\beta}\|P_{\boldsymbol{y}^{(t)}|\tilde{\psi}^{\kappa(j)},\beta}\right) = \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left(\beta^\top\psi^{(j)}(x_i^{(t)}) - \beta^\top\tilde{\psi}^{(\kappa(j))}(x_i^{(t)})\right)^2.$$

Hence, its expectation with respect to $\beta, \mathbf{X}^{(t)}$ becomes

$$\mathbb{E}_{\mathbf{X}^{(t)},\beta}\left[D\left(P_{\boldsymbol{y}^{(t)}|\psi^{(j)},\beta}\|P_{\boldsymbol{y}^{(t)}|\tilde{\psi}^{\kappa(j)},\beta}\right)\right] = \frac{n\sigma_\beta^2}{2\sigma^2}\|\psi^{(j)} - \tilde{\psi}^{\kappa(j)}\|^2_{L^2(\mathcal{P}_\mathcal{X})} \le \frac{n\sigma_\beta^2}{2\sigma^2}\varepsilon^2_{n,1},$$

In the same manner, we have that

$$\begin{aligned}
&D\left(P_{\boldsymbol{y}^{(T+1)}|\psi^{(j)},\beta^{(j)}_{T+1}}\|P_{\boldsymbol{y}^{(T+1)}|\tilde{\psi}^{\kappa(j)},\tilde{\beta}^{(\kappa(j))}_{T+1}}\right) \\
&= \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\beta^{(j)\top}_{T+1}\psi^{(j)}(x_i^{(t)}) - \tilde{\beta}^{(\kappa(j))\top}_{T+1}\tilde{\psi}^{(\kappa(j))}(x_i^{(t)})\right)^2 \\
&\le \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left[\left(\beta^{(j)}_{T+1} - \tilde{\beta}^{(\kappa(j))}_{T+1}\right)^\top\tilde{\psi}^{(\kappa(j))}(x_i^{(t)})\right]^2 + \sum_{i=1}^{n}\frac{1}{2\sigma^2}\left[\beta^{(j)\top}_{T+1}(\psi^{(j)}(x_i^{(t)}) - \tilde{\psi}^{(\kappa(j))}(x_i^{(t)}))\right]^2.
\end{aligned}$$

The expectation of the right-hand side with respect to $\mathbf{X}^{(T+1)}, \beta^{(j)}_{T+1}$ is bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}^{(T+1)},\beta^{(j)}_{T+1}}\left[D\left(P_{\boldsymbol{y}^{(T+1)}|\psi^{(j)},\beta^{(j)}_{T+1}}\|P_{\boldsymbol{y}^{(t)}|\tilde{\psi}^{\kappa(j)},\tilde{\beta}^{(\kappa(j))}_{T+1}}\right)\right] \\
&\le \frac{C_2 n}{2\sigma^2}\|\beta^{(j)}_{T+1} - \tilde{\beta}^{(\kappa(j))}_{T+1}\|^2 + \frac{n}{2\sigma^2}\sigma_\beta^2\|\psi^{(j)} - \tilde{\psi}^{(\kappa(j))}\|^2_{L^2(\mathcal{P}_\mathcal{X})} \\
&\le \frac{C_2 n}{2\sigma^2}\varepsilon^2_{n,2} + \frac{n}{2\sigma^2}\sigma_\beta^2\varepsilon^2_{n,1} = \frac{n}{2\sigma^2}\varepsilon^2_n.
\end{aligned}$$

Therefore, the expected mutual information can be bounded as

$$\mathbb{E}_X[I_{\mathbf{X}^{(1:T+1)}}(\Theta, \boldsymbol{y}^{(1:T+1)})] \le \log Q_1 + \log Q_2 + \frac{nT}{2\sigma^2}\sigma_\beta^2\varepsilon^2_{n,1} + \frac{n}{2\sigma^2}\varepsilon^2_n.$$

Applying Proposition F.1 together with (7) concludes the proof. □

F.2 Lower Bound in Besov Space

Here, we derive the minimax lower bound when $\mathcal{F}^\circ = \mathbb{U}(B^\alpha_{p,q}(\mathcal{X}))$. Recall that in this setting $s = \alpha/d$. We fix a resolution K and then consider the set of B-splines $\omega^d_{K,\ell}, \ell \in I^d_K$ of cardinality $N' \asymp 2^{Kd}$. Considering the basis pairs $(\omega^d_{K,1}, \omega^d_{K,2}), \dots, (\omega^d_{K,N'-1}, \omega^d_{K,N'})$, we can determine which one is employed to construct the basis $\psi^{(j)}$. The Varshamov-Gilbert bound yields that for $\Omega = \{0,1\}^{N'/2}$, we can construct a subset $\Omega' = \{w_1, \dots, w_{2^{N'/16}}\} \subset \Omega$ such that $|\Omega| = 2^{N'/16}$ and $w \neq w' \in \Omega'$ has a Hamming distance not less than $N'/16$. Using this Ω', we set $N = N'/2$ and $M = 2^{N'/16}$ and define $(\psi^{(j)})_{j=1}^M$ as $\psi^{(j)}_i = \omega^d_{K,2i-1}$ if $w_{j,i} = 0$ and $\psi^{(j)}_i = \omega^d_{K,2i}$ if $w_{j,i} = 1$. We use the same B-spline bases with resolution more than K for $\psi^{(j)}_i$ $(i \geq N)$ across all j.

By the construction of $(\psi^{(j)})$, if we set $\beta^{(1)} = (\sigma_\beta, \dots, \sigma_\beta, 0, 0, \dots)$, then

$$\|\beta^{(1)\top}\psi^{(j)} - \beta^{(1)\top}\psi^{(j')}\|^2_{L^2(\mathcal{P}_\mathcal{X})} \geq \sigma^2_\beta N/8.$$

Hence, for $\delta^2_n \leq \sigma^2_\beta N/8 \lesssim 1$, the δ_n-packing number is not less than $2^{N/8}$. Moreover, the logarithmic δ_n-packing number of $\{\beta^\top\psi^{(j)} \mid \beta \in \mathcal{B}\}$ for a fixed j is $\Theta(\min\{\delta_n^{-1/s}, N\log(1/\delta_n)\})$ by the standard argument.

Hence taking $\delta_n = N^{-s}$, we obtain $\log M \gtrsim N$ and the upper bound of the covering numbers $\log Q_1 + \log Q_2 \lesssim N$ for $\sigma^2_\beta\varepsilon^2_{n,1} \leq \delta^2_n$ and $\varepsilon^2_{n,2} = C\delta^2_n$ where C is a constant. Then, by choosing C appropriately and $\varepsilon_{n,1} \lesssim N^{-1-s}$ (so that $\log Q_1 \lesssim N$), as long as

$$nT\sigma^2_\beta\varepsilon^2_{n,1} + n\delta^2_n \lesssim \log Q_1 + \log Q_2 \lesssim N$$

is satisfied, the minimax rate is lower bounded by δ^2_n. Taking $N \asymp n^{\frac{1}{2s+1}}$, we obtain the lower bound

$$\delta^2_n \gtrsim n^{-\frac{2s}{2s+1}}. \tag{23}$$

F.3 Lower Bound with Coarser Basis

We consider a generalized setting where $\mathcal{X} = \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{(N+1)\text{times}}$ and take $\psi^{(j)}_i \in \mathbb{U}(B^\tau_{p,q}(\mathbb{R}^d))$ and assume that $\beta_1 \in [-1,1]$ and $\beta_j \in [-\sigma_\beta, \sigma_\beta]$ where $\sigma^2_\beta = \tilde{\Theta}(N^{-2s-1})$. Since the logarithmic $\tilde{\varepsilon}_1$-covering and packing numbers of $\mathbb{U}(B^\tau_{p,q}(\mathbb{R}^d))$ are $\Theta(\tilde{\varepsilon}_1^{-d/\tau})$, those for the basis functions on $j = 2, \dots, N+1$ become $\Theta(N\tilde{\varepsilon}_1^{-d/\tau})$, and those for $\mathcal{B}$ are $\Theta(N\log(1 + \frac{N\sigma^2_\beta}{\varepsilon^2_{n,2}}))$. Therefore, by taking $\varepsilon^2_{n,1} = N\tilde{\varepsilon}^2_1$ we see that

$$nT(\varepsilon^2_{n,1} + \sigma^2_\beta\varepsilon^2_{n,1}) + n\varepsilon^2_{n,2} \lesssim \varepsilon_{n,1}^{-d/\tau} + N\left(\frac{\varepsilon_{n,1}}{\sqrt{N}}\right)^{-d/\tau} + N\log\left(1 + \frac{N\sigma^2_\beta}{\varepsilon^2_{n,2}}\right)$$

should be satisfied. Moreover, by taking $\varepsilon_{n,1}^{(2\tau+d)/\tau} \asymp 1/nT$ and $\varepsilon^2_{n,2} \asymp N\log(1 + N^{-2s}/\varepsilon^2_{n,2})/n$ we can balance both sides. In particular, we may set

$$\varepsilon^2_{n,1} \asymp (nT)^{-\frac{2\tau}{2\tau+d}}, \quad \varepsilon^2_{n,2} \asymp \frac{N}{n} \wedge N^{-2s}.$$

Taking the balance with respect to N to maximize $\varepsilon^2_{n,2}$, we have $N \asymp n^{\frac{d}{2\alpha+d}}$ and $\varepsilon^2_{n,1} \asymp (nT)^{-\frac{2\tau}{2\tau+d}}$. Therefore, the minimax rate is lower bounded as

$$\delta^2_n \simeq (1 + \sigma^2_\beta)\varepsilon^2_{n,1} + \varepsilon^2_{n,2} \simeq n^{-\frac{2\alpha}{2\alpha+d}} + (nT)^{-\frac{2\tau}{2\tau+d}}. \tag{24}$$

F.4 Lower Bound in Piecewise γ-smooth Class

Suppose that we utilize the basis functions up to resolution K. Then, by the argument by Nishimura and Suzuki (2024), the number of basis functions ψ_r in the K-th resolution is $N' \asymp 2^{K/a^\dagger}$. Moreover, the δ_n-packing number of the γ-smooth class is also lower bounded by

$$\log M \geq N' \operatorname{polylog}(\delta_n, N'). \tag{25}$$

Here, by noticing the approximation error bound in Appendix D.2, we take $N' \asymp \delta_n^{-1/a^\dagger}$ where the basis functions are chosen from the Kth resolution. As in the case of the Besov space, we construct $(\psi^{(j)})_{j=1}^{M'}$ where $M' = 2^{N'/16}$ and $\psi^{(j)}(x) \in \mathbb{R}^N$ for $N = N'/2$ and $\|\psi^{(j)} - \psi^{(j')}\|_{L^2(P_{\mathcal{X}})}^2 \geq N/8$ for $j \neq j'$. Following the same argument as in the Besov case, we need to take $\varepsilon_{n,1}$ and $\varepsilon_{n,2}$ as

$$nT\sigma_\beta^2 \varepsilon_{n,1}^2 + n\varepsilon_{n,2}^2 \lesssim \delta_n^{-1/a^\dagger}$$

up to logarithmic factors. This is satisfied by taking $\varepsilon_{n,2}^2 = C\delta_n^2 \asymp n^{-\frac{2a^\dagger}{2a^\dagger+1}}$ with a constant C and balancing N so that $\sigma_\beta^2 \varepsilon_{n,1}^2 = (nT)^{-\frac{2a^\dagger}{2a^\dagger+1}} \sigma_\beta^{\frac{2}{2a^\dagger+1}} \asymp T^{-1} n^{-\frac{2a^\dagger}{2a^\dagger+1}}$. Combining this evaluation and (25) yields that the minimax lower bound is given by

$$\delta_n^2 \gtrsim n^{-\frac{2a^\dagger}{2a^\dagger+1}}. \tag{26}$$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes]

   Justification: See the Our Contributions paragraph for details.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Limitations paragraph.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Assumptions 1, 2, 3, 4, 5, 6. All proofs were provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix E, Figure 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: All experiments are toy simulations and data is i.i.d. random.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix E, Figure 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The median over 5 runs is reported in Figure 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: All experiments are toy simulations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research is theoretical and raises no ethical concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The research is theoretical and raises no concerns on societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.