Transformers are Minimax Optimal Nonparametric In-Context Learners

Juno Kim [1 2] **Tai Nakamaki** [1] **Taiji Suzuki** [1 2]

Abstract

We study the efficacy of in-context learning (ICL) from the viewpoint of statistical learning theory. We develop approximation and generalization analyses 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. In particular, 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, establishing upper and lower bounds w.r.t. both the number of tasks and in-context examples. These findings shed light on the effectiveness of few-shot prompting and the roles of task diversity and representation learning for ICL.

1. Introduction

In-context learning (ICL) refers to the ability of pretrained LLMs to perform a new task by being provided with a few examples within 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), while Raventos et al. (2023) show that sufficient pretraining task diversity is needed for the emergence of ICL. However, an understanding of the statistical foundations of ICL and few-shot prompting is still lacking. Existing theoretical analyses generally focus on a single attention layer and do not take into account the depth of transformers; see Appendix A for an overview of related works.

Our contributions. In this paper, we analyze the optimality of ICL from the perspective of statistical learning theory. We study a multi-layer 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. All proofs are deferred to the appendix.

- We develop a general framework for upper bounding the in-context estimation error of the empirical risk minimizer. We also derive an information-theoretic lower bound for the minimax risk in Appendix F.
- In the Besov space setting which includes the Hölder and Sobolev spaces, we show that **ICL with DNNs achieve 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 show that **ICL can improve upon the a priori optimal rate** when the task class basis resides in a coarser Besov space by learning informative representations, emphasizing the importance of pretraining on diverse tasks.
- We extend the optimality guarantees to nearly dimension-independent rates in the anisotropic Besov space and also to learning sequential data with transformers in the piecewise γ-smooth function class, further justifying the efficacy of ICL in extremely high-dimensional settings.

2. Problem Setup

2.1. Nonparametric Regression

We analyze ICL of a transformer pretrained on examples from a *family* of regression tasks, which we describe below. Let $\mathcal{X} \subseteq \mathbb{R}^d$ ($d \leq \infty$) be the input space endowed with a distribution $\mathcal{P}_\mathcal{X}$ and $(\psi_j^\circ)_{j=1}^\infty$ a discrete subset of $L^2(\mathcal{P}_\mathcal{X})$. A regression function $F_\beta^\circ : \mathcal{X} \to \mathbb{R}$ is generated for each task by sampling a coefficient sequence $\beta \in \mathbb{R}^\infty$ from a

[1] University of Tokyo, Tokyo, Japan [2] Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Juno Kim <junokim@g.ecc.u-tokyo.ac.jp>.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

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 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 output $\tilde{y}$ in the same manner.

We proceed to state our assumptions for the regression model. 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 the covariance matrices

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

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

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

Furthermore, $\operatorname{Tr}(\Sigma_{\bar{\beta},N})$ is bounded for all N and

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

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.

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, there exist $\phi_{\underline{N}}^*, \cdots, \phi_{\bar{N}}^* \in \mathcal{F}_N$ so that $\|\psi_j^\circ - \phi_j^*\|_{L^\infty(\mathcal{P}_\mathcal{X})} \le \delta_N$, $\underline{N} \le j \le \bar{N}$ for some $\delta_N > 0$.*

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 $\mathcal{S}_N \subset \mathbb{R}^{N\times N}$,

$$\check{f}_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}) = \frac{1}{n}\sum_{k=1}^n y_k \phi(x_k)^\top \Gamma^\top \phi(\tilde{x}) = \left\langle \frac{\Gamma\phi(\mathbf{X})\boldsymbol{y}}{n}, \phi(\tilde{x}) \right\rangle$$

where $\Theta = (\Gamma, \phi) \in \mathcal{S}_N \times \mathcal{F}_N$. Finally, the output is clipped by applying $\operatorname{clip}_{\bar{B}}(u) := \max\{\min\{u, \bar{B}\}, -\bar{B}\}$, yielding $f_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}) := \operatorname{clip}_{\bar{B}}(\check{f}_\Theta(\mathbf{X}, \boldsymbol{y}, \tilde{x}))$. Here, 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 reparametrization of linear attention widely used in theoretical analyses (Zhang et al., 2023; Wu et al., 2024, see e.g.), where the query and key matrices are consolidated into one matrix Γ. The form 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 $F_{\beta^{(t)}}^\circ \in \mathcal{F}^\circ$, $\beta^{(t)} \sim \mathcal{P}_\beta$ 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 in Section 2.1, and trained to minimize the empirical risk:

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

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

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

[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.

3. Risk Bounds for In-Context Learning

In this section, we outline our framework for upper bounding 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 $\mathcal{C}$.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 $\rho(c_i, c_j) \geq \delta$ for all $i \neq j$. The metric is omitted if $\rho = \|\cdot\|_{L^\infty}$. The transformer 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)**.** *For any $\epsilon > 0$ such that $\mathcal{V}(\mathcal{T}_N, \epsilon) \geq 1$,*

$$\bar{R}(\widehat{\Theta}) \lesssim \inf_{\Theta} R(\Theta) + \frac{B^2 + \sigma^2}{T} \mathcal{V}(\mathcal{T}_N, \epsilon) + (B + \sigma)\epsilon.$$

See Appendix B.1 for the proof. Here, the second term is the *pretraining* generalization error w.r.t. number of tasks T; the *in-context* generalization error w.r.t. prompt length n manifests as part of the first term. This separation allows us to compare the difficulty of the two types of learning.

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

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

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 B.

Bounding generalization error. To estimate the metric entropy of $\mathcal{T}_N$, we first reduce to the representation class $\mathcal{F}_N$. The proof is given in Appendix C.2.

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

$$\mathcal{V}(\mathcal{T}_N, \epsilon) \lesssim N^2 \log \frac{B_N'^2}{\epsilon} + \mathcal{V}\left(\mathcal{F}_N, \frac{\epsilon}{D B_N' \sqrt{N}} \right).$$

4. Minimax Optimality of In-Context Learning

4.1. Besov Space and DNNs

In this section, we apply our theory to the sample complexity of ICL when $\mathcal{F}_N$ consists of (clipped, see (6)) DNNs. We define the set of DNNs with depth L, width W, sparsity S, norm bound M and ReLU activation $\eta(x) = x \vee 0$ as

$$\mathcal{F}_{\mathrm{DNN}}(L, W, S, M) = \{(\mathbf{W}^{(L)}\eta + b^{(L)}) \circ \cdots \circ (\mathbf{W}^{(1)}x + b^{(1)}) \mid \mathbf{W}^{(\ell)} \in \mathbb{R}^{W \times W}, \mathbf{W}^{(L)} \in \mathbb{R}^W, b^{(\ell)} \in \mathbb{R}^W, b^{(L)} \in \mathbb{R}, \textstyle\sum_{\ell=1}^L \|\mathbf{W}^{(\ell)}\|_0 + \|b^{(\ell)}\|_0 \leq S, \|\mathbf{W}^{(\ell)}\|_\infty \vee \|b^{(\ell)}\|_\infty \leq M\}.$$

The *Besov space* is a very general function class including the Hölder and Sobolev spaces (see Appendix D.1.1) which captures spatial inhomogeneity in smoothness, and provides a natural setting in which to study the expressive power of DNNs (Suzuki, 2019). We fix $\mathcal{X} = [0, 1]^d$ for simplicity.

Definition 4.1 (Besov space)**.** For $2 \leq p \leq \infty, 0 < q \leq \infty$, smoothness $\alpha > 0$ and $r = \lfloor \alpha \rfloor + 1$, the rth modulus of $f \in L^p(\mathcal{X})$ is $w_{r,p}(f, t) := \sup_{\|h\|_2 \leq t} \|\Delta_h^r(f)\|_p$ where

$$\Delta_h^r(f)(x) = 1_{\{x, x+rh \in \mathcal{X}\}} \textstyle\sum_{j=0}^r \binom{r}{j} (-1)^{r-j} f(x + jh).$$

The Besov space is defined as $B_{p,q}^\alpha(\mathcal{X}) = \{f \in L^p(\mathcal{X}) \mid \|f\|_{B_{p,q}^\alpha} := \|f\|_{L^p} + |f|_{B_{p,q}^\alpha} < \infty\}$ where

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

$\mathbb{U}(B_{p,q}^\alpha(\mathcal{X}))$ denotes the unit ball in $(B_{p,q}^\alpha(\mathcal{X}), \|\cdot\|_{B_{p,q}^\alpha})$.

A natural basis system for $B_{p,q}^\alpha(\mathcal{X})$ is formed by the *B-splines*, a type of wavelet decomposition or multiresolution analysis (DeVore & Popov, 1988).

Definition 4.2. The tensor product B-spline of order $m \in \mathbb{N}$, $m > \alpha + 1 - 1/p$, resolution $k \in \mathbb{Z}_{\geq 0}^d$ and location $\ell \in I_k^d = \prod_{i=1}^d [-m : 2^{k_i}]$ is $\omega_{k,\ell}^d(x) = \prod_{i=1}^d \iota_m(2^{k_i} x_i - \ell_i)$ where $\iota_m(x)$ is the $(m+1)$-fold convolution of $1_{\{x \in [0,1]\}}$. We abuse notation and write $\omega_{k,\ell}^d = \omega_{(k,\cdots,k),\ell}^d$ for $k \in \mathbb{Z}_{\geq 0}$.

The difficulty of learning a regression function in the Besov space is quantified by the following minimax risk.

Proposition 4.3 (Donoho & 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*

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

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 $\{\psi_j^\circ \mid j \in \mathbb{N}\} =$

$\{2^{kd/2}\omega^d_{k,\ell} \mid k \in \mathbb{Z}_{\geq 0}, \ell \in I^d_k\}$ the set of all B-spline wavelets ordered 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 $2^{kd/2}\omega^d_{k,\ell}$. 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 with a certain decay (Proposition D.11). In Appendix D.1.1, we verify Assumptions 1 and 2 with $r = 1/2, s = \alpha/d$ under:

Assumption 4. *$\mathcal{F}^\circ = \mathbb{U}(B^\alpha_{p,q}(\mathcal{X}))$, $\alpha > d/p$ and $\mathcal{P}_\mathcal{X}$ has positive Lebesgue density $\rho_\mathcal{X}$ bounded above and below on $\mathcal{X}$. Also, for all $k \geq 0$ and $\ell \in I^d_k$, $\beta_{k,\ell}$ are independent and*

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

We obtain the following entropy bound in Appendix D.1.2.

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

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

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 entropy is bounded as $\mathcal{V}(\mathcal{F}_N, \|\cdot\|_{L^\infty}, \epsilon) \lesssim N \log \frac{N}{\delta_N \epsilon}$.

Combining these results, we conclude in Appendix D.1.3:

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

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

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

The first term is equal to the N-term optimal error (Dũng, 2011a). For the generalization gaps, $n = \widetilde{\Omega}(N)$ is enough to learn the basis expansion in context while $T = \widetilde{\Omega}(N^2)$ is needed 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 examples.

A limitation of ICL. While DNNs can adapt to task smoothness in supervised settings (Suzuki, 2019), ICL and other meta-learning schemes are fundamentally constrained to non-adaptive approximation since they cannot update at inference time, and hence are lower bounded by the *Kolmogorov width* $N^{-\frac{\alpha}{d}+(\frac{1}{p}-\frac{1}{2})_+}$ (Vybíral, 2008) which is strictly worse than the optimal rate in the regime $1 \leq p < 2$.

Avoiding the curse of dimensionality. The above optimal rate inevitably suffers from the curse of dimensionality as d appears in the exponent. In Appendix D.2, we extend Theorem 4.5 to the *anisotropic Besov space* (Nikol'skii, 1975), a generalization to differing degrees of smoothness $(\alpha_1, \cdots, \alpha_d)$ in each coordinate. Then the ICL rate only depends on the quantity $\widetilde{\alpha} := (\sum_i \alpha_i^{-1})^{-1}$, and becomes independent of d if only a few coordinates are important.

Sequential input and transformers. In Appendix E, we also study a more complex setting where inputs $x \in [0,1]^{d\times\infty}$ are unbounded *sequences* of tokens (e.g. whole documents) and ϕ is itself a deep transformer network. Here, $\mathcal{F}^\circ$ is taken to be the *piecewise γ-smooth class* (Takakura & Suzuki, 2023) where positional smoothness may vary depending on the input. Even in this infinite-dimensional setting, we show that ICL still achieves near-optimal risk, further justifying the efficacy of ICL for sequential data.

4.3. Learning a Coarser Basis

As another application of our framework, we illustrate how pretraining can actively *mitigate* the complexity of ICL. Consider the case where $(\psi^\circ_j)^\infty_{j=1}$ is no longer the B-spline basis of $B^\alpha_{p,q}(\mathcal{X})$ but instead is chosen from some wider function space, say the unit ball of $B^\tau_{p,q}(\mathcal{X})$ for a smaller smoothness $\tau < \alpha$. Without knowledge of the basis, the sample complexity of regression is a priori lower bounded by the minimax rate $n^{-\frac{2\tau}{2\tau+d}}$ by Proposition 4.3. With ICL, however, we obtain the modified risk bound:

Corollary 4.6 (ICL for coarser basis)**.** *Suppose $\alpha > \tau > d/p$, the basis $(\psi^\circ_j)^\infty_{j=1} \subset \mathbb{U}(B^\tau_{p,q}(\mathcal{X}))$ and Assumptions 1, 2 hold with $r = 1/2$, $s = \alpha/d$. Then if $n \gtrsim N \log N$,*

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

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 gap is now dominated by the higher entropy of the DNN class and strictly worse than Theorem 4.5. Nonetheless, observe that the burden is entirely carried by T; with sufficient pretraining, the third term can be made arbitrarily small and the 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 the coarser basis during pretraining.

5. Conclusion

We analyzed ICL of a transformer consisting of a DNN and a linear attention layer pretrained on nonparametric regression tasks from the perspective of statistical learning theory. We proved that ICL achieves nearly minimax optimal risk in the (anisotropic) Besov space and γ-smooth class. We also showed that ICL can improve upon the a priori optimal rate by learning informative representations. Our work opens up new ways of understanding emergent properties of LLMs.

Acknowledgments

JK was partially supported by JST CREST (JPMJCR2015). TS was partially supported by JSPS KAKENHI (20H00576) and JST CREST (JPMJCR2115).

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mahankali, A., Hashimoto, T. B., and Ma, T. 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.

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

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

Nishimura, Y. and Suzuki, T. 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.

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

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

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

Suzuki, T. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zhang, R., Wu, J., and Bartlett, P. L. 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. Related Works

Theory of ICL. 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.

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 & 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.

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 & Nitanda, 2021), convolutional neural networks for infinite-dimensional input (Okumoto & Suzuki, 2022), and transformers for sequence-to-sequence functions (Takakura & Suzuki, 2023).

B. Proof of Proposition 3.2

B.1. Proof of Theorem 3.1

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 C.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 $f^\circ(z) = F_\beta^\circ(\tilde{x})$ 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. □

B.2. 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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x}) - \phi^*(\tilde{x})^\top \frac{\Gamma^*\,\Phi^*\boldsymbol{y}}{n}\right)^2\right]\\
&\lesssim \mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x}) - \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}\bar{\beta}_{\underline{N}:\bar{N}}}{n}\right)^2\right] \qquad (7)\\
&\quad + \mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top \Gamma^* \frac{\Psi^\circ_{\underline{N}:\bar{N}}(\Psi^{\circ\top}_{\bar{N}:\infty}\beta_{\bar{N}:\infty} + \xi)}{n}\right)^2\right] \qquad (8)\\
&\quad + \mathbb{E}\left[\left((\phi^*(\tilde{x}) - \psi^\circ_{\underline{N}:\bar{N}}(\tilde{x}))^\top \Gamma^* \frac{(\Phi^* - \Psi^\circ_{\underline{N}:\bar{N}})\boldsymbol{y}}{n}\right)^2\right] \qquad (9)
\end{aligned}$$

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

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

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

B.3. Bounding Term (7)

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

$$\begin{aligned}
&\mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}\bar{\beta}_{\underline{N}:\bar{N}}}{n}\right)^2\right]\\
&=\mathbb{E}\left[\left(F^\circ_{\beta,\bar{N}}(\tilde{x})-\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\left(\frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-(\Gamma^*)^{-1}+(\Gamma^*)^{-1}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^2\right]\\
&=\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\left(\frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-\Sigma_{\Psi,N}-\frac{1}{n}\Sigma^{-1}_{\bar{\beta},N}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^2\right]\\
&\leq 2\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\left(\frac{\Psi^\circ_{\underline{N}:\bar{N}}\Psi^{\circ\top}_{\underline{N}:\bar{N}}}{n}-\Sigma_{\Psi,N}\right)\bar{\beta}_{\underline{N}:\bar{N}}\right)^2\right] &(12)\\
&\qquad+2\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Sigma^{-1}_{\bar{\beta},N}}{n}\bar{\beta}_{\underline{N}:\bar{N}}\right)^2\right]. &(13)
\end{aligned}$$

Denote the operator norm (with respect to L^2 norm of vectors) by $\|\cdot\|_{\mathrm{op}}$. For (12), 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}\Bigg[\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]\\
&\qquad\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}}\Bigg]\\
&=\mathbb{E}_{\mathbf{X}}\Bigg[\mathrm{Tr}\Bigg[\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}\\
&\qquad\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}\Bigg]\Bigg]\\
&\lesssim\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)^2\Sigma^{1/2}_{\Psi,N}\Sigma_{\bar{\beta},N}\Sigma^{1/2}_{\Psi,N}\right]\right]
\end{aligned}$$

due to the independence of $\mathbf{X},\tilde{x}$ and β. For the last inequality, we have used the fact that both the $\Sigma^{1/2}_{\Psi,N}\Gamma^*\Sigma_{\Psi,N}\Gamma^*\Sigma^{1/2}_{\Psi,N}$ 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 B.6:

Lemma B.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 (12) is bounded as

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

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

Moreover for (13), we compute

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

B.4. Bounding Term (8)

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

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\Psi^\circ_{\underline{N}:\bar{N}}(\Psi^{\circ\top}_{\bar{N}:\infty}\beta_{\bar{N}:\infty}+\xi)}{n}\right)^2\right]\\
&=\mathbb{E}\left[\left(\sum_{i=1}^n\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\frac{\psi^\circ_{\underline{N}:\bar{N}}(x_i)(\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x_i)+\xi_i)}{n}\right)^2\right]\\
&=\frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x_i)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x_i)+\sum_{i=1}^n\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x_i)\xi_i\right)^2\right]
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{n}\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\right)^2\right] && (14)\\
&\quad + \frac{n-1}{n}\mathbb{E}\left[\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x)\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x)\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x')\beta^\top_{\bar{N}:\infty}\psi^\circ_{\bar{N}:\infty}(x')\right] && (15)\\
&\quad + \frac{\sigma^2}{n}\mathbb{E}\left[\left(\psi^\circ_{\underline{N}:\bar{N}}(\tilde{x})^\top\Gamma^*\psi^\circ_{\underline{N}:\bar{N}}(x)\right)^2\right] && (16)
\end{aligned}$$

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

For (14), 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 (15), 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 (16), we have that

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

$$\begin{aligned}
&\leq \frac{\sigma^2}{n}\operatorname{Tr}\left[\Gamma^*\Sigma_{\Psi,N}\Gamma^*\left(\Sigma_{\Psi,N}+\frac{1}{n}\Sigma_{\bar{\beta},N}^{-1}\right)\right]\\
&= \frac{\sigma^2}{n}\operatorname{Tr}\left[\Gamma^*\Sigma_{\Psi,N}\right] \leq \frac{\sigma^2 N}{n}.
\end{aligned}$$

Therefore, we obtain the following bound:

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

B.5. Bounding Terms (9)-(11)

For (9), we use the Cauchy-Schwarz inequality and Assumption 3 to bound

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

Similarly for (10) and (11), we have

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

and

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

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

This concludes the proof of Proposition 3.2.

B.6. Proof of Lemma B.1

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

Theorem B.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 B.3. *For matrices $\mathbf{S}_1,\cdots,\mathbf{S}_n$ satisfying the conditions of Theorem B.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 B.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

$$
\left\|\mathbb{E}_x\left[\Sigma_{\Psi,N}^{-1/2}\psi^\circ_{\underline{N}:\bar{N}}(x)\psi^\circ_{\underline{N}:\bar{N}}(x)^\top\Sigma_{\Psi,N}^{-1}\psi^\circ_{\underline{N}:\bar{N}}(x)\psi^\circ_{\underline{N}:\bar{N}}(x)^\top\Sigma_{\Psi,N}^{-1/2}\right]\right\|_{\mathrm{op}}
$$

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

again by Assumption 1.

Hence we may apply Corollary B.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 B.3. From Theorem B.2 and with the change of variables $\lambda = t^2/n^2$, we have

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

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

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

and the expectation can be controlled as

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

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

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

as was to be shown. □

C. Proofs on Metric Entropy

C.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 & 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. □

C.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

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

Then it holds that

$$|f_{\Theta_1}(\mathbf{X}, \boldsymbol{y}, \tilde{x}) - f_{\Theta_2}(\mathbf{X}, \boldsymbol{y}, \tilde{x})|$$

$$\begin{aligned}
&\le |\check{f}_{\Theta_1}(\mathbf{X},\boldsymbol{y},\tilde{x}) - \check{f}_{\Theta_2}(\mathbf{X},\boldsymbol{y},\tilde{x})| \\
&= \left| \left\langle \frac{\Gamma_1\phi_1(\mathbf{X})\boldsymbol{y}}{n}, \phi_1(\tilde{x}) \right\rangle - \left\langle \frac{\Gamma_2\phi_2(\mathbf{X})\boldsymbol{y}}{n}, \phi_2(\tilde{x}) \right\rangle \right| \\
&= \frac{1}{n}\Big| \phi_1(\tilde{x})^\top \Gamma_1\phi_1(\mathbf{X})\boldsymbol{y} - \phi_2(\tilde{x})^\top \Gamma_1\phi_1(\mathbf{X})\boldsymbol{y} + \phi_2(\tilde{x})^\top \Gamma_1\phi_1(\mathbf{X})\boldsymbol{y} \\
&\qquad - \phi_2(\tilde{x})^\top \Gamma_2\phi_1(\mathbf{X})\boldsymbol{y} + \phi_2(\tilde{x})^\top \Gamma_2\phi_1(\mathbf{X})\boldsymbol{y} - \phi_2(\tilde{x})^\top \Gamma_2\phi_2(\mathbf{X})\boldsymbol{y} \Big| \\
&\le \frac{1}{n}\|\phi_1(\tilde{x}) - \phi_2(\tilde{x})\| \|\Gamma_1\|_{\mathrm{op}} \|\phi_1(\mathbf{X})\boldsymbol{y}\| + \frac{1}{n}\|\phi_2(\tilde{x})\| \|\Gamma_1 - \Gamma_2\|_{\mathrm{op}} \|\phi_1(\mathbf{X})\boldsymbol{y}\| \\
&\qquad + \frac{1}{n}\|\phi_2(\tilde{x})\| \|\Gamma_2\|_{\mathrm{op}} \|(\phi_1(\mathbf{X}) - \phi_2(\mathbf{X}))\boldsymbol{y}\| \\
&\le \left(\frac{\sqrt{N}\delta_2 C_2}{n} + \frac{B'_N\delta_1}{n} \right) \sum_{i=1}^n \|\phi_1(x_i)\| |y_i| + \frac{B'_N C_2}{n} \sum_{i=1}^n \|\phi_1(x_i) - \phi_2(x_i)\| |y_i| \\
&\le (\sqrt{N}\delta_2 C_2 + B'_N\delta_1) B'_N (B+\sigma) + B'_N C_2 \sqrt{N}\delta_2 (B+\sigma) \\
&= B'^2_N (B+\sigma)\delta_1 + 2B'_N(B+\sigma)C_2\sqrt{N}\delta_2.
\end{aligned}$$

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

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

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 C.3.

Lemma C.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

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

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. □

C.3. Proof of Lemma C.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 C.2 (Szarek (1981), Proposition 6)**.** *There exist universal constants $c_1, c_2 > 0$ such that for all $N \in \mathbb{N}$ and $\delta \in (0, 2]$,*

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

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. □

D. Details on Besov-type Spaces

D.1. Besov Space

D.1.1. Verification of Assumptions

It is known that the Hölder space $C^\alpha(\mathcal{X}) = B^\alpha_{\infty,\infty}(\mathcal{X})$ for $\alpha > 0, \alpha \notin \mathbb{N}$ and the Sobolev space $W^m_2(\mathcal{X}) = B^m_{2,2}(\mathcal{X})$ for $m \in \mathbb{N}$ as well as the embeddings $B^m_{p,1}(\mathcal{X}) \hookrightarrow W^m_p(\mathcal{X}) \hookrightarrow B^m_{p,\infty}(\mathcal{X})$; if $\alpha > d/p$, $B^\alpha_{p,q}(\mathcal{X})$ compactly embeds into the space of continuous functions on $\mathcal{X}$. See Triebel (1983); Giné & Nickl (2015) for more details.

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 D.1 (DeVore & 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*

$$f = \sum_{k=0}^\infty \sum_{\ell \in I^d_k} \tilde{\beta}_{k,\ell} \omega^d_{k,\ell}$$

such that the coefficients satisfy

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

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.

We can now check that we have not given ourselves an easier learning problem with (5). In particular, Lemma D.1 implies that for any $f \in \mathbb{U}(B^\alpha_{p,q}(\mathcal{X}))$ the p-norm average of $\tilde{\beta}_{k,\ell}$ for $\ell \in I^d_k$ at resolution k is bounded as

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

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

$$\left(\frac{1}{|I^d_k|} \sum_{\ell \in I^d_k} |\beta_{k,\ell}|^p\right)^{1/p} \lesssim (2^{kd})^{-\alpha/d - 1/2}. \tag{17}$$

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 (17). This matches Assumption 4 up to the logarithmic factor and hence the rate is nearly tight in variance, even though (17) 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 $(\omega_{K,\ell}^d)_{\ell\in I_K^d}$ at each resolution K is linearly independent; for any $g\in L^2(\mathcal{X})$ that can be expressed as

$$g=\sum_{\ell\in I_K^d}\beta_{K,\ell}2^{Kd/2}\omega_{K,\ell}^d$$

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

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

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 $(\omega_{K,\ell}^d)_{\ell\in I_K^d}$ by repeatedly applying the following relation.

Lemma D.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*

$$\omega_{k,\ell}^d=\sum_{r_1,\cdots,r_d=0}^m 2^{(-m+1)d}\prod_{i=1}^d\binom{m}{r_i}\cdot\omega_{k+1,2\ell+r-\frac{m}{2}\mathbf{1}}^d. \tag{18}$$

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

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

as was to be shown. □

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

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

and

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

It is straightforward to see that $0 \leq \omega_{k,\ell}^d(x) \leq 1$ 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:

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

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

$$\sum_{j=\underline{N}}^{\bar{N}} \psi_j^\circ(x)^2 = \sum_{\ell \in I_K^d} 2^{Kd} \omega_{K,\ell}^d(x)^2 \leq 2^{Kd} \sum_{\ell \in \mathbb{Z}^d} \omega_{K,\ell}^d(x) = 2^{Kd} \lesssim N,$$

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

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

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

Furthermore, the convergence rate of the truncated approximation $F_{\beta,\bar{N}}^\circ$ is determined by the decay rate of β in Lemma D.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 (17). 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 D.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 & 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 D.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 D.11, stated and proved in Appendix D.3 below, by restricting to wavelets with uniform resolution across dimensions. Using Corollary D.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

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

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

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

D.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 D.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*

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

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

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

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

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

□

D.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. □

Remark D.7. We can see that 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 required task diversity may be lessened by considering low-rank structure or approximation of attention heads (Bhojanapalli et al., 2020; Chen et al., 2021).

D.2. Anisotropic Besov Space

D.2.1. Definitions and Results

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

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

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

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

For the anisotropic Besov space, we need to redefine the wavelet basis so that the sensitivity to resolution $k \in \mathbb{Z}_{\ge 0}$ 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 $k \ge 0$ 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

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

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 & Nitanda, 2021, Theorem 4). Our result for in-context learning is as follows.

Theorem D.8 (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*

$$\mathbb{E}_\beta[\beta_{k,\ell}] = 0, \quad \mathbb{E}_\beta[\beta_{k,\ell}^2] \lesssim 2^{-k\underline{\alpha}(2+1/\widetilde{\alpha})} k^{-2}, \quad \forall k \ge 0, \; \ell \in I_k^{d,\alpha}. \tag{19}$$

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.

D.2.2. Proof of Theorem D.8

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

Lemma D.9 (Suzuki & 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 D.2. See Proposition D.11 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,

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

and the scaled coefficients decay in average as

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

For Assumption 2, we can check that

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

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

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

Thus we may set $r = 1/2, s = \widetilde{\alpha}$ and take the variance decay rate (19) as

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

For boundedness of $\mathrm{Tr}(\Sigma_{\bar{\beta},N})$, we use the following result which is also obtained from Proposition D.11 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 D.10. *For any $0 \leq k < k'$ there exists constants $\gamma_{k,k',\ell,\ell'} \geq 0$ for $\ell \in I_k^{d,\alpha}$, $\ell' \in I_{k'}^{d,\alpha}$ such that*

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

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

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

We apply Corollary D.10 to refine all components of $F_{\beta,\bar{N}}^\circ$ to resolution K:

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

Hence it follows that

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

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. □

D.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 D.11. *For any $k,k'\in\mathbb{Z}^d_{\geq 0}$ such that $k'-k\in\mathbb{Z}^d_{\geq 0}$ there exists constants $\gamma_{k,k',\ell,\ell'}\geq 0$ for $\ell\in I^d_k$, $\ell'\in I^d_{k'}$ such that*

$$\sum_{\ell\in I^d_k}\beta_{k,\ell}2^{\|k\|_1/2}\omega^d_{k,\ell}=\sum_{\ell'\in I^d_{k'}}\bar{\beta}_{k',\ell'}2^{\|k'\|_1/2}\omega^d_{k',\ell'},\quad \bar{\beta}_{k',\ell'}=\sum_{\ell\in I^d_k}\gamma_{k,k',\ell,\ell'}\beta_{k,\ell} \tag{20}$$

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

$$\sum_{\ell\in I^d_k}\gamma_{k,k',\ell,\ell'}\leq 2^{-\|k'-k\|_1/2},\qquad \sum_{\ell'\in I^d_{k'}}\gamma_{k,k',\ell,\ell'}\leq 2^{\|k'-k\|_1/2}.$$

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

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

Since $\ell+(\ell_j+r-\frac{m}{2})e_j$ matches a given location vector $\ell'\in I^d_{k+e_j}$ if and only if $\ell_i=\ell'_i$ $(i\neq j)$ and $\ell'_j=2\ell_j+r-\frac{m}{2}$, comparing coefficients in (20) yields

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

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

$$\begin{aligned}\sum_{\ell\in I^d_k}\gamma_{k,k+e_j,\ell,\ell'}&\leq\sum_{\ell_j\in\mathbb{Z}}2^{-m+1/2}\binom{m}{\ell'_j-2\ell_j+\frac{m}{2}}\leq 2^{-1/2},\\ \sum_{\ell'\in I^d_{k+e_j}}\gamma_{k,k+e_j,\ell,\ell'}&\leq\sum_{\ell'_j\in\mathbb{Z}}2^{-m+1/2}\binom{m}{\ell'_j-2\ell_j+\frac{m}{2}}\leq 2^{1/2},\end{aligned}$$

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

$$\sum_{\ell\in I^d_k}\beta_{k,\ell}2^{\|k\|_1/2}\omega^d_{k,\ell}=\sum_{\ell'\in I^d_{k'}}\bar{\beta}_{k',\ell'}2^{\|k'\|_1/2}\omega^d_{k',\ell'}=\sum_{\ell''\in I^d_{k'+1}}\bar{\bar{\beta}}_{k'+1,\ell''}2^{(\|k'\|_1+1)/2}\omega^d_{k'+e_j,\ell''}$$

where

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

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

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

and furthermore

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

This concludes the proof. □

D.4. Proof of Corollary 4.6

In order to approximate arbitrary $\psi^\circ_j\in\mathbb{U}(B^\tau_{p,q}(\mathcal{X}))$, we need the following construction instead of Lemma D.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 D.12 (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*

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

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 D.6, after some algebra we obtain the following counterpart to Lemma 4.4.

Corollary D.13. *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. □

E. Sequential Input and Transformers

E.1. Definitions and Results

We now consider a more involved setting where the inputs $x \in [0,1]^{d\times\infty}$ are bidirectional sequences of tokens and ϕ is itself a transformer network.[2] 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 & Suzuki (2023), which allows for arbitrary bounded permutations among input tokens. They show in such a setting that transformer networks still achieve near-optimal sample complexity due to their parameter sharing and dynamical feature extraction capabilities; we extend this guarantee to ICL of trained transformers.

γ-smooth class. We first define the γ-smooth function class introduced by Okumoto & 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 $p \ge 2, q \ge 1$, 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 $\alpha \in \mathbb{R}^{d\times\infty}_{>0}$, 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.

[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 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.

- 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 & 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

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

where

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

Transformer network. 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]

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

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

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

Next, we state the set of assumptions inherited from Takakura & 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* $C_\mu, \varrho > 0$*.*

[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.

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,*

$$\mathbb{E}_\beta[\beta_r] = 0, \quad \mathbb{E}_\beta[\beta_r^2] \lesssim 2^{-(2+1/\alpha^\dagger)\gamma(s)}\gamma(s)^{-2}. \tag{21}$$

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 E.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*

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

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

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.

E.2. Proof of Theorem E.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 E.2 (Takakura & 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 E.3 (Takakura & Suzuki (2023), Theorem 5.3)**.** *For $\epsilon > 0$ and $B \ge 1$ it holds that*

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

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 E.4 (Okumoto & Suzuki (2022), Lemma 17)**.** *For $1 \le q \le 2$ 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 & 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 (21) is justified. Moreover for large K,

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

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 E.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 & 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 E.2 and E.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. □

F. Minimax Lower Bounds for ICL

In this section, we provide a scheme for obtaining lower bounds for the minimax rate in the ICL setup by extending the theory of Yang & Barron (1999). The result is purely information-theoretic and hence applies to not just ICL but any meta-learning scheme for the regression problem of Section 2.1 depending on the data $\mathcal{D}_{n,T} = \{(\mathbf{X}^{(t)}, \boldsymbol{y}^{(t)}, \tilde{x}^{(t)})\}_{t=1}^{T+1}$, 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. We also suppose the support of $\mathcal{P}_\beta$ is included in $\mathcal{B} := \{\beta \in \mathbb{R}^\infty \mid |\beta_j| \lesssim j^{-2s-1}(\log j)^{-2},\ j \in \mathbb{N}\}$ 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 F.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:*

$$\begin{gathered}\tfrac{1}{2\sigma^2}\left(n(T+1)\sigma_\beta^2\varepsilon_{n,1}^2 + C_2 n\varepsilon_{n,2}^2\right) \le \log Q_1 + \log Q_2, \\ 8(\log Q_1 + \log Q_2) \le \log M, \quad 4\log 2 \le \log M.\end{gathered} \tag{22}$$

Then the minimax rate is lower bounded as

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

Finally, Proposition F.1 is applied to obtain concrete lower bounds for the settings analyzed above throughout Appendices F.2-F.4. We prove the following rates: $n^{-\frac{2\alpha}{2\alpha+d}}$ for the Besov space, $n^{-\frac{2\alpha}{2\alpha+d}}+(nT)^{-\frac{2\tau}{2\tau+d}}$ for the coarser basis setting, and $n^{-\frac{2\alpha^\dagger}{2\alpha^\dagger+1}}$ for the piecewise γ-smooth class. In particular, we can check that insufficient pretraining indeed leads to a worse complexity for the second setting, while large T retrieves the faster rate $n^{-\frac{2\alpha}{2\alpha+d}}$. These results verify the optimality of our previous results from the perspective of the sample complexity of meta-learning.

F.1. Proof of Proposition F.1

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 & Barron, 1999).

Proposition F.2. *Let Θ be a random variable uniformly distributed over $\{(\psi^{(j)},\beta_{T+1}^{(j)})\}_{j=1}^M$. Then, it holds that*

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

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. 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

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

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 & 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 $\tilde{P}_{\boldsymbol{y}^{(1:T+1)}}=\frac{1}{Q}\sum_{j=1}^Q P_{\boldsymbol{y}^{(1:T+1)}|\tilde{\theta}^{(j)}}$. If we let

$$\kappa(j):=\arg\min_{1\leq k\leq Q}D(P_{\boldsymbol{y}^{(1:T+1)}|\theta^{(j)}}\|P_{\boldsymbol{y}^{(1:T+1)}|\tilde{\theta}^{(k)}}),$$

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

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

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

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

In the same manner, we have that

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

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

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

Therefore, the expected mutual information can be bounded as

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

Applying Proposition F.2 together with (22) 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 $\Theta(\delta_n^{-1/s})$ 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 $Q_1 \lesssim 2^N$), 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 $\beta_j \in [-\sigma_\beta, \sigma_\beta]$. 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 $\varepsilon_{n,1}^{(2\tau+d)/\tau} = 1/nT$ and $\varepsilon^2_{n,2} = N\log(1 + N^{-2s}/\varepsilon^2_{n,2})/n$ we can balance both sides. In particular, we may set

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

Taking the balance with respect to N to maximize $\varepsilon^2_{n,2}$, we have $N = n^{\frac{d}{2\alpha+d}}$ and $\varepsilon^2_{n,1} = (nT)^{-\frac{2\tau}{2\tau+d}}$. 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 & 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 E.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

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