Optimal In-context Adaptivity and Distributional Robustness of Transformers

Tianyi Ma
Statistical Laboratory
University of Cambridge
tm681@cam.ac.uk

Tengyao Wang
Department of Statistics
London School of Economics and Political Science
t.wang59@lse.ac.uk

Richard J. Samworth
Statistical Laboratory
University of Cambridge
r.samworth@statslab.cam.ac.uk

Abstract

We study in-context learning problems where a Transformer is pretrained on tasks drawn from a mixture distribution $\pi = \sum_{\alpha \in \mathcal{A}} \lambda_\alpha \pi_\alpha$, called the pretraining prior, in which each mixture component π_α is a distribution on tasks of a specific difficulty level indexed by α. Our goal is to understand the performance of the pretrained Transformer when evaluated on a different test distribution μ, consisting of tasks of difficulty $\beta \in \mathcal{A}$, and with potential distribution shift relative to π_β. In particular, we consider nonparametric regression problems with random smoothness, and multi-index models with both random smoothness and random effective dimension. We prove that a Transformer pretrained on such mixture distributions can adapt to the difficulty of a new task in-context and achieve the optimal prediction rate, uniformly over test distributions in a chi-squared divergence ball. Thus, the pretrained Transformer is able to achieve faster rates of convergence on easier tasks and is robust to distribution shift at test time. Finally, we prove that even if an estimator had access to the test distribution μ, the convergence rate of its expected risk over μ could not be faster than that of our pretrained Transformers, thereby providing a more appropriate optimality guarantee than minimax lower bounds.

1 Introduction

Transformers (Vaswani et al., 2017) have emerged as one of the dominant architectures in modern machine learning, as they have achieved state-of-the-art performance in many domains such as natural language processing (Devlin et al., 2019; Brown et al., 2020), computer vision (Dosovitskiy et al., 2021; Liu et al., 2021) and protein prediction (Jumper et al., 2021). A striking ability of pretrained Transformers, first observed by Brown et al. (2020) in Large Language Models (LLMs), is the phenomenon of *in-context learning* (ICL): given a prompt (context) containing examples and a query, a Transformer can learn the underlying pattern from the examples and produce accurate output for the query, without updating its parameters. The ICL ability of Transformers is recently exploited by Hollmann et al. (2025) to build tabular foundation models, which they show to outperform many machine learning methods, such as tree-based and boosting algorithms, on regression and classification problems for tabular data.

Extensive studies on ICL have been conducted from various perspectives (e.g. Garg et al., 2022; Xie et al., 2022; von Oswald et al., 2023; Bai et al., 2023; Kim, Nakamaki and Suzuki, 2024); we refer the readers to Section 1.2 for a more detailed literature review. Most theoretical analyses of ICL

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

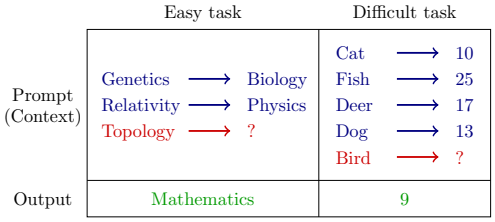


Figure 1: Two ICL tasks. Blue indicates the (in-context) examples, red indicates the queries, and green indicates the outputs.

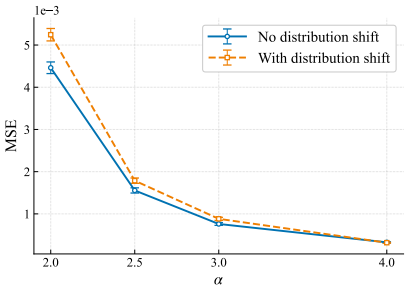


Figure 2: Test-time MSE against the smoothness α of the regression function.

have focused on the setting where the distribution of pretraining data (which we call the *pretraining prior*) and the distribution of the test data (which we call the *test distribution*) are the same. However, in practice, Transformers are pretrained on a variety of tasks of different difficulties, and the test distribution will typically exhibit *distribution shift* from the pretraining prior. For example, Figure 1 provides an illustration of two ICL tasks. The in-context examples in the left column are mappings from subfields to their corresponding fields. This is a relatively easy task and LLMs can provide correct answers when given two such examples. In the right column of Figure 1, each word is mapped to (alphabetical position of its first letter) $\times$ (total number of letters) $+ 1$, e.g. 'C' is the third letter in the alphabet and 'Cat' has three letters in total, so 'Cat' is mapped to $3 \times 3 + 1 = 10$. This is a harder problem and LLMs typically require more than two examples to correctly answer the question; see Appendix H. Thus, LLMs can adapt to the difficulties of the tasks in-context, in the sense that fewer examples are needed (or equivalently we witness faster rates of convergence) for easy tasks. Moreover, at test time we may focus more on a specific type of tasks, which represents a distribution shift from the pretraining prior.

To formulate a theoretical framework, we focus on regression problems (in line with most prior work on ICL theory). As an empirical preview, Figure 2 shows the test-time mean squared error (MSE) of a Transformer pretrained on a mixture of regression tasks with different smoothness levels. At test time, the Transformer is evaluated on new tasks of a certain smoothness, both under the original task distribution and under a distribution shift (see Section 5 for more details). We see that the MSE decreases as the smoothness increases (higher smoothness means easier tasks), while the shifted and unshifted curves remain close, suggesting both in-context adaptivity to task difficulty and robustness to distribution shift. The main goal of this paper is to study these phenomena theoretically, and we provide an informal theorem below (see Section 4 for rigorous statements).

Informal theorem. *A Transformer pretrained on a mixture of regression tasks of varying smoothness and effective dimension adapts to the difficulty of a new test task in-context and achieves the corresponding optimal prediction rate. These guarantees also hold under certain distribution shifts between the pretraining and test tasks.*

Our main idea is to view a pretrained Transformer as learning the posterior regression function induced by the pretraining prior. This perspective connects in-context learning with adaptive nonparametric Bayesian regression: once the Transformer approximates this posterior predictor, posterior contraction theory yields task-specific optimal rates, including adaptation to unknown smoothness and effective dimension. Moreover, the same Bayesian viewpoint allows us to establish robustness to distribution shift between the pretraining prior and the test distribution. To the best of our knowledge, our work provides the first proof that Transformer-based in-context learning can be optimally adaptive to task difficulty, while at the same time being robust to distribution shifts.

1.1 Contributions

- In Proposition 2.1, we provide a key test-time risk upper bound, which decomposes into a term that captures both the degree of distribution shift and the proximity of the pretrained Transformer to the posterior regression function, and another term that represents the rate of convergence of the posterior regression function to the true regression function.

- In Section 3.2, we prove a universal approximation theorem for Transformers with softmax attention and non-polynomial activation in FFN layers. We further show that a large Transformer pretrained on sufficient data can learn the posterior regression function arbitrarily well when the activation in FFN layers is one of ReLU, GELU or SiLU.
- In Sections 4.1 and 4.2, we consider general nonparametric regression problems and multi-index models respectively, with regression functions belonging to Besov smoothness classes. We prove that the posterior regression function is optimally adaptive to the difficulty of every task in the support of the pretraining prior (and hence robust to distribution shift). In combination with the results mentioned above, this establishes the adaptivity and distributional robustness of pretrained Transformers.
- In Section 4.3, we show that the expected risk of any estimator over the test distribution μ cannot achieve a faster rate of convergence than our pretrained Transformers (even if the estimator is allowed to depend on μ). This provides a more appropriate form of optimality guarantee for pretrained Transformers than minimax lower bounds usually considered in ICL literature.
- Finally, in Section 5, we verify our theoretical results with empirical experiments. These confirm our main conclusions that pretrained Transformers are able to adapt to different task difficulties and are robust to distribution shifts at test time.

1.2 Related work

In-context learning There is a large body of theoretical research that aims to explain the ICL ability of Transformers. One line of work shows that Transformers can perform ICL by approximating certain algorithms, such as gradient descent (Ahn et al., 2023; Akyürek et al., 2023; Bai et al., 2023; von Oswald et al., 2023), functional gradient descent (Cheng, Chen and Sra, 2024), basis function regression (Kim, Nakamaki and Suzuki, 2024), reinforcement learning (Lin, Bai and Mei, 2024) and domain adaptation (Hataya, Matsui and Imaizumi, 2024). Other lines of work analyse ICL through the lens of Bayesian inference (Xie et al., 2022; Zhang et al., 2022; Nagler, 2023; Wang et al., 2023; Deora et al., 2025; Zhang et al., 2025), meta-learning (Dai et al., 2023; Jeon et al., 2024; Wu, Wang and Yao, 2025) and training dynamics (Oko et al., 2024; Zhang, Frei and Bartlett, 2024; Kumano, Kera and Yamasaki, 2025; Kuwataka and Suzuki, 2025). In particular, Bai et al. (2023) show that Transformers can perform gradient descent on penalised linear regression, generalised linear models and two-layer neural networks, as well as implement in-context algorithm selection, while Kim, Nakamaki and Suzuki (2024) prove that Transformers are minimax optimal for nonparametric regression. Both papers consider the case where the test distribution coincides with the pretraining prior. On the other hand, Xie et al. (2022) study the setting where the test prompts have different formats compared to the pretraining data, and provide consistency results under a Bayesian hidden Markov model. The ability of Transformers to generalise to out-of-distribution tasks has been theoretically studied, under linear regression settings, by Ahuja and Lopez-Paz (2023); Demir and Dogan (2025); Goddard et al. (2025); Wang et al. (2025); Kwon et al. (2026). There are also numerous empirical studies on the behaviour and properties of ICL (e.g. Garg et al., 2022; Kirsch et al., 2022; Panwar, Ahuja and Goyal, 2024; Reuter et al., 2025). Finally, we note the concurrent and independent work of Wakayama and Suzuki (2026), who study the adaptivity of ICL from a Bayesian perspective. Their results are complementary to ours, and neither subsumes the other.

Posterior contraction theory Posterior contraction theory is an active area of research in statistics that studies the concentration of the posterior distribution around a true data generating distribution (i.e. from a frequentist perspective). Pioneered by Ghosal, Ghosh and van der Vaart (2000) and Shen and Wasserman (2001), posterior contraction for Bayesian nonparametric models has been investigated in different problems (e.g. Ghosal, Lember and van der Vaart, 2008; van der Vaart and van Zanten, 2009; Finocchio and Schmidt-Hieber, 2023; Egels and Castillo, 2025). In this paper, we modify the tools developed by Ghosal, Lember and van der Vaart (2008) to derive rates of convergence for the posterior regression function. Interested readers are referred to the book by Ghosal and van der Vaart (2017) for an introduction to the field.

2 Problem set-up and ICL excess risk decomposition

We first remark that commonly used notation in the main text are summarised in Appendix A.1. Unless stated otherwise, all sets and functions are assumed to be measurable. We now present the generating mechanism of our pretraining data. Let $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}$ and let $\mathcal{P}$ be a space of

probability distributions on $\mathcal{X} \times \mathcal{Y}$ (dominated by a common σ-finite measure). For $T \in \mathbb{N}$, define $[T] \coloneqq \{1, \dots, T\}$. The *pretraining prior* π is a distribution on $\mathcal{P}$, and our pretraining data take the form $(\mathcal{D}_n^{(t)}, X^{(t)}, Y^{(t)})_{t=1}^T$, generated as follows:

1. Draw random distributions $P_1, \dots, P_T \overset{\text{iid}}{\sim} \pi$.
2. For each $t \in [T]$, draw $\mathcal{D}_n^{(t)} \coloneqq (X_i^{(t)}, Y_i^{(t)})_{i=1}^n \sim P_t^{\otimes n}$ and $(X^{(t)}, Y^{(t)}) \sim P_t$ independently.

Thus, for each $t \in [T]$, $\mathcal{D}_n^{(t)}$ consists of n (in-context) examples, $X^{(t)}$ is our query and our goal is to predict its corresponding output $Y^{(t)}$; see Figure 1. We emphasise that the *test distribution* μ on $\mathcal{P}$ of new prompts at test time will typically be different from the (possibly unknown) distribution π that generated our pretraining data.

Next, for a class $\mathcal{F}$ of functions from $\mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{Y}$ (e.g. Transformers), define the *empirical risk minimiser* $\widehat{f}_T$ over $\mathcal{F}$ by

$$\widehat{f}_T \in \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{\mathcal{R}}_T(f), \tag{1}$$

where $\widehat{\mathcal{R}}_T(f) \coloneqq \frac{1}{T} \sum_{t=1}^T \big\{Y^{(t)} - f(X^{(t)}, \mathcal{D}_n^{(t)})\big\}^2$ and $(\mathcal{D}_n^{(t)}, X^{(t)}, Y^{(t)})_{t=1}^T$ are generated according to π. We remark that, given a new set of examples $D_n \in (\mathcal{X} \times \mathcal{Y})^n$, the prediction at a query $x \in \mathcal{X}$ is given by $\widehat{f}_T(x, D_n)$, and we do not need to update $\widehat{f}_T$ accordingly at test time.

For a distribution ν on $\mathcal{P}$, we define the *ν-risk* of $g : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$ as[1]

$$\mathcal{R}_\nu(g) \coloneqq \mathbb{E}_{P \sim \nu} \mathbb{E}_{(\mathcal{D}_n, X, Y) | P \sim P^{\otimes(n+1)}} \big[\big\{Y - g(X, \mathcal{D}_n)\big\}^2\big] \equiv \mathbb{E}_\nu \mathbb{E}_P \big[\big\{Y - g(X, \mathcal{D}_n)\big\}^2\big].$$

If $\widetilde{f}_T$ is random and depends on the pretraining data $(\mathcal{D}_n^{(t)}, X^{(t)}, Y^{(t)})_{t=1}^T$, then we interpret $\mathcal{R}_\nu(\widetilde{f}_T)$ as the expectation conditional on the pretraining data, which is assumed to be independent of the test data $(\mathcal{D}_n, X, Y)$, so $\mathcal{R}_\nu(\widetilde{f}_T)$ is also random. For a test distribution μ on $\mathcal{P}$, we further define the *ICL excess μ-risk* of $\widetilde{f}_T$ by

$$\mathcal{R}_\mu^{\text{ICL}}(\widetilde{f}_T) \coloneqq \mathbb{E}\mathcal{R}_\mu(\widetilde{f}_T) - \mathbb{E}_\mu \mathbb{E}_P \big[\big\{Y - \mathbb{E}_P(Y \mid X)\big\}^2\big] = \mathbb{E}\, \mathbb{E}_\mu \mathbb{E}_P \big[\big\{\widetilde{f}_T(X, \mathcal{D}_n) - \mathbb{E}_P(Y \mid X)\big\}^2\big], \tag{2}$$

where the term $\mathbb{E}_\mu \mathbb{E}_P \big[\big\{Y - \mathbb{E}_P(Y \mid X)\big\}^2\big]$ is the oracle risk assuming knowledge of the distribution P that generates (X, Y), and in particular of the true regression function $\mathbb{E}_P(Y \mid X)$ under P; the equality (2) is shown in the proof of Proposition 2.1, see (13). Finally, we define the *posterior regression function* $g_\pi : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$ (with respect to π) by

$$g_\pi(x, D_n) \coloneqq \mathbb{E}_\pi \mathbb{E}_P \big(Y \mid X = x,\, \mathcal{D}_n = D_n\big) \tag{3}$$

for $x \in \mathcal{X}$ and $D_n \in (\mathcal{X} \times \mathcal{Y})^n$. Thus $g_\pi(x, D_n)$ is the posterior mean of Y given $X = x$ and $\mathcal{D}_n = D_n$, where $(\mathcal{D}_n, X, Y) \mid P \sim P^{\otimes(n+1)}$ and $P \sim \pi$.

The following proposition provides a key upper bound on the ICL excess μ-risk, which is crucial for our study of the adaptivity and distributional robustness of ICL. Its proof can be found in Appendix B.

Proposition 2.1. *Let $R > 0$, and suppose that for all $P \in \mathcal{P}$, we have $|\mathbb{E}_P(Y \mid X)| \le R$ almost surely. Further let $\widehat{f}_T$ be defined as in (1), let $\widetilde{f}_T \coloneqq \mathsf{clip}_R \circ \widehat{f}_T$ and let $\mathcal{E}(\widetilde{f}_T) \coloneqq \mathbb{E}\mathcal{R}_\pi(\widetilde{f}_T) - \mathcal{R}_\pi(g_\pi)$. Then we have*

$$\mathcal{R}_\mu^{\text{ICL}}(\widetilde{f}_T) \le 4R \sqrt{\big\{\chi^2(\mu, \pi) + 1\big\} \mathcal{E}(\widetilde{f}_T)} + 2\mathbb{E}_\mu \mathbb{E}_P \big[\big\{g_\pi(X, \mathcal{D}_n) - \mathbb{E}_P(Y \mid X)\big\}^2\big]. \tag{4}$$

The upper bound in (4) consists of two terms. The first involves both the extent of the distribution shift measured by the χ^2-divergence, as well as the expected difference between the π-risk of $\widetilde{f}_T$ and that of the posterior regression function g_π. It is well-known that g_π minimises the π-risk $\mathcal{R}_\pi$ over all measurable functions, so intuitively, if $\widetilde{f}_T$ is a large Transformer pretrained on a sufficient corpus of

[1] We adopt the convention that for $P \in \mathcal{P}$, $\mathbb{E}_P$ means that the random variables are generated from P, i.e. $(\mathcal{D}_n, X, Y) \sim P^{\otimes(n+1)}$, whereas for a distribution ν on $\mathcal{P}$, $\mathbb{E}_\nu \mathbb{E}_P$ means that the distribution P is also random and drawn from ν.

data, then $\mathcal{E}(\widetilde{f}_T)$ should be negligible. Indeed, this is proved in our Proposition 3.2. The second term reflects the proximity of the posterior regression function to the oracle regression function. In fact, in Sections 4.1 and 4.2, we will control $\mathbb{E}_P\big[\{g_\pi(X, \mathcal{D}_n) - \mathbb{E}_P(Y \mid X)\}^2\big]$ uniformly for all P in the support of μ. This reveals the interesting feature that only the first term on the right-hand side of (4) depends on the degree of distribution shift, and this term can be made arbitrarily small by increasing the size of the Transformer and the pretraining dataset, so that the second term ultimately determines the rate of convergence.

Bai et al. (2023) showed that Transformers can approximate algorithm selection based on a train-validation split. Their arguments proceeded via a different risk decomposition to that in Proposition 2.1, where Transformers are regarded as approximators of the oracle regression functions $\mathbb{E}_P(Y \mid X)$; see also Kim, Nakamaki and Suzuki (2024). By contrast, in Proposition 2.1 we view Transformers as approximators of the posterior regression function, which has three key advantages. First, we are able to control the ICL excess risk under distribution shift between the pretraining prior and the test distribution. Second, we are able to demonstrate adaptivity of Transformers to the difficulty of the test task, instead of obtaining a rate of convergence that is a weighted average over difficulty levels (typically dominated by the difficult tasks). Finally, by approximating the posterior regression function, Proposition 2.1 can be applied to settings where no simple closed form algorithms are available or optimal, thereby (in combination with Theorem 3.1 below) explaining the empirically-observed superiority of Transformer-based ICL compared to classical algorithms for nonparametric regression (e.g. Hollmann et al., 2025).

3 Transformers, universal approximation and learnability

3.1 Transformers

In this section, we set up basic notation for Transformer models that we consider. Given $d_{\mathrm{model}} \geq d + 2$, called the *model dimension*, examples $D_n = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and a query $x \in \mathcal{X}$, the input matrix $Z_{\mathrm{in}} \in \mathbb{R}^{(n+1)\times d_{\mathrm{model}}}$ is defined as

$$Z_{\mathrm{in}} \equiv Z_{\mathrm{in}}(D_n, x) := \begin{pmatrix} x_1^\top & y_1 & 0_{d_{\mathrm{model}}-d-2}^\top & 0 \\ \vdots & \vdots & \vdots & \vdots \\ x_n^\top & y_n & 0_{d_{\mathrm{model}}-d-2}^\top & 0 \\ x^\top & 0 & 0_{d_{\mathrm{model}}-d-2}^\top & 1 \end{pmatrix} \tag{5}$$

where the last column of Z_{in} is our positional encoding used to identify the query, and we pad each row with zeros to make them d_{model}-dimensional.

Let $N \in \mathbb{N}$ denote an upper bound on the context length. Our universal approximation theory (Theorem 3.1) holds uniformly for all $n \in [N]$, so it is convenient to consider Transformers as functions from[2] $\bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\mathrm{model}}}$ to $\bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\mathrm{model}}}$. We consider (encoder-only) Transformers with softmax attention, and the activation function in the FFN layers is denoted by ρ. Let $\mathcal{F}_{\mathrm{TF}}(L, H, d_{\mathrm{model}}, d_{\mathrm{ffn}})$ denote the set of all Transformers with L Transformer blocks, H attention heads, model dimension d_{model} and FFN width d_{ffn}; see Appendix C.1 for precise definitions.

For a Transformer $\mathsf{TF} \in \mathcal{F}_{\mathrm{TF}}(L, H, d_{\mathrm{model}}, d_{\mathrm{ffn}})$, examples $D_n = (x_i, y_i)_{i=1}^n$ and a query x, we will slightly abuse our notation by writing $\mathsf{TF}(x, D_n) := \mathsf{TF}\big(Z_{\mathrm{in}}(D_n, x)\big)$ where the input matrix $Z_{\mathrm{in}}(D_n, x)$ is defined by (5). Finally, we define[3] $\mathsf{Read} : \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\mathrm{model}}} \to \mathbb{R}$ by $\mathsf{Read}(Z) := Z_{n+1,d+1}$ for $Z \in \mathbb{R}^{(n+1)\times d_{\mathrm{model}}}$ and $n \in [N]$. Thus, $\mathsf{Read} \circ \mathsf{TF}(x, D_n)$ can be interpreted as the predicted value of the response at the query x based on examples D_n.

3.2 Universal approximation and learnability

We now consider pretraining priors induced by randomly drawn regression functions, and prove universal approximation and learnability results for Transformers. Let $\mathcal{G}$ be a space of real-valued functions on $\mathcal{X} \subseteq \mathbb{R}^d$ that are uniformly bounded by $R > 0$ and let $\widetilde{\pi}$ be a distribution on $\mathcal{G}$. Let X

[2] The space $\bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\mathrm{model}}}$ is equipped with the disjoint union topology and its corresponding Borel σ-algebra.
[3] More generally, $\mathsf{Read}(Z)$ can be an affine transformation of the last row of Z.

be a random variable on $\mathcal{X}$ with distribution P_X, and let $\xi \perp\!\!\!\perp X$ be a zero-mean random variable on $\mathbb{R}$. For $g \in \mathcal{G}$, let P_g be the distribution of (X, Y_g) where $Y_g = g(X) + \xi$. Further let π be the distribution of the random measure $P_{\widetilde{g}}$, where the randomness comes from $\widetilde{g} \sim \widetilde{\pi}$. Suppose throughout this subsection that there exist $R' > 0$ and a (σ-finite) measure ν on $\mathbb{R}$ such that for all $g \in \mathcal{G}$ and almost all realisations of $g(X)$, the conditional distribution of $Y_g \,|\, g(X)$ is absolutely continuous with respect to ν, with Radon–Nikodym derivative bounded by R'. For example, if the noise ξ has a $N(0, \sigma^2)$ distribution, then the conditional distribution of $Y_g \,|\, \{g(X) = z\}$ has bounded Lebesgue density given by $y \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-z)^2/(2\sigma^2)}$.

Furthermore, assume that the activation function $\rho : \mathbb{R} \to \mathbb{R}$ in the FFN layers is Borel measurable and locally essentially bounded (i.e. for every compact set $A \subseteq \mathbb{R}$, there exists $C_A > 0$ such that $|f(x)| \leq C_A$ for Lebesgue almost all $x \in A$), that the closure of the set of discontinuities of ρ has zero Lebesgue measure, and that there is no polynomial h such that $\rho = h$ Lebesgue almost everywhere. The following theorem shows that Transformers are universal approximators for the posterior regression function g_π given by (3).

Theorem 3.1. *For any $L \geq 3$, $H \geq 1$, $\epsilon > 0$ and $N \in \mathbb{N}$, there exist $d_{\text{model}} \geq d + 2$, $d_{\text{ffn}} \in \mathbb{N}$ and a Transformer $\mathsf{TF} \in \mathcal{F}_{\text{TF}}(L, H, d_{\text{model}}, d_{\text{ffn}})$ such that*

$$\max_{n \in [N]} \mathbb{E}_\pi \mathbb{E}_P \Big\{ \big(\mathsf{Read} \circ \mathsf{TF}(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n) \big)^2 \Big\} \leq \epsilon.$$

In Appendix C.2, we in fact show that there exists a Transformer with $L = 3$ and $H = 1$ such that Theorem 3.1 holds. In practice, deeper Transformers will have more expressive power.

From now on, we assume that the activation function ρ in the FFN layers belongs to the set $\{\mathsf{ReLU}, \mathsf{GELU}, \mathsf{SiLU}\}$, which are commonly used activation functions in modern Transformer models (see (9) for their definitions), while the attention layers use softmax activation. For $R > 0$, define $\mathsf{clip}_R(x) := -R \vee x \wedge R$ for $x \in \mathbb{R}$. The following proposition shows that, when the noise ξ is sub-Gaussian, a large Transformer pretrained with sufficient data can achieve a π-risk that is arbitrarily close to the π-risk of g_π. Note that we do not require the input matrix or the parameters of the Transformer to be bounded. We achieve this by showing that the pseudo-dimension of the class is finite using results from a branch of mathematical logic called model theory, see Appendix C.3.

Proposition 3.2. *Suppose that the noise ξ is sub-Gaussian. For any $L \geq 3$, $H \geq 1$, $n \in \mathbb{N}$ and $\epsilon > 0$, there exist $d^\circ_{\text{model}} \geq d + 2$ and $d^\circ_{\text{ffn}} \in \mathbb{N}$ such that the following holds. Suppose $d_{\text{model}} \geq d^\circ_{\text{model}}$, $d_{\text{ffn}} \geq d^\circ_{\text{ffn}}$, and $\widehat{f}_T$ is defined as in (1) with $\mathcal{F} := \mathsf{Read} \circ \mathcal{F}_{\text{TF}}(L, H, d_{\text{model}}, d_{\text{ffn}})$ and pretraining data generated according to π. Let $\widetilde{f}_T := \mathsf{clip}_R \circ \widehat{f}_T$. Then, for all T sufficiently large,*

$$\mathcal{E}(\widetilde{f}_T) := \mathbb{E}\mathcal{R}_\pi(\widetilde{f}_T) - \mathcal{R}_\pi(g_\pi) \leq \epsilon.$$

This proposition confirms formally the intuition provided after Proposition 2.1 that the first term in the decomposition (4) can be made negligible. In other words, Transformers can learn the posterior regression function from data, without knowing the pre-training prior.

4 Optimal in-context adaptivity and distributional robustness

In this section, we consider mixture distribution pretraining priors $\pi = \sum_{\alpha \in \mathcal{A}} \lambda_\alpha \pi_\alpha$ for some finite index set $\mathcal{A}$, and test distributions μ on $\mathcal{P}$ with $\chi^2(\mu, \pi_\beta) \leq \kappa$ for some $\beta \in \mathcal{A}$, so that draws from μ contain tasks of difficulty β with potential distribution shift relative to π_β. We have already seen that Proposition 2.1 provides a decomposition of the ICL excess μ-risk, and Proposition 3.2 shows that the first term on the right-hand side of (4) can be made negligible. Our goal here, then, is to control the key quantity in the other term, namely $\mathbb{E}_P\big[\{g_\pi(X, \mathcal{D}_n) - \mathbb{E}_P(Y \,|\, X)\}^2\big]$, uniformly over all P in the support of π_β. To this end, we specialise the general posterior contraction theory for nonparametric regression that we provide in Appendix E.2 to Besov regression functions with random smoothness (in Section 4.1) and multi-index models (in Section 4.2). All proofs are deferred to Appendix F.

4.1 In-context learning for Besov functions

For this subsection, we take[4] $\mathcal{X} := [0,1]^d, \mathcal{Y} := \mathbb{R}$. To capture the smoothness of regression functions, we consider Besov spaces (see Appendix D.1), which capture a broad range of smoothness behaviours and are among the most widely used function classes for studying optimal rates in nonparametric regression; see, e.g., Donoho and Johnstone (1998), Suzuki (2019) and Giné and Nickl (2021).

To keep the main text focused on our core statistical message, we defer the precise construction of our random Besov functions to Appendix F.1. Informally, for each smoothness level $\alpha > 0$, we construct a distribution $\widetilde{\pi}_\alpha$ on a class $\mathcal{G}_\alpha$ of regression functions with Besov smoothness α. The construction is based on random wavelet coefficients chosen so that, for some $C_0 > 0$,

$$B^\alpha_{\infty,\infty}([0,1]^d, C_0) \subseteq \mathcal{G}_\alpha \subseteq B^\alpha_{\infty,\infty}([0,1]^d, 2C_0).$$

Given $g : [0,1]^d \to \mathbb{R}$, let P_g denote the distribution of (X, Y_g) where $X \sim P_X$ satisfies Assumption F.1,

$$Y_g = g(X) + \xi, \quad \xi \sim N(0, \sigma^2) \quad \text{and} \quad \xi \perp\!\!\!\perp X.$$

We then define π_α to be the distribution of the random measure $P_{\widetilde{g}_\alpha}$, where the randomness comes from the random regression function $\widetilde{g}_\alpha \sim \widetilde{\pi}_\alpha$, and let $\mathcal{P}_\alpha := \mathrm{supp}(\pi_\alpha)$. Hence π_α is a distribution over regression tasks whose regression functions have smoothness α.

Let $\mathcal{A} \subseteq (0, S)$ be a finite set, where $S > 0$ is the regularity of the wavelet basis in Appendix F.1, and let $(\lambda_\alpha)_{\alpha\in\mathcal{A}}$ be a sequence of positive weights such that $\sum_{\alpha\in\mathcal{A}} \lambda_\alpha = 1$. Our pretraining prior π is defined to be the mixture distribution

$$\pi := \sum_{\alpha\in\mathcal{A}} \lambda_\alpha \pi_\alpha. \tag{6}$$

Thus π consists of random regression tasks where the smoothness is randomly drawn from $\mathcal{A}$ with probabilities given by $(\lambda_\alpha)_{\alpha\in\mathcal{A}}$. Finally, we define $\mathcal{P} := \bigcup_{\alpha\in\mathcal{A}} \mathcal{P}_\alpha$ and $\mathcal{G} := \bigcup_{\alpha\in\mathcal{A}} \mathcal{G}_\alpha$.

The following proposition provides the rate of convergence of the posterior regression function, when the true regression function belongs to $\mathcal{G}_\beta$ for some smoothness level $\beta \in \mathcal{A}$.

Proposition 4.1. *For any $\beta \in \mathcal{A}$ and $g^\circ \in \mathcal{G}_\beta$, write $P_0 \equiv P_{g^\circ}$. Let π be the prior distribution defined in (6) and let g_π denote the posterior regression function with respect to π. Then there exists $C > 0$, not depending on n, such that for all $n \in \mathbb{N}$, we have*

$$\mathbb{E}_{P_0}\left[\{g_\pi(X, \mathcal{D}_n) - g^\circ(X)\}^2\right] \le C n^{-2\beta/(2\beta+d)}.$$

Combining Propositions 2.1, 3.2 and 4.1, we are now ready to state the main theorem of this section.

Theorem 4.2. *For any $L \ge 3$, $H \ge 1$, $n \in \mathbb{N}$ and $\kappa > 0$, there exist $d^\circ_{\mathrm{model}} \ge d + 2$ and $d^\circ_{\mathrm{ffn}} \in \mathbb{N}$ such that the following holds. Suppose $d_{\mathrm{model}} \ge d^\circ_{\mathrm{model}}$, $d_{\mathrm{ffn}} \ge d^\circ_{\mathrm{ffn}}$, $\widehat{f}_T$ is defined as in (1) with $\mathcal{F} := \mathsf{Read} \circ \mathcal{F}_{\mathrm{TF}}(L, H, d_{\mathrm{model}}, d_{\mathrm{ffn}})$ and pretraining prior π defined in (6). Let $R := \sup_{g\in\mathcal{G}} \|g\|_\infty$ and $\widetilde{f}_T := \mathsf{clip}_R \circ \widehat{f}_T$. Then, for all T sufficiently large and $\beta \in \mathcal{A}$,*

$$\sup_{\mu\,:\,\chi^2(\mu,\pi_\beta)\le\kappa} \mathcal{R}^{\mathrm{ICL}}_\mu(\widetilde{f}_T) \le C n^{-2\beta/(2\beta+d)},$$

where $C > 0$ does not depend on n and κ.

The Transformer $\widetilde{f}_T$ in Theorem 4.2 is pretrained on the mixture distribution $\pi = \sum_{\alpha\in\mathcal{A}} \lambda_\alpha \pi_\alpha$ with different smoothness levels, whereas the test distribution μ consists of random regression functions of a fixed smoothness level β, with a distribution shift such that $\chi^2(\mu, \pi_\beta) \le \kappa$. When we evaluate the performance of $\widetilde{f}_T$ under the test distribution μ, Theorem 4.2 shows that the ICL excess μ-risk of $\widetilde{f}_T$ is adaptive to the unknown smoothness β, in the sense that its rate of convergence $n^{-2\beta/(2\beta+d)}$ is optimal even if β were known; see Theorem 4.4. Moreover, this risk bound holds uniformly over test distributions in a χ^2-divergence ball around π_β. Thus, although the Transformer is pretrained on tasks of different levels of difficulties, it is able to achieve faster (and optimal) rates of convergence

[4] Our results in this section extend to any compact domain $\mathcal{X} \subseteq \mathbb{R}^d$ by a scaling argument.

on easier tasks (without knowing the difficulty of the tasks) and is robust to distribution shift. Finally, we remark that d°_{model}, d°_{ffn} and T should increase with κ in order to keep C independent of κ. Thus, for a Transformer with fixed architecture, fine tuning for downstream tasks may still be beneficial in cases where (i) the Transformer is not large enough, (ii) the amount of pretraining data is not sufficient, (iii) there is significant distribution shift (κ is large), or (iv) there is shift in the support of the pretraining prior and the test distribution.

4.2 In-context learning for multi-index models

To avoid the curse of dimensionality in the previous section, we will now consider multi-index models with[5] $\mathcal{X} = \mathbb{B}^d := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ and $\mathcal{Y} = \mathbb{R}$. In general, a multi-index model takes the form $x \mapsto g\big(A(x)\big)$ where $A : \mathbb{R}^d \to \mathbb{R}^p$ is an affine function, $p \in [d]$, and $g : \mathbb{R}^p \to \mathbb{R}$ has smoothness α. Here, p is called the effective dimension and α is the smoothness of the multi-index model.

The formal definition of our random multi-index model is stated in Appendix F.2. Informally, we define $\mathcal{G}_{\alpha,p}$ to be a set of multi-index models with smoothness α and effective dimension p, and let[6] $\pi_{\alpha,p}$ be a distribution that generates random tasks with random regression functions drawn from $\mathcal{G}_{\alpha,p}$. For a finite set $\mathcal{A}' \subseteq (0, S) \times [d]$ and a sequence of positive weights $(\lambda_{\alpha,p})_{(\alpha,p)\in\mathcal{A}'}$ such that $\sum_{(\alpha,p)\in\mathcal{A}'} \lambda_{\alpha,p} = 1$, our pretraining prior π for multi-index models is defined to be the mixture distribution

$$\pi := \sum_{(\alpha,p)\in\mathcal{A}'} \lambda_{\alpha,p}\pi_{\alpha,p}. \tag{7}$$

Thus π generates random multi-index regression tasks, where the smoothness and effective dimensions of the regression functions are randomly drawn from $\mathcal{A}'$ with probabilities given by $(\lambda_{\alpha,p})_{(\alpha,p)\in\mathcal{A}'}$. Finally, we define $\mathcal{P} := \bigcup_{(\alpha,p)\in\mathcal{A}'} \mathcal{P}_{\alpha,p}$ and $\mathcal{G} := \bigcup_{(\alpha,p)\in\mathcal{A}'} \mathcal{G}_{\alpha,p}$. Proposition F.7, stated and proved in Appendix F.4, is an analogous result to Proposition 4.1 but for multi-index models. Combining Propositions 2.1, 3.2 and F.7, we can now conclude that a pretrained Transformer simultaneously adapts to both the effective dimension of the multi-index models and the smoothness of the regression function, and is robust to distribution shift.

Theorem 4.3. *For any $L \geq 3$, $H \geq 1$, $n \in \mathbb{N}$ and $\kappa > 0$, there exist $d^{\circ}_{\text{model}} \geq d + 2$ and $d^{\circ}_{\text{ffn}} \in \mathbb{N}$ such that the following holds. Suppose $d_{\text{model}} \geq d^{\circ}_{\text{model}}$, $d_{\text{ffn}} \geq d^{\circ}_{\text{ffn}}$, $\widehat{f}_T$ is defined as in (1) with $\mathcal{F} := \mathsf{Read} \circ \mathcal{F}_{\text{TF}}(L, H, d_{\text{model}}, d_{\text{ffn}})$ and pretraining prior π defined in (7). Let $R := \sup_{g\in\mathcal{G}} \|g\|_\infty$ and $\widetilde{f}_T := \mathsf{clip}_R \circ \widehat{f}_T$. Then, for all T sufficiently large and $(\beta, r) \in \mathcal{A}'$,*

$$\sup_{\mu\,:\,\chi^2(\mu,\pi_{\beta,r})\leq\kappa} \mathcal{R}^{\text{ICL}}_{\mu}(\widetilde{f}_T) \leq Cn^{-2\beta/(2\beta+r)},$$

where $C > 0$ does not depend on n and κ.

The rate in Theorem 4.3 depends on the effective dimension r rather than the ambient dimension d, thus avoiding the curse of dimensionality incurred in the setting of Theorem 4.2.

4.3 Lower bounds on the ICL excess risk

Our final theoretical contribution is to provide lower bounds on the ICL excess risk. Prior work has sought to establish the optimality of ICL algorithms via lower bounds on the minimax risk

$$\inf_{\widehat{g}_n\in\widehat{\mathcal{G}}_n} \sup_{P\in\text{supp}(\mu)} \mathbb{E}_P\big[\big\{\widehat{g}_n(X, \mathcal{D}_n) - \mathbb{E}_P(Y \mid X)\big\}^2\big], \tag{8}$$

where $\widehat{\mathcal{G}}_n$ denotes the set of all measurable functions from $\mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{Y}$. Although the minimax risk is often regarded as the gold standard in traditional statistical learning problems, the ICL excess μ-risk $\mathcal{R}^{\text{ICL}}_{\mu}$ is an expectation over $P \sim \mu$, not a worst-case risk over $P \in \text{supp}(\mu)$. Thus lower bounds on (8) do not provide lower bounds for the ICL excess μ-risk. Our Theorem 4.4 below, however, does indeed provide a lower bound on the ICL excess μ-risk, and in a strong sense: even an estimator that has knowledge of the test distribution μ cannot achieve a better rate of convergence than the pretrained Transformers in Theorems 4.2 and 4.3.

[5] Again, our results in this section extend to any compact domain $\mathcal{X} \subseteq \mathbb{R}^d$ by a scaling argument.

[6] Note that we index difficulty by the pair of parameters (α, p) in this subsection.

Theorem 4.4. *(a) Under the setting of Section 4.1, let $\beta \in \mathcal{A}$ and let $\mu \ll \pi_\beta$ be such that $\sup_{P\in\mathcal{P}_\beta} \frac{\mathrm{d}\mu}{\mathrm{d}\pi_\beta}(P) < \infty$. Then there exists $c > 0$, not depending on n, such that*

$$\inf_{\widehat{g}_n \in \widehat{\mathcal{G}}_n} \mathcal{R}_\mu^{\mathrm{ICL}}(\widehat{g}_n) \geq cn^{-2\beta/(2\beta+d)}.$$

(b) In the setting of Section 4.2, let $(\beta, r) \in \mathcal{A}'$ and let $\mu \ll \pi_{\beta,r}$ be such that $\sup_{P\in\mathcal{P}_{\beta,r}} \frac{\mathrm{d}\mu}{\mathrm{d}\pi_{\beta,r}}(P) < \infty$. Then there exists $c' > 0$, not depending on n, such that

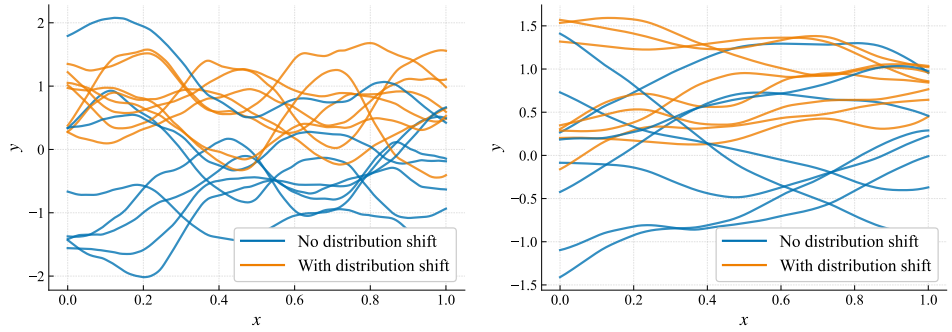
$$\inf_{\widehat{g}_n \in \widehat{\mathcal{G}}_n} \mathcal{R}_\mu^{\mathrm{ICL}}(\widehat{g}_n) \geq c'n^{-2\beta/(2\beta+r)}.$$

We remark that, if $\widetilde{f}_T$ is defined as in Theorems 4.2 or 4.3, then conditional on the pretraining data, we have $\widetilde{f}_T \in \widehat{\mathcal{G}}_n$.

5 Simulations

The goal of this section is to verify our theoretical results by synthetic experiments. To this end, we consider single-index models with $n = 30$, $d = 5$, $p = 1$, $\sigma = 0.01$ and four levels of Besov smoothness $\alpha \in \{2,\ 2.5,\ 3,\ 4\}$. For our Transformer, we use a GPT-2 medium-style model (Radford et al., 2019) with $L = 24$, $H = 16$, $d_{\mathrm{model}} = 1024$, $d_{\mathrm{ffn}} = 4096$ and GELU activation function in the FFN layers. Further experimental details are provided in Appendix I. The code for our simulations is available at `https://github.com/tianyima2000/Adaptive_ICL`.

At test time, we randomly generate new single-index models with fixed smoothness levels under two settings: (i) no distribution shift, and (ii) with distribution shift. See Figure 3 for random Besov functions with and without distribution shift. Details for the data generating mechanism are provided in Appendix I.1. Figure 4(a) shows the test-time MSE of our pretrained Transformer for different levels of α. We see that our Transformer is able to adapt to the smoothness and achieve a faster rate of convergence as α increases; moreover, our Transformer is also robust to distribution shift at test time. Figure 4(b) plots the logarithm of the MSE against $2\alpha/(2\alpha+1)$, revealing that the log MSE is approximately linear in $2\alpha/(2\alpha+1)$. This is in line with the conclusion of Theorem 4.3. We also provide additional empirical studies in Appendix I.3, including further simulations on within-support and out-of-support distribution shifts, as well as simulations plotting the logarithm of the MSE against the logarithm of the number of pretraining sequences.

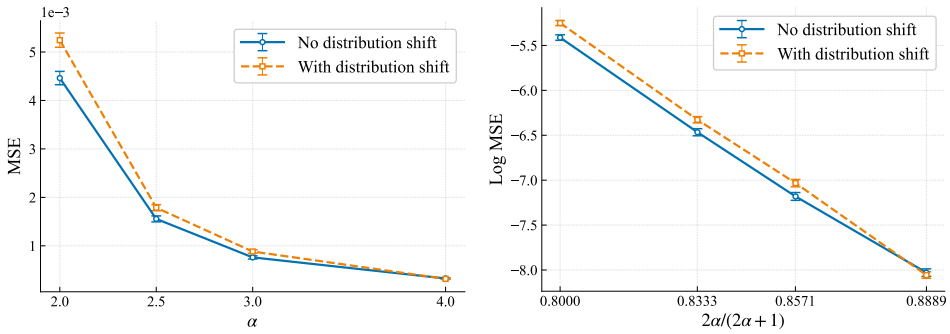


(a) $\alpha = 2$. (b) $\alpha = 4$.

Figure 3: Random Besov functions with smoothness $\alpha \in \{2, 4\}$.

6 Conclusion

In this work, we show that pretrained Transformers adapt to different task difficulties at optimal rates, and are robust to distribution shifts at test time. Our results connect ICL with posterior contraction theory, and provide a rigorous explanation for the empirical success of ICL.



(a) MSE against α.

(b) Log MSE against $\frac{2\alpha}{2\alpha+1}$.

Figure 4: MSE and log MSE at test time based on 100,000 random test functions. Error bars represent 90% confidence intervals.

A natural extension of this work would be to combine our theory with the quantitative approximation theory derived in independent, concurrent work by Wakayama and Suzuki (2026) to provide more explicit values of d°_{model}, d°_{ffn} and T that guarantee the conclusions of Theorems 4.2 and 4.3. More broadly, our work opens the door to providing adaptive and distributionally robust ICL guarantees for Transformers in other settings, including heavy-tailed noise, classification and dependent data.

Acknowledgement

The research of the first and third author was supported by RJS's European Research Council Advanced Grant 101019498.

References

Ahn, K., Cheng, X., Daneshmand, H. and Sra, S. (2023) Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, **36**, 45614–45650.

Ahuja, K. and Lopez-Paz, D. (2023) A Closer Look at In-Context Learning under Distribution Shifts. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

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

Anderson, J. and Pitt, L. (1989) Probabilistic behaviour of functions in the Zygmund spaces Λ^* and λ^*. *Proceedings of the London Mathematical Society*, **3**, 558–592.

Bai, Y., Chen, F., Wang, H., Xiong, C. and Mei, S. (2023) Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in Neural Information Processing Systems*, **36**, 57125–57211.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

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. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems*, **33**, 1877–1901.

Chen, X., Guntuboyina, A. and Zhang, Y. (2016) On Bayes risk lower bounds. *Journal of Machine Learning Research*, **17**, 1–58.

Cheng, X., Chen, Y. and Sra, S. (2024) Transformers implement functional gradient descent to learn non-linear functions in context. *Proceedings of the 41st International Conference on Machine Learning*, 8002–8037.

Chikuse, Y. (2003) *Statistics on Special Manifolds*. Springer Science & Business Media.

Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, **1**, 54–81.

Dahmen, W. (1997) Wavelet and multiscale methods for operator equations. *Acta Numerica*, **6**, 55–228.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z. and Wei, F. (2023) Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. *Findings of the Association for Computational Linguistics: ACL 2023*, 4005–4019.

Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, **41**, 909–996.

Demir, S. and Dogan, Z. (2025) Optimal Attention Temperature Enhances In-Context Learning under Distribution Shift. *arXiv preprint arXiv:2511.01292*.

Deora, P., Vasudeva, B., Behnia, T. and Thrampoulidis, C. (2025) In-context Occam's Razor: How Transformers prefer simpler hypotheses on the fly. In *Second Conference on Language Modeling*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Egels, P. and Castillo, I. (2025) Posterior and variational inference for deep neural networks with heavy-tailed weights. *Journal of Machine Learning Research*, **26**, 1–58.

Finocchio, G. and Schmidt-Hieber, J. (2023) Posterior contraction for deep Gaussian process priors. *Journal of Machine Learning Research*, **24**, 1–49.

Folland, G. B. (1999) *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons.

Garg, S., Tsipras, D., Liang, P. S. and Valiant, G. (2022) What can transformers learn in-context? A case study of simple function classes. *Advances in Neural Information Processing Systems*, **35**, 30583–30598.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000) Convergence rates of posterior distributions. *Annals of Statistics*, **28**, 500–531.

Ghosal, S., Lember, J. and van der Vaart, A. (2008) Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, **2**, 63–89.

Ghosal, S. and van der Vaart, A. W. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Giné, E. and Nickl, R. (2011) Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Annals of Statistics*, **39**, 2883–2911.

Giné, E. and Nickl, R. (2021) *Mathematical Foundations of Infinite-dimensional Statistical Models*. Cambridge University Press.

Goddard, C., Smith, L. M., Ngampruetikorn, V. and Schwab, D. J. (2025) When can in-context learning generalize out of task distribution? In *Forty-second International Conference on Machine Learning*.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002) *A Distribution-free Theory of Nonparametric Regression*. Springer.

Hataya, R., Matsui, K. and Imaizumi, M. (2024) Automatic domain adaptation by transformers in in-context learning. In *ICML 2024 Workshop on In-Context Learning*.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T. and Hutter, F. (2025) Accurate predictions on small data with a tabular foundation model. *Nature*, **637**, 319–326.

Jeon, H. J., Lee, J. D., Lei, Q. and Roy, B. V. (2024) An information-theoretic analysis of in-context learning. In *Forty-first International Conference on Machine Learning*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Karpinski, M. and Macintyre, A. (1995) Polynomial bounds for VC dimension of sigmoidal neural networks. In *Proceedings of the Twenty-seventh Annual ACM Symposium on Theory of Computing*, 200–208.

Kim, J., Nakamaki, T. and Suzuki, T. (2024) Transformers are minimax optimal nonparametric in-context learners. *Advances in Neural Information Processing Systems*, **37**, 106667–106713.

Kirsch, L., Harrison, J., Sohl-Dickstein, J. and Metz, L. (2022) General-purpose in-context learning by meta-learning Transformers. In *Sixth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*.

Kumano, S., Kera, H. and Yamasaki, T. (2025) Adversarially pretrained Transformers may be universally robust in-context learners. *arXiv preprint arXiv:2505.14042*.

Kuwataka, K. and Suzuki, T. (2025) Test time training enhances in-context learning of nonlinear functions. *arXiv preprint arXiv:2509.25741*.

Kwon, S. M., Xu, A. S., Yaras, C., Balzano, L. and Qu, Q. (2026) Out-of-Distribution Generalization of In-Context Learning: A Low-Dimensional Subspace Perspective. In *The 29th International Conference on Artificial Intelligence and Statistics*.

Laskowski, M. C. (1992) Vapnik–Chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, **2**, 377–384.

Leshno, M., Lin, V. Y., Pinkus, A. and Schocken, S. (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, **6**, 861–867.

Lin, L., Bai, Y. and Mei, S. (2024) Transformers as Decision Makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I. and Hutter, F. (2019) Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ma, T., Wang, T. and Samworth, R. J. (2025) Deep learning with missing data. *arXiv preprint arXiv:2504.15388*.

Marker, D. (2002) *Model Theory: An Introduction*. Springer.

Nagler, T. (2023) Statistical Foundations of Prior-Data Fitted Networks. In *Proceedings of the 40th International Conference on Machine Learning*, 25660–25676.

Oko, K., Song, Y., Suzuki, T. and Wu, D. (2024) Pretrained transformer efficiently learns low-dimensional target functions in-context. *Advances in Neural Information Processing Systems*, **37**, 77316–77365.

Panwar, M., Ahuja, K. and Goyal, N. (2024) In-context learning through the Bayesian prism. In *The Twelfth International Conference on Learning Representations*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019) Language models are unsupervised multitask learners. *OpenAI blog*, **1**, 9.

Reiß, M. and Schmidt-Hieber, J. (2020) Posterior contraction rates for support boundary recovery. *Stochastic Processes and their Applications*, **130**, 6638–6656.

Reuter, A., Rudner, T. G. J., Fortuin, V. and Rügamer, D. (2025) Can transformers learn full Bayesian inference in context? In *Forty-second International Conference on Machine Learning*.

Rudin, W. (1987) *Real and Complex Analysis*. McGraw-Hill, Inc.

Shen, X. and Wasserman, L. (2001) Rates of convergence of posterior distributions. *Annals of Statistics*, **29**, 687–714.

Speissegger, P. (1999) The Pfaffian closure of an o-minimal structure. *Journal für die Reine und Angewandte Mathematik*, **508**, 189–211.

Suzuki, T. (2019) 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*.

Triebel, H. (1983) *Theory of Function Spaces*. Birkhäuser Basel.

van der Vaart, A. W. and van Zanten, J. H. (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Annals of Statistics*, **37**, 2655–2675.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A. and Vladymyrov, M. (2023) Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 35151–35174.

Wakayama, T. and Suzuki, T. (2026) In-context learning is provably Bayesian inference: a generalization theory for meta-learning. *International Conference on Machine Learning*.

Wang, Q., Wang, Y., Ying, X. and Wang, Y. (2025) Can In-context Learning Really Generalize to Out-of-distribution Tasks? In *The Thirteenth International Conference on Learning Representations*.

Wang, X., Zhu, W., Saxon, M., Steyvers, M. and Wang, W. Y. (2023) Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, **36**, 15614–15638.

Wu, S., Wang, Y. and Yao, Q. (2025) Why in-context learning models are good few-shot learners? In *The Thirteenth International Conference on Learning Representations*.

Xie, S. M., Raghunathan, A., Liang, P. and Ma, T. (2022) An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L. and Liu, T. (2020) On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 10524–10533.

Zhang, R., Frei, S. and Bartlett, P. L. (2024) Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, **25**, 1–55.

Zhang, Y., Liu, B., Cai, Q., Wang, L. and Wang, Z. (2022) An analysis of attention via the lens of exchangeability and latent variable models. *arXiv preprint arXiv:2212.14852*.

Zhang, Y., Zhang, F., Yang, Z. and Wang, Z. (2025) What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, 1684–1692.

Zygmund, A. (2002) *Trigonometric Series*. Cambridge University Press.

A Notation

A.1 Notation for the main text

For $n \in \mathbb{N}$, we write $[n] \coloneqq \{1, \dots, n\}$. For $m, n \in \mathbb{N}$ with $m < n$, we write $[m : n] \coloneqq \{m, m+1, \dots, n\}$ and similarly $[n : \infty) \coloneqq \{n, n+1, \dots\}$. Given $a, b \in \mathbb{R}$, we write $a \wedge b \coloneqq \min\{a, b\}$ and $a \vee b \coloneqq \max\{a, b\}$. For $d \in \mathbb{N}$, we define $0_d \coloneqq (0, \dots, 0)^\top \in \mathbb{R}^d$, $1_d \coloneqq (1, \dots, 1)^\top \in \mathbb{R}^d$ and $\mathbb{B}^d \coloneqq \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. For $R > 0$, we define the clipping operator $\mathsf{clip}_R : \mathbb{R} \to [-R, R]$ by $\mathsf{clip}_R(x) \coloneqq -R \vee x \wedge R$ for $x \in \mathbb{R}$. Given a Borel measurable set $A \subseteq \mathbb{R}^d$, we write $\mathrm{Vol}_d(A)$ for the d-dimensional Lebesgue measure of A. For two probability measures P and Q on a measurable space $(\mathcal{X}, \Sigma)$, we say that P is absolutely continuous with respect to Q, written as $P \ll Q$, if $P(A) = 0$ for all $A \in \Sigma$ such that $Q(A) = 0$, and in that case write $\frac{\mathrm{d}P}{\mathrm{d}Q}$ for the Radon–Nikodym derivative. When P and Q are probability measures, the χ^2-divergence from Q to P is defined as $\chi^2(P, Q) \coloneqq \int_{\mathcal{X}} \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q - 1$ if $P \ll Q$, and $\chi^2(P, Q) \coloneqq \infty$ otherwise. Further, we write $P^{\otimes n}$ for the n-fold product measure of P and $\mathrm{supp}(P)$ for the support of P, i.e. the intersection of all closed $A \in \Sigma$ with $P(A) = 1$. For a set $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ and a set $\mathcal{G}$ of functions from $\mathcal{Y}$ to $\mathcal{Z}$, we write $\mathcal{F} \circ \mathcal{G} \coloneqq \{f \circ g : f \in \mathcal{F},\, g \in \mathcal{G}\}$, where $f \circ g$ denotes the composition of f and g. For a function $f : \mathcal{X} \to \mathbb{R}$, we write $\|f\|_\infty \coloneqq \sup_{x \in \mathcal{X}} |f(x)|$.

A.2 Notation for the Appendix

Let $(\mathcal{P}, d)$ be a metric space, $\epsilon > 0$ and $\mathcal{P}' \subseteq \mathcal{P}$, we write $\mathcal{N}(\epsilon, \mathcal{P}', d)$ for the minimum number of closed ϵ-balls in d needed to cover $\mathcal{P}'$. For $m, n \in \mathbb{N}$, let $I_n \in \mathbb{R}^{n \times n}$ be the identity matrix; for a matrix $A \in \mathbb{R}^{n \times m}$, define the operator norm of A by $\|A\|_{\mathrm{op}} \coloneqq \sup_{x \in \mathbb{R}^m : \|x\|_2 = 1} \|Ax\|_2$ and define its entrywise ℓ_∞-norm by $\|A\|_\infty \coloneqq \max_{i \in [n], j \in [m]} |A_{ij}|$.

Let P and Q be probability measures on $\mathcal{Z}$, and let ν be a measure on $\mathcal{Z}$ such that $P \ll \nu$ and $Q \ll \nu$ (such a measure always exists, e.g. $\nu = P + Q$). Write $p = \frac{\mathrm{d}P}{\mathrm{d}\nu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\nu}$, then the ***Hellinger distance*** between P and Q is defined as

$$d_{\mathrm{H}}(P, Q) \coloneqq \left(\int_{\mathcal{Z}} \left(\sqrt{p(z)} - \sqrt{q(z)} \right)^2 \mathrm{d}\nu(z) \right)^{1/2}.$$

The definition is independent of the choice of the dominating measure ν, and d_{H} is a metric on the space of distributions on $\mathcal{Z}$. The ***Kullback–Leibler divergence*** from Q to P is defined as

$$\mathrm{KL}(P, Q) \coloneqq \mathbb{E}_P \left(\log \frac{\mathrm{d}P}{\mathrm{d}Q}(Z) \right) = \int_{\mathcal{Z}} \log \frac{\mathrm{d}P}{\mathrm{d}Q}(z)\, \mathrm{d}P(z)$$

if $P \ll Q$, and infinity otherwise; this is the P-expectation of the log-likelihood ratio between P and Q. We further define

$$\mathrm{V}_2(P, Q) \coloneqq \mathrm{Var}_P \left(\log \frac{\mathrm{d}P}{\mathrm{d}Q}(Z) \right) = \int_{\mathcal{Z}} \left(\log \frac{\mathrm{d}P}{\mathrm{d}Q}(z) - \mathrm{KL}(P, Q) \right)^2 \mathrm{d}P(z)$$

if $P \ll Q$, and infinity otherwise; this is the P-variance of the log-likelihood ratio between P and Q.

For $d \in \mathbb{N}$, $p \in [1, \infty)$, a Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$ and a Borel measure μ on $\mathbb{R}^d$, we define

$$\|f\|_{L^p(\mu)} \coloneqq \left(\int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}\mu(x) \right)^{1/p}.$$

and write $L^p(\mu)$ for the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ with $\|f\|_{L^p(\mu)} < \infty$. If the domain of f is restricted to $\mathcal{X} \subseteq \mathbb{R}^d$ and μ is the Lebesgue measure restricted to a Borel set $\mathcal{X}$, then we write $\|f\|_{L^p(\mathcal{X})} \equiv \|f\|_{L^p(\mu)}$ and $L^p(\mathcal{X}) \equiv L^p(\mu)$; when $\mathcal{X}$ is clear from context, we further write $\|f\|_{L^p} \equiv \|f\|_{L^p(\mathcal{X})}$.

For a zero-mean random variable X on $\mathbb{R}$, we say that X is ***sub-Gamma in the right tail*** with variance parameter $\sigma^2 > 0$ and scale parameter $c > 0$ if for all $\lambda \in [0, 1/c)$, we have

$$\mathbb{E} e^{\lambda X} \leq \exp\left(\frac{\sigma^2 \lambda^2}{2(1 - c\lambda)} \right).$$

Similarly, X is *sub-Gamma in the left tail* if $-X$ is sub-Gamma in the right tail. And we say X is *sub-Gamma in both tails* is it is both sub-Gamma in the right tail and sub-Gamma in the left tail.

Finally, we define three commonly used activation functions in Transformers. For $x \in \mathbb{R}$, define

$$\mathsf{ReLU}(x) \coloneqq \max\{0, x\}, \quad \mathsf{GELU}(x) \coloneqq x\Phi(x) \quad \text{and} \quad \mathsf{SiLU}(x) \coloneqq \frac{x}{1+e^{-x}}, \tag{9}$$

where $\Phi(x) \coloneqq \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t$ is the distribution function of standard Gaussian distribution.

B Proof of Proposition 2.1

Proof of Proposition 2.1. We may assume without loss of generality that $\mu \ll \pi$ since otherwise, the right-hand side of (4) is infinite. Note that

$$\begin{aligned}
\mathbb{E}\mathcal{R}_\pi(\widetilde{f}_T) - \mathcal{R}_\pi(g_\pi) &= \mathbb{E}\,\mathbb{E}_\pi\mathbb{E}_P\big[\big\{Y - \widetilde{f}_T(X, \mathcal{D}_n)\big\}^2 - \big\{Y - g_\pi(X, \mathcal{D}_n)\big\}^2\big] \\
&= \mathbb{E}\,\mathbb{E}_\pi\mathbb{E}_P\big\{-2Y\widetilde{f}_T(X, \mathcal{D}_n) + \widetilde{f}_T^2(X, \mathcal{D}_n) + 2Y g_\pi(X, \mathcal{D}_n) - g_\pi^2(X, \mathcal{D}_n)\big\},
\end{aligned}$$

where the outer expectation is taken over the randomness of $\widetilde{f}_T$. Moreover,

$$\begin{aligned}
\mathbb{E}_\pi\mathbb{E}_P\big\{Y\widetilde{f}_T(X, \mathcal{D}_n)\big\} &= \mathbb{E}_\pi\mathbb{E}_P\big[\mathbb{E}_{P\sim\pi,\,(\mathcal{D}_n,X,Y)\sim P^{\otimes(n+1)}}\big\{Y\widetilde{f}_T(X, \mathcal{D}_n)\,\big|\,X, \mathcal{D}_n\big\}\big] \\
&= \mathbb{E}_\pi\mathbb{E}_P\big\{\widetilde{f}_T(X, \mathcal{D}_n)\cdot\mathbb{E}_{P\sim\pi,\,(\mathcal{D}_n,X,Y)\sim P^{\otimes(n+1)}}\big(Y\,\big|\,X, \mathcal{D}_n\big)\big\} \\
&= \mathbb{E}_\pi\mathbb{E}_P\big\{\widetilde{f}_T(X, \mathcal{D}_n)g_\pi(X, \mathcal{D}_n)\big\},
\end{aligned}$$

and similarly

$$\begin{aligned}
\mathbb{E}_\pi\mathbb{E}_P\big\{Y g_\pi(X, \mathcal{D}_n)\big\} &= \mathbb{E}_\pi\mathbb{E}_P\big[\mathbb{E}_{P\sim\pi,\,(\mathcal{D}_n,X,Y)\sim P^{\otimes(n+1)}}\big\{Y g_\pi(X, \mathcal{D}_n)\,\big|\,X, \mathcal{D}_n\big\}\big] \\
&= \mathbb{E}_\pi\mathbb{E}_P\big\{g_\pi^2(X, \mathcal{D}_n)\big\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{E}(\widetilde{f}_T) &= \mathbb{E}\mathcal{R}_\pi(\widetilde{f}_T) - \mathcal{R}_\pi(g_\pi) = \mathbb{E}\,\mathbb{E}_\pi\mathbb{E}_P\big\{-2g_\pi(X, \mathcal{D}_n)\widetilde{f}_T(X, \mathcal{D}_n) + \widetilde{f}_T^2(X, \mathcal{D}_n) + g_\pi^2(X, \mathcal{D}_n)\big\} \\
&= \mathbb{E}\,\mathbb{E}_\pi\mathbb{E}_P\big[\big\{\widetilde{f}_T(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n)\big\}^2\big].
\end{aligned}$$

By (3), we have $|g_\pi(X, \mathcal{D}_n)| \le R$ almost surely. Now, for every fixed $P \in \mathcal{P}$, define $Z_P \coloneqq \mathbb{E}\,\mathbb{E}_P\big[\big\{\widetilde{f}_T(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n)\big\}^2\big]$, so that by Fubini's theorem and the Cauchy–Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\,\mathbb{E}_\mu\mathbb{E}_P\big[\big\{\widetilde{f}_T(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n)\big\}^2\big] &= \mathbb{E}_\mu(Z_P) = \mathbb{E}_\pi\bigg(Z_P\cdot\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(P)\bigg) \\
&\le \sqrt{\mathbb{E}_\pi(Z_P^2)\cdot\mathbb{E}_\pi\bigg\{\bigg(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(P)\bigg)^2\bigg\}} \le \sqrt{4R^2\mathbb{E}_\pi(Z_P)\big\{\chi^2(\mu,\pi)+1\big\}} \\
&= 2R\sqrt{\big\{\chi^2(\mu,\pi)+1\big\}\mathcal{E}(\widetilde{f}_T)}.
\end{aligned} \tag{10}$$

Further,

$$\begin{aligned}
\mathbb{E}_\mu\mathbb{E}_P\big\{Y\widetilde{f}_T(X, \mathcal{D}_n)\big\} &= \mathbb{E}_\mu\mathbb{E}_P\big[\mathbb{E}_P\big\{Y\widetilde{f}_T(X, \mathcal{D}_n)\,\big|\,X, \mathcal{D}_n\big\}\big] \\
&= \mathbb{E}_\mu\mathbb{E}_P\big\{\widetilde{f}_T(X, \mathcal{D}_n)\cdot\mathbb{E}_P(Y\,|\,X, \mathcal{D}_n)\big\} \\
&= \mathbb{E}_\mu\mathbb{E}_P\big\{\widetilde{f}_T(X, \mathcal{D}_n)\cdot\mathbb{E}_P(Y\,|\,X)\big\},
\end{aligned} \tag{11}$$

and similarly,

$$\mathbb{E}_\mu\mathbb{E}_P\big\{Y\mathbb{E}_P(Y\,|\,X)\big\} = \mathbb{E}_\mu\mathbb{E}_P\big[\mathbb{E}_P\big\{Y\mathbb{E}_P(Y\,|\,X)\,\big|\,X\big\}\big] = \mathbb{E}_\mu\mathbb{E}_P\big\{\big(\mathbb{E}_P(Y\,|\,X)\big)^2\big\}. \tag{12}$$

Hence, we deduce that

$$\begin{aligned}
\mathcal{R}_\mu^{\mathrm{ICL}}(\widetilde{f}_T) &= \mathbb{E}\mathcal{R}_\mu(\widetilde{f}_T) - \mathbb{E}_\mu\mathbb{E}_P\big[\big\{Y - \mathbb{E}_P(Y\,|\,X)\big\}^2\big] \\
&= \mathbb{E}\,\mathbb{E}_\mu\mathbb{E}_P\big\{-2Y\widetilde{f}_T(X, \mathcal{D}_n) + \widetilde{f}_T^2(X, \mathcal{D}_n) + 2Y\mathbb{E}_P(Y\,|\,X) - \big(\mathbb{E}_P(Y\,|\,X)\big)^2\big\} \\
&\overset{(i)}{=} \mathbb{E}\,\mathbb{E}_\mu\mathbb{E}_P\big[\big\{\widetilde{f}_T(X, \mathcal{D}_n) - \mathbb{E}_P(Y\,|\,X)\big\}^2\big] \\
&\le 2\mathbb{E}\,\mathbb{E}_\mu\mathbb{E}_P\big[\big\{\widetilde{f}_T(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n)\big\}^2\big] + 2\mathbb{E}_\mu\mathbb{E}_P\big[\big\{g_\pi(X, \mathcal{D}_n) - \mathbb{E}_P(Y\,|\,X)\big\}^2\big] \\
&\overset{(ii)}{\le} 4R\sqrt{\big\{\chi^2(\mu,\pi)+1\big\}\mathcal{E}(\widetilde{f}_T)} + 2\mathbb{E}_\mu\mathbb{E}_P\big[\big\{g_\pi(X, \mathcal{D}_n) - \mathbb{E}_P(Y\,|\,X)\big\}^2\big],
\end{aligned} \tag{13}$$

where (i) follows from (11) and (12), and (ii) follows from (10). $\square$

C Transformer definition and proofs for Section 3

C.1 Transformer definition

In this section, we formally define the classes of (encoder-only) Transformers that we will consider. For a vector $v = (v_1, \dots, v_d)^\top \in \mathbb{R}^d$, let $\mathsf{Softmax}(v) \in \mathbb{R}^d$ denote the vector whose jth entry is given by $\exp(v_j)/\sum_{\ell=1}^d \exp(v_\ell)$. Often, we will apply the $\mathsf{Softmax}$ function to matrices, by which we mean that the $\mathsf{Softmax}$ function is applied row-wise. A Transformer iteratively applies attention layers and feed forward network (FFN) layers to the input matrix Z_{in} (5).

Definition C.1 (Attention layer)**.** Let $H \in \mathbb{N}$ and for $h \in [H]$, let $Q_h, K_h, V_h \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$. The *attention layer* with H heads and parameters $\theta_{\text{attn}} := (Q_h, K_h, V_h)_{h=1}^H$ is the function $\mathsf{Attn}_{\theta_{\text{attn}}} : \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}} \to \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}}$ such that for $n \in [N]$ and $Z \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$,

$$\mathsf{Attn}_{\theta_{\text{attn}}}(Z) := Z + \sum_{h=1}^H \mathsf{Softmax}\left(\frac{ZQ_hK_h^\top Z^\top}{\sqrt{d_{\text{model}}}}\right) ZV_h \in \mathbb{R}^{(n+1)\times d_{\text{model}}}.$$

If $Z \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$ has rows $Z_1^\top, \dots, Z_{n+1}^\top \in \mathbb{R}^{1\times d_{\text{model}}}$, then the ith row $[\mathsf{Attn}_{\theta_{\text{attn}}}(Z)]_{i,:}$ of $\mathsf{Attn}_{\theta_{\text{attn}}}(Z)$ can be written as

$$[\mathsf{Attn}_{\theta_{\text{attn}}}(Z)]_{i,:} = Z_i^\top + \sum_{h=1}^H \sum_{j=1}^{n+1} Z_j^\top V_h \cdot \frac{\exp(Z_i^\top Q_h K_h^\top Z_j/\sqrt{d_{\text{model}}})}{\sum_{\ell=1}^{n+1} \exp(Z_i^\top Q_h K_h^\top Z_\ell/\sqrt{d_{\text{model}}})} \in \mathbb{R}^{1\times d_{\text{model}}}.$$

Definition C.2 (FFN layer)**.** Let $d_{\text{ffn}} \in \mathbb{N}$, $W_1 \in \mathbb{R}^{d_{\text{model}}\times d_{\text{ffn}}}$, $W_2 \in \mathbb{R}^{d_{\text{ffn}}\times d_{\text{model}}}$, $v \in \mathbb{R}^{d_{\text{ffn}}}$ and let $\rho : \mathbb{R} \to \mathbb{R}$. The *FFN layer* with parameters $\theta_{\text{ffn}} := (W_1, W_2, v)$ (and activation function ρ) is the function $\mathsf{FFN}_{\theta_{\text{ffn}}} : \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}} \to \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}}$ such that for $n \in [N]$ and $Z \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$,

$$\mathsf{FFN}_{\theta_{\text{ffn}}}(Z) := Z + \rho(ZW_1 + 1_{n+1}v^\top)W_2 \in \mathbb{R}^{(n+1)\times d_{\text{model}}},$$

where ρ is applied entrywise.

If $Z \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$ has rows $Z_1^\top, \dots, Z_{n+1}^\top \in \mathbb{R}^{1\times d_{\text{model}}}$, then the ith row $[\mathsf{FFN}_{\theta_{\text{ffn}}}(Z)]_{i,:}$ of $\mathsf{FFN}_{\theta_{\text{ffn}}}(Z)$ can be written as

$$[\mathsf{FFN}_{\theta_{\text{ffn}}}(Z)]_{i,:} = \left(Z_i + W_2^\top \rho(W_1^\top Z_i + v)\right)^\top.$$

Thus the FFN layer applies a single-hidden layer neural network to each row of Z.

Now, a Transformer block is a composition of an attention layer and an FFN layer.

Definition C.3 (Transformer block)**.** Let θ_{attn} and θ_{ffn} be as in Definitions C.1 and C.2 respectively. The *Transformer block* with parameters $(\theta_{\text{attn}}, \theta_{\text{ffn}})$ is the function $\mathsf{TFBlock}_{(\theta_{\text{attn}},\theta_{\text{ffn}})} : \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}} \to \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}}$ given by

$$\mathsf{TFBlock}_{(\theta_{\text{attn}},\theta_{\text{ffn}})} := \mathsf{FFN}_{\theta_{\text{ffn}}} \circ \mathsf{Attn}_{\theta_{\text{attn}}}.$$

We say that the Transformer block $\mathsf{TFBlock}_{(\theta_{\text{attn}},\theta_{\text{ffn}})}$ has H heads, model dimension d_{model} and FFN width d_{ffn}.

Finally, a Transformer is a composition of Transformer blocks.

Definition C.4 (Transformer)**.** Let $\mathcal{F}_{\text{TF}}(L, H, d_{\text{model}}, d_{\text{ffn}})$ be the set of all *Transformers* $\mathsf{TF} : \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}} \to \bigcup_{n=1}^N \mathbb{R}^{(n+1)\times d_{\text{model}}}$ of the form

$$\mathsf{TF} = \mathsf{TFBlock}^{(L)} \circ \mathsf{TFBlock}^{(L-1)} \circ \cdots \circ \mathsf{TFBlock}^{(1)},$$

where $\mathsf{TFBlock}^{(\ell)} \equiv \mathsf{TFBlock}_{(\theta_{\text{attn}}^{(\ell)},\theta_{\text{ffn}}^{(\ell)})}$ is a Transformer block with H heads, model dimension d_{model} and FFN width d_{ffn} for all $\ell \in [L]$.

C.2 Proof of Theorem 3.1

We work under the assumptions of Section 3.2 and assume that ρ satisfies the conditions stated immediately before Theorem 3.1. For $g \in \mathcal{G}$, $(X, Y_g) \sim P_g$ and $z \in \mathbb{R}$, let $y \mapsto f_\xi(y, z)$ denote the conditional density of $Y_g \,|\, \{g(X) = z\}$ with respect to the dominating measure ν. Further let $\bar{f}_\xi \coloneqq f_\xi / R'$. By Bayes' theorem and (39), the posterior regression function g_π can be written as

$$g_\pi(x, D_n) = \frac{\int_{\mathcal{G}} g(x) \prod_{i=1}^n \bar{f}_\xi\big(y_i - g(x_i)\big) \,\mathrm{d}\widetilde{\pi}(g)}{\int_{\mathcal{G}} \prod_{i=1}^n \bar{f}_\xi\big(y_i - g(x_i)\big) \,\mathrm{d}\widetilde{\pi}(g)}. \tag{14}$$

Throughout this section, we assume that $\big((X_1, Y_1), \dots, (X_N, Y_N), (X, Y)\big) \,|\, P \sim P^{\otimes(N+1)}$ with $P \sim \pi$. For $n \in [N]$ and $d_{\mathrm{model}} \geq d + 2$, define

$$Z_{\mathrm{in}}^{(n, d_{\mathrm{model}})} \coloneqq \begin{pmatrix} X_1^\top & Y_1 & 0_{d_{\mathrm{model}}-d-2}^\top & 0 \\ \vdots & \vdots & \vdots & \vdots \\ X_n^\top & Y_n & 0_{d_{\mathrm{model}}-d-2}^\top & 0 \\ X^\top & 0 & 0_{d_{\mathrm{model}}-d-2}^\top & 1 \end{pmatrix} \in \mathbb{R}^{(n+1) \times d_{\mathrm{model}}}. \tag{15}$$

Lemma C.5. *For any $\epsilon, \delta > 0$, there exist $M \in \mathbb{N}$ and $g_1, \dots, g_M \in \mathcal{G}$ such that with probability at least $1 - \delta$, we have*

$$\max_{n \in [N]} \left| \frac{1}{M} \sum_{m=1}^M g_m(X) \prod_{i=1}^n \bar{f}_\xi\big(Y_i, g_m(X_i)\big) - \int_{\mathcal{G}} g(X) \prod_{i=1}^n \bar{f}_\xi\big(y_i, g(x_i)\big) \,\mathrm{d}\widetilde{\pi}(g) \right| \leq \epsilon$$

and

$$\max_{n \in [N]} \left| \frac{1}{M} \sum_{m=1}^M \prod_{i=1}^n \bar{f}_\xi\big(Y_i, g_m(X_i)\big) - \int_{\mathcal{G}} \prod_{i=1}^n \bar{f}_\xi\big(y_i, g(x_i)\big) \,\mathrm{d}\widetilde{\pi}(g) \right| \leq \epsilon.$$

Proof. Let $M \coloneqq 2\epsilon^{-2}(R^2 \vee 1) \log(2N/\delta)$ and let $\widetilde{g}_1, \dots, \widetilde{g}_M \overset{\mathrm{iid}}{\sim} \widetilde{\pi}$ be independent of $(X_i, Y_i)_{i=1}^N$ and (X, Y). For $n \in [N]$, define events

$$A_n \coloneqq \left\{ \left| \frac{1}{M} \sum_{m=1}^M \widetilde{g}_m(X) \prod_{i=1}^n \bar{f}_\xi\big(Y_i, \widetilde{g}_m(X_i)\big) - \int_{\mathcal{G}} g(X) \prod_{i=1}^n \bar{f}_\xi\big(y_i, g(x_i)\big) \,\mathrm{d}\widetilde{\pi}(g) \right| \leq \epsilon \right\}$$

and

$$B_n \coloneqq \left\{ \left| \frac{1}{M} \sum_{m=1}^M \prod_{i=1}^n \bar{f}_\xi\big(Y_i, \widetilde{g}_m(X_i)\big) - \int_{\mathcal{G}} \prod_{i=1}^n \bar{f}_\xi\big(y_i, g(x_i)\big) \,\mathrm{d}\widetilde{\pi}(g) \right| \leq \epsilon \right\}.$$

We have by Hoeffding's inequality that

$$\mathbb{P}(A_n^{\mathrm{c}} \,|\, X_1, Y_1, \dots, X_n, Y_n, X) \leq \frac{\delta}{2N} \quad \text{and} \quad \mathbb{P}(B_n^{\mathrm{c}} \,|\, X_1, Y_1, \dots, X_n, Y_n, X) \leq \frac{\delta}{2N}.$$

Then by a union bound,

$$\begin{aligned} \mathbb{E}\left\{ \mathbb{P}\left(\bigcup_{n=1}^N (A_n^{\mathrm{c}} \cup B_n^{\mathrm{c}}) \,\middle|\, \widetilde{g}_1, \dots, \widetilde{g}_M \right) \right\} &= \mathbb{P}\left(\bigcup_{n=1}^N (A_n^{\mathrm{c}} \cup B_n^{\mathrm{c}}) \right) \\ &= \mathbb{E}\left\{ \mathbb{P}\left(\bigcup_{n=1}^N (A_n^{\mathrm{c}} \cup B_n^{\mathrm{c}}) \,\middle|\, X_1, Y_1, \dots, X_N, Y_N, X, Y \right) \right\} \leq \delta. \end{aligned}$$

Therefore,

$$\mathbb{P}\left\{ \mathbb{P}\left(\bigcup_{n=1}^N (A_n^{\mathrm{c}} \cup B_n^{\mathrm{c}}) \;\middle|\; \widetilde{g}_1, \dots, \widetilde{g}_M \right) \leq \delta \right\} > 0.$$

This proves the claim. □

Lemma C.6. *For all $n \in [N]$, we have*

$$\int_{\mathcal{G}} \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big)\, \mathrm{d}\widetilde{\pi}(g) > 0$$

almost surely.

Proof. The density (with respect to ν) of $(Y_1, \ldots, Y_n)$ conditional on $(X_1, \ldots, X_n)$ is given by

$$(y_1, \ldots, y_n) \mapsto \int_{\mathcal{G}} \prod_{i=1}^{n} f_\xi\big(y_i, g(X_i)\big)\, \mathrm{d}\widetilde{\pi}(g).$$

Hence, by definition,

$$\mathbb{P}\bigg(\int_{\mathcal{G}} \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big)\, \mathrm{d}\widetilde{\pi}(g) = 0\bigg) = \mathbb{E}\bigg\{\mathbb{P}\bigg(\int_{\mathcal{G}} \prod_{i=1}^{n} f_\xi\big(y_i, g(X_i)\big)\, \mathrm{d}\widetilde{\pi}(g) = 0 \;\bigg|\; X_1, \ldots, X_n\bigg)\bigg\} = 0.$$

This holds for all $n \in [N]$, which proves the claim. □

Lemma C.7. *Let $d_1, d_2 \in \mathbb{N}$, let $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be Borel measurable and let μ be a Borel probability measure on $\mathbb{R}^{d_1}$. Then for every $\epsilon > 0$ and $\delta \in (0, 1)$, there exist $d' \in \mathbb{N}$, $W_1 \in \mathbb{R}^{d_1 \times d'}$, $W_2 \in \mathbb{R}^{d' \times d_2}$ and $v \in \mathbb{R}^{d'}$ such that*

$$\mu\Big(\big\{x \in \mathbb{R}^{d_1} : \big\|f(x) - W_2^\top \rho(W_1^\top x + v)\big\|_\infty \leq \epsilon\big\}\Big) \geq 1 - \delta.$$

Proof. We first prove the claim for $d_2 = 1$. Let $K_0 \subseteq \mathbb{R}^{d_1}$ be a compact set such that $\mu(K_0) \geq 1 - \delta/3$. By Lusin's theorem (Folland, 1999, Theorems 7.8 and 7.10), there exists a (compactly supported) continuous function $h : \mathbb{R}^{d_1} \to \mathbb{R}$ such that $\mu\big(\{x \in K_0 : f(x) \neq h(x)\}\big) \leq \delta/3$. Since one-hidden layer neural networks with non-polynomial activation are dense in the set of continuous functions on compact domains equipped with the uniform norm (Leshno et al., 1993, Theorem 1), there exist $d' \in \mathbb{N}$, $W_1 \in \mathbb{R}^{d_1 \times d'}$, $W_2 \in \mathbb{R}^{d' \times d_2}$ and $v \in \mathbb{R}^{d'}$ such that for all $x \in K_0$,

$$\big|h(x) - W_2^\top \rho(W_1^\top x + v)\big| \leq \frac{\delta\epsilon}{3}.$$

It follows by Markov's inequality that

$$\begin{aligned}
&\mu\Big(\big\{x \in \mathbb{R}^{d_1} : \big|f(x) - W_2^\top \rho(W_1^\top x + v)\big| \geq \epsilon\big\}\Big) \\
&\qquad \leq \mu\Big(\big\{x \in K_0 : \big|f(x) - W_2^\top \rho(W_1^\top x + v)\big| \geq \epsilon\big\}\Big) + \frac{\delta}{3} \\
&\qquad \leq \mu\Big(\big\{x \in K_0 : \big|h(x) - W_2^\top \rho(W_1^\top x + v)\big| \geq \epsilon\big\}\Big) + \frac{\delta}{3} + \frac{\delta}{3} \leq \delta.
\end{aligned}$$

This proves the claim when $d_2 = 1$. For $d_2 > 1$, we may approximate the component functions in parallel and apply a union bound to deduce the final result. □

Lemma C.8. *Let $d_{\text{model}} \geq d + 2$ and let $f_1 : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d_{\text{model}}}$ and $f_2 : \mathbb{R}^d \to \mathbb{R}^{d_{\text{model}}}$ be Borel measurable. For $n \in [N]$, let $Z_{\text{in}}^{(n, d_{\text{model}})} \in \mathbb{R}^{(n+1) \times d_{\text{model}}}$ be defined as in (15) and let*

$$U^{(n)} := \begin{pmatrix} f_1^\top(X_1, Y_1) \\ \vdots \\ f_1^\top(X_n, Y_n) \\ f_2^\top(X) \end{pmatrix} \in \mathbb{R}^{(n+1) \times d_{\text{model}}}.$$

Then for $\delta \in (0, 1)$ and $\epsilon > 0$, there exists an FFN layer $\mathsf{FFN}_{\theta_{\text{ffn}}}$ such that with probability at least $1 - \delta$,

$$\max_{n \in [N]} \big\|\mathsf{FFN}_{\theta_{\text{ffn}}}(Z_{\text{in}}^{(n, d_{\text{model}})}) - U^{(n)}\big\|_\infty \leq \epsilon.$$

Proof. Define a Borel measurable function $f : \mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_{\text{model}}}$ by $f(x, y, 0_{d_{\text{model}}-d-2}, 0) \coloneqq f_1(x, y) - (x, y, 0_{d_{\text{model}}-d-2}, 0)$ and $f(x, 0, 0_{d_{\text{model}}-d-2}, 1) \coloneqq f_2(x) - (x, 0, 0_{d_{\text{model}}-d-2}, 1)$ for $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ and set f to be zero otherwise. Let μ_1 denote the distribution of (X, Y) and let μ_2 denote the marginal distribution of X. Define a distribution μ on $\mathbb{R}^{d_{\text{model}}}$ by

$$\mu\Big(\prod_{j=1}^{d_{\text{model}}} A_j\Big) \coloneqq \frac{\mu_1(\prod_{j=1}^{d+1} A_j)\mathbb{1}_{\{0\in A_{d_{\text{model}}}\}} + \mu_2(\prod_{j=1}^{d} A_j)\mathbb{1}_{\{1\in A_{d_{\text{model}}}\}}}{2} \cdot \prod_{j=d+2}^{d_{\text{model}}-1} \mathbb{1}_{\{0\in A_j\}},$$

for all measurable sets $A_1, \ldots, A_{d_{\text{model}}} \subseteq \mathbb{R}$. By Lemma C.7 and the definition of μ above, there exists an FFN layer $\mathsf{FFN}_{\theta_{\text{ffn}}}$ such that

$$\mathbb{P}\Big(\big\|\mathsf{FFN}_{\theta_{\text{ffn}}}(X, Y, 0_{d_{\text{model}}-d-2}, 0) - f_1(X, Y)\big\|_\infty > \epsilon\Big) + \mathbb{P}\Big(\big\|\mathsf{FFN}_{\theta_{\text{ffn}}}(X, 0, 0_{d_{\text{model}}-d-2}, 1) - f_2(X)\big\|_\infty > \epsilon\Big) \le \frac{\delta}{N+1}.$$

We conclude the claim by a union bound. □

Proof of Theorem 3.1. It suffices to prove the theorem with $L = 3$ and $H = 1$. Let $\delta \coloneqq \frac{7\epsilon}{80R^2} \wedge 1$. By Lemma C.6 and a union bound, there exists $b \in (0, 1]$ such that the event

$$\mathcal{E}_0 \coloneqq \Big\{\min_{n\in[N]} \int_{\mathcal{G}} \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big)\, \mathrm{d}\widetilde{\pi}(g) \ge b\Big\}$$

has probability at least $1 - \delta$. Let $\epsilon_1 \coloneqq \frac{b}{8} \wedge \frac{b\sqrt{\epsilon}}{64(R\vee 1)}$. By Lemma C.5, there exist $M \ge d$ and $g_1, \ldots, g_M \in \mathcal{G}$ such that the event

$$\mathcal{E}_1 \coloneqq \Big\{\max_{n\in[N]}\Big|\frac{1}{M}\sum_{m=1}^{M} g_m(X)\prod_{i=1}^{n}\bar{f}_\xi\big(Y_i, g_m(X_i)\big) - \int_{\mathcal{G}} g(X)\prod_{i=1}^{n}\bar{f}_\xi\big(Y_i, g(X_i)\big)\,\mathrm{d}\widetilde{\pi}(g)\Big| \le \epsilon_1\Big\} \bigcap \Big\{\max_{n\in[N]}\Big|\frac{1}{M}\sum_{m=1}^{M}\prod_{i=1}^{n}\bar{f}_\xi\big(Y_i, g_m(X_i)\big) - \int_{\mathcal{G}}\prod_{i=1}^{n}\bar{f}_\xi\big(Y_i, g(X_i)\big)\,\mathrm{d}\widetilde{\pi}(g)\Big| \le \epsilon_1\Big\}$$

has probability at least $1 - \delta$. Now let $d_{\text{model}} \coloneqq 4M + 1$ and define $f_1 = (f_{1,1}, \ldots, f_{1,d_{\text{model}}}) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d_{\text{model}}}$ by

$$f_{1,j}(x, y) \coloneqq \begin{cases} \log \bar{f}_\xi\big(y, g_j(x)\big) & \text{if } j \in [M] \\ \log \bar{f}_\xi\big(y, g_{j-M}(x)\big) & \text{if } j \in [M+1 : 2M] \\ 0 & \text{if } j \in [2M+1 : 4M+1]. \end{cases}$$

Also define $f_2 = (f_{2,1}, \ldots, f_{2,d_{\text{model}}}) : \mathbb{R}^d \to \mathbb{R}^{d_{\text{model}}}$ by

$$f_{2,j}(x) \coloneqq \begin{cases} \log g_j(x) & \text{if } j \in [M] \\ 0 & \text{if } j \in [M+1 : 4M] \\ 1 & \text{if } j = 4M+1. \end{cases}$$

Finally, for $n \in [N]$, define $U^{(n)} \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$ by

$$U^{(n)} \coloneqq \begin{pmatrix} f_1(X_1, Y_1) \\ \vdots \\ f_1(X_n, Y_n) \\ f_2(X) \end{pmatrix}.$$

By Lemma C.8, there exists an FFN layer $\mathsf{FFN}_{\theta^{(1)}_{\text{ffn}}}$ with $\theta^{(1)}_{\text{ffn}} = (W_1, W_2, v)$, such that writing $\epsilon_2 \coloneqq \frac{b}{32(N+1)} \wedge \frac{b\sqrt{\epsilon}}{256R(N+1)}$, the event

$$\mathcal{E}_2 \coloneqq \Big\{\max_{n\in[N]}\big\|\mathsf{FFN}_{\theta^{(1)}_{\text{ffn}}}(Z^{(n,d_{\text{model}})}_{\text{in}}) - U^{(n)}\big\|_\infty \le \epsilon_2\Big\}$$

has probability at least $1-\delta$. Since the final column of W_2 only affects the final column of $\mathsf{FFN}_{\theta_{\text{ffn}}^{(1)}}(Z_{\text{in}}^{(n,d_{\text{model}})})$, we can choose the final column of W_2 to be zero, so that

$$\big[\mathsf{FFN}_{\theta_{\text{ffn}}^{(1)}}(Z_{\text{in}}^{(n,d_{\text{model}})})\big]_{i,d_{\text{model}}} = \big[Z_{\text{in}}^{(n,d_{\text{model}})}\big]_{i,d_{\text{model}}} = \mathbb{1}_{\{i=n+1\}} = \big[U^{(n)}\big]_{i,d_{\text{model}}} \tag{16}$$

for $i \in [n+1]$. Next, we define $Q^{(2)} = K^{(2)} \coloneqq 0_{d_{\text{model}} \times d_{\text{model}}}$ and define $V^{(2)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ by $V^{(2)}_{j,2M+j} \coloneqq 1$ for $j \in [2M]$, $V^{(2)}_{4M+1,4M+1} \coloneqq 1$ and zero otherwise. Let $\theta^{(2)}_{\text{attn}} \coloneqq (Q^{(2)}, K^{(2)}, V^{(2)})$, $\widetilde{Z}^{(n)} \coloneqq \mathsf{FFN}_{\theta_{\text{ffn}}^{(1)}}(Z_{\text{in}}^{(n,d_{\text{model}})}) \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$ and $\bar{Z}^{(n)} \coloneqq \mathsf{Attn}_{\theta^{(2)}_{\text{attn}}}(\widetilde{Z}^{(n)}) \in \mathbb{R}^{(n+1)\times d_{\text{model}}}$. Then by construction and (16), the $(n+1)$th row of $\bar{Z}^{(n)}$ is given by

$$\bar{Z}^{(n)}_{n+1,j} = \begin{cases} \widetilde{Z}^{(n)}_{n+1,j} & \text{if } j \in [2M] \\ \widetilde{Z}^{(n)}_{n+1,j} + \frac{1}{n+1}\sum_{i=1}^{n+1} \widetilde{Z}^{(n)}_{i,j-2M} & \text{if } j \in [2M+1:4M] \\ 1 + \frac{1}{n+1} & \text{if } j = 4M+1. \end{cases} \tag{17}$$

Thus, for $n \in [N]$ and $m \in [M]$, we have that on $\mathcal{E}_2$,

$$\begin{aligned} &\left|\bar{Z}^{(n)}_{n+1,2M+m} - \frac{1}{n+1}\sum_{i=1}^{n} \log \bar{f}_\xi\big(Y_i, g_m(X_i)\big) - \frac{1}{n+1}\log g_m(X)\right| \\ &\qquad = \left|\widetilde{Z}^{(n)}_{n+1,2M+m} + \frac{1}{n+1}\sum_{i=1}^{n+1} \widetilde{Z}^{(n)}_{i,m} - \frac{1}{n+1}\sum_{i=1}^{n+1} U^{(n)}_{i,m}\right| \\ &\qquad \le \big|\widetilde{Z}^{(n)}_{n+1,2M+m} - 0\big| + \frac{1}{n+1}\sum_{i=1}^{n+1}\big|\widetilde{Z}^{(n)}_{i,m} - U^{(n)}_{i,m}\big| \le 2\epsilon_2. \end{aligned} \tag{18}$$

Similarly, for $n \in [N]$ and $m \in [M]$, we have that on $\mathcal{E}_2$,

$$\left|\bar{Z}^{(n)}_{n+1,3M+m} - \frac{1}{n+1}\sum_{i=1}^{n} \log \bar{f}_\xi\big(Y_i, g_m(X_i)\big)\right| \le 2\epsilon_2. \tag{19}$$

Now define a continuous function $f_3 : \mathbb{R}^{4M+1} \to \mathbb{R}$ by

$$f_3(z_1, \ldots, z_{4M+1}) \coloneqq \frac{\sum_{m=1}^{M} \exp\{z_{2M+m}/(z_{4M+1}-1)\}}{\sum_{m=1}^{M} \exp\{z_{3M+m}/(z_{4M+1}-1)\}}.$$

For $n \in [N]$, let $\bar{P}^{(n)}$ be the distribution of the last row of $\bar{Z}^{(n)}$. By applying Lemma C.7 with $\mu = \frac{1}{N}\sum_{n=1}^{N} \bar{P}^{(n)}$ therein, there exists an FFN layer $\mathsf{FFN}_{\theta^{(2)}_{\text{ffn}}}$ such that the event

$$\mathcal{E}_3 \coloneqq \left\{\max_{n\in[N]}\left|\big[\mathsf{FFN}_{\theta^{(2)}_{\text{ffn}}}(\bar{Z}^{(n)})\big]_{n+1,d+1} - f_3\big(\bar{Z}^{(n)}_{n+1,1}, \ldots, \bar{Z}^{(n)}_{n+1,d_{\text{model}}}\big)\right| \le \sqrt{\epsilon}/4\right\}$$

has probability at least $1-\delta$. Since $\|g_m\|_\infty \le R$ and $\|\bar{f}_\xi\|_\infty \le 1$, we have by (17) and (18) that on $\mathcal{E}_1 \cap \mathcal{E}_2$, for all $n \in [N]$,

$$\begin{aligned} &\left|\frac{1}{M}\sum_{m=1}^{M} \exp\big\{\bar{Z}^{(n)}_{n+1,2M+m}/(\bar{Z}^{(n)}_{n+1,4M+1}-1)\big\} - \int_{\mathcal{G}} g(X)\prod_{i=1}^{n}\bar{f}_\xi\big(y_i, g(x_i)\big)\,\mathrm{d}\widetilde{\pi}(g)\right| \\ &\le \left|\frac{1}{M}\sum_{m=1}^{M} \exp\big\{\bar{Z}^{(n)}_{n+1,2M+m}/(\bar{Z}^{(n)}_{n+1,4M+1}-1)\big\} - \frac{1}{M}\sum_{m=1}^{M} g_m(X)\prod_{i=1}^{n}\bar{f}_\xi\big(Y_i, g_m(X_i)\big)\right| + \epsilon_1 \\ &\le \frac{1}{M}\sum_{m\in[M]}\left|\exp\big\{(n+1)\bar{Z}^{(n)}_{n+1,2M+m}\big\} - g_m(X)\prod_{i=1}^{n}\bar{f}_\xi\big(Y_i, g_m(X_i)\big)\right| + \epsilon_1 \\ &\le R(e^{2(N+1)\epsilon_2} - 1) + \epsilon_1 \le \frac{b\sqrt{\epsilon}}{32} =: \epsilon_3. \end{aligned} \tag{20}$$

Similarly, on $\mathcal{E}_1 \cap \mathcal{E}_2$ we have by (17) and (19) that for all $n \in [N]$,

$$\left| \frac{1}{M} \sum_{m=1}^{M} \exp\{\bar{Z}^{(n)}_{n+1,3M+m} / (\bar{Z}^{(n)}_{n+1,4M+1} - 1)\} - \int_{\mathcal{G}} \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big) \, \mathrm{d}\widetilde{\pi}(g) \right| \\ \leq (e^{2(N+1)\epsilon_2} - 1) + \epsilon_1 \leq \frac{b\sqrt{\epsilon}}{32R} \wedge \frac{b}{4} =: \epsilon_4. \tag{21}$$

Now, on the event $\mathcal{E}_0 \cap \mathcal{E}_1$, since $\epsilon_1 \leq b/2$, we have that

$$\min_{n \in [N]} \frac{1}{M} \sum_{m=1}^{M} \prod_{i=1}^{n} \bar{f}_\xi\big(Y_i, g_m(X_i)\big) \geq \frac{b}{2}. \tag{22}$$

Therefore, on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$, we have by (20), (21), (22) and (14) that for all $n \in [N]$

$$\begin{aligned} &\big| f_3(\bar{Z}^{(n)}_{n+1,1}, \ldots, \bar{Z}^{(n)}_{n+1,d_{\mathrm{model}}}) - g_\pi(X, \mathcal{D}_n) \big| \\ &= \left| \frac{\sum_{m=1}^{M} \exp\{\bar{Z}^{(n)}_{n+1,2M+m} / (\bar{Z}^{(n)}_{n+1,4M+1} - 1)\}}{\sum_{m=1}^{M} \exp\{\bar{Z}^{(n)}_{n+1,3M+m} / (\bar{Z}^{(n)}_{n+1,4M+1} - 1)\}} - \frac{\int_{\mathcal{G}} g(X) \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big) \, \mathrm{d}\widetilde{\pi}(g)}{\int_{\mathcal{G}} \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big) \, \mathrm{d}\widetilde{\pi}(g)} \right| \\ &\leq \left(\frac{1}{1 - 2\epsilon_4/b} - 1 \right) \left| \frac{\int_{\mathcal{G}} g(X) \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big) \, \mathrm{d}\widetilde{\pi}(g)}{\int_{\mathcal{G}} \prod_{i=1}^{n} \bar{f}_\xi\big(y_i, g(x_i)\big) \, \mathrm{d}\widetilde{\pi}(g)} \right| + \frac{\epsilon_3}{b/2 - \epsilon_4} \\ &\leq \left(\frac{1}{1 - 2\epsilon_4/b} - 1 \right) R + \frac{2\epsilon_3}{b - 2\epsilon_4} \leq \frac{4R\epsilon_4}{b} + \frac{4\epsilon_3}{b} \leq \frac{\sqrt{\epsilon}}{4}, \end{aligned} \tag{23}$$

where in the penultimate inequality, we used the fact that $\frac{1}{1-x} - 1 \leq 2x$ for $x \in [0, 1/2]$ and that $\epsilon_4 \leq b/4$, and the final inequality uses the definitions of ϵ_3 and ϵ_4. If we set the first attention layer $\mathsf{Attn}_{\theta^{(1)}_{\mathrm{attn}}}$ to be the identity map (i.e. set the parameters of the attention layer to zero), and let $\check{Z}^{(n)} := \mathsf{FFN}_{\theta^{(2)}_{\mathrm{ffn}}} \circ \mathsf{Attn}_{\theta^{(2)}_{\mathrm{attn}}} \circ \mathsf{FFN}_{\theta^{(1)}_{\mathrm{ffn}}} \circ \mathsf{Attn}_{\theta^{(1)}_{\mathrm{attn}}}(Z_{\mathrm{in}}^{(n,d_{\mathrm{model}})})$, then by (23) and the triangle inequality, we have on $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ that for all $n \in [N]$

$$\left| \check{Z}^{(n)}_{n+1,d+1} - g_\pi(X, \mathcal{D}_n) \right| = \left| \big[\mathsf{FFN}_{\theta^{(2)}_{\mathrm{ffn}}}(\bar{Z}^{(n)})\big]_{n+1,d+1} - g_\pi(X, \mathcal{D}_n) \right| \leq \frac{\sqrt{\epsilon}}{2}. \tag{24}$$

For $n \in [N]$, let $\check{P}^{(n)}$ be the distribution of the last row of $\check{Z}^{(n)}$. By applying Lemma C.7 with $\mu = \frac{1}{N} \sum_{n=1}^{N} \check{P}^{(n)}$ therein, there exists an FFN layer $\mathsf{FFN}_{\theta^{(3)}_{\mathrm{ffn}}}$ such that the event

$$\mathcal{E}_4 := \left\{ \max_{n \in [N]} \left| \big[\mathsf{FFN}_{\theta^{(3)}_{\mathrm{ffn}}}(\check{Z}^{(n)})\big]_{n+1,d+1} - \mathsf{clip}_R(\check{Z}^{(n)}_{n+1,d+1}) \right| \leq \sqrt{\epsilon}/4 \right\}$$

has probability at least $1 - \delta$. Finally, let $\mathsf{Attn}_{\theta^{(3)}_{\mathrm{attn}}}$ be the identity map and let

$$\mathsf{TF} := \mathsf{FFN}_{\theta^{(3)}_{\mathrm{ffn}}} \circ \mathsf{Attn}_{\theta^{(3)}_{\mathrm{attn}}} \circ \mathsf{FFN}_{\theta^{(2)}_{\mathrm{ffn}}} \circ \mathsf{Attn}_{\theta^{(2)}_{\mathrm{attn}}} \circ \mathsf{FFN}_{\theta^{(1)}_{\mathrm{ffn}}} \circ \mathsf{Attn}_{\theta^{(1)}_{\mathrm{attn}}}.$$

Since $\|g_\pi\|_\infty \leq R$, we have by (24) that on $\mathcal{E} := \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$, for all $n \in [N]$,

$$\begin{aligned} &\big| \mathsf{Read} \circ \mathsf{TF}(Z_{\mathrm{in}}^{(n,d_{\mathrm{model}})}) - g_\pi(X, \mathcal{D}_n) \big| \\ &\leq \left| \big[\mathsf{FFN}_{\theta^{(3)}_{\mathrm{ffn}}}(\check{Z}^{(n)})\big]_{n+1,d+1} - \mathsf{clip}_R(\check{Z}^{(n)}_{n+1,d+1}) \right| + \left| \mathsf{clip}_R(\check{Z}^{(n)}_{n+1,d+1}) - g_\pi(X, \mathcal{D}_n) \right| \leq \frac{3\sqrt{\epsilon}}{4}. \end{aligned}$$

By a union bound, the event $\mathcal{E}$ has probability at least $1 - 5\delta$. We conclude that

$$\max_{n \in [N]} \mathbb{E}_{P \sim \pi, (\mathcal{D}_n, X, Y) | P \sim P^{\otimes(n+1)}} \left\{ \big(\mathsf{Read} \circ \mathsf{TF}(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n) \big)^2 \right\} \leq \frac{9\epsilon}{16} + 5R^2\delta \leq \epsilon,$$

as required. □

C.3 Proof of Proposition 3.2

The main idea behind the proof of Proposition 3.2 is to use empirical process theory to upper bound $\mathcal{E}(\widetilde{f}_T)$ by the approximation error and a complexity measure of $\mathcal{F}$. In the literature, the complexity measure is often taken to be the covering number with respect to the uniform norm, and upper bounds on this covering number for Transformers are available, provided the input matrix and parameters are bounded. There, the proof strategy is to argue that for any input matrix from a bounded set, the output of the Transformer with bounded parameters is uniformly Lipschitz in its parameters. However, this argument does not work in our setting, since our input matrix and parameters are unbounded. Instead, we use the pseudo-dimension of $\mathcal{F}$ (see the definition below) as a complexity measure and show that the pseudo-dimension of Transformers is finite. This result (Lemma C.12 below) may be of independent interest.

Definition C.9 (VC-dimension and pseudo-dimension)**.** Let X be a set and let $\mathcal{C}$ be a collection of subsets of X. For $S \coloneqq \{x_1, \ldots, x_n\} \subseteq X$, we say that $\mathcal{C}$ shatters S if for every $T \subseteq S$, there exists $C \in \mathcal{C}$ such that $C \cap S = T$. The *VC-dimension* of $\mathcal{C}$ is the cardinality of the largest subset of X that can be shattered by $\mathcal{C}$.

Now let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$. The *pseudo-dimension* $\mathrm{Pdim}(\mathcal{G})$ of $\mathcal{G}$ is the VC-dimension of its subgraph class $\big\{\{(x, y) \in \mathcal{X} \times \mathbb{R} : y \leq g(x)\} : g \in \mathcal{G}\big\}$.

Lemma C.10 below controls $\mathcal{E}(\widetilde{f}_T)$ in terms of the approximation error and pseudo-dimension of $\mathcal{F}$.

Lemma C.10. *Under the setting of Proposition 3.2, there exists $C > 0$, depending only on $\|\xi\|_{\psi_2}$ and R, such that for $T \geq 2$,*

$$\mathcal{E}(\widetilde{f}_T) \leq 2 \inf_{f \in \mathcal{F}} \mathbb{E}_{P \sim \pi,\, (\mathcal{D}_n, X, Y) | P \sim P^{\otimes(n+1)}} \Big\{ \big(f(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n)\big)^2 \Big\} + \frac{C \log^3(T) \cdot \mathrm{Pdim}(\mathcal{F})}{T}.$$

Proof. Let $P \sim \pi$, $(\mathcal{D}_n, X, Y) | P \sim P^{\otimes(n+1)}$ and let $(\mathcal{D}_n^{(t)}, X^{(t)}, Y^{(t)})_{t \in [T]}$ be independent copies of $(\mathcal{D}_n, X, Y)$. For $t \in [T]$, let $Z^{(t)} \coloneqq (X^{(t)}, \mathcal{D}_n^{(t)})$ and let $Z \coloneqq (X, \mathcal{D}_n)$. Now note that $g_\pi(Z) = \mathbb{E}(Y \mid Z)$,

$$\widehat{f}_T \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \big(Y^{(t)} - f(Z^{(t)})\big)^2$$

and

$$\mathcal{E}(\widetilde{f}_T) = \mathbb{E}\big\{\big(Y - \mathsf{clip}_R \widehat{f}_T(Z)\big)^2\big\} - \mathbb{E}\big\{\big(Y - g_\pi(Z)\big)^2\big\}.$$

We have therefore reduced our problem to bounding the excess risk of a nonparametric regression estimator in a problem with bounded regression function and sub-Gaussian noise. The conclusion now follows by applying, for example, Ma, Wang and Samworth (2025, Theorem 1) with the following minor modifications: (i) we ignore their Ω vectors since our data do not have missing values, (ii) by inspecting their proof, we may truncate $\widehat{f}_T$ at level R since this is a known bound on the regression function, (iii) the optimisation error term (the first term in their upper bound) is zero since we consider the empirical risk minimiser, and (iv) by inspecting their proof and using Györfi et al. (2002, Theorem 9.4), we can replace the third term in their upper bound by $\frac{C \log^3(T) \cdot \mathrm{Pdim}(\mathcal{F})}{T}$, where $C > 0$ depends only on $\|\xi\|_{\psi_2}$ and R. □

Next we will show that the class of Transformers with fixed architecture has finite pseudo-dimension. The proof uses terminology and results from model theory, a branch of mathematical logic. Since this field will probably be unfamiliar to most readers, we now provide a minimal glossary (sufficient for our proof) of the relevant terms; we refer the readers to Marker (2002) for a formal introduction to model theory. We will be concerned with an object called a *structure*, which is a set equipped with a collection of distinguished functions, relations and elements that together give meaning to the symbols of a formal language. Our structure of interest is $\mathbb{R}_{\exp,\Phi} \coloneqq (\mathbb{R}; 0, 1, +, -, \cdot, <, \exp, \Phi)$, where $\cdot$ refers to multiplication, $\exp : \mathbb{R} \to (0, \infty)$ refers to the exponential function and $\Phi : \mathbb{R} \to [0, 1]$ refers to the distribution function of the standard Gaussian distribution. Our goal is to show that Transformers can be defined within $\mathbb{R}_{\exp,\Phi}$, in a sense that we will make precise below.

Given a countable set $\mathcal{V}$ of variable symbols, which we intrepret as syntactic placeholders, the set of *terms* in the structure $\mathbb{R}_{\exp,\Phi} := (\mathbb{R}; 0, 1, +, -, \cdot, <, \exp, \Phi)$ is the smallest set $\mathcal{T}$ such that (i) $0, 1 \in \mathcal{T}$, (ii) $\mathcal{V} \subseteq \mathcal{T}$, (iii) if $t_1, t_2 \in \mathcal{T}$, then $t_1 + t_2$, $t_1 - t_2$, $t_1 \cdot t_2 \in \mathcal{T}$, and (iv) if $t \in \mathcal{T}$, then $\exp(t), \Phi(t) \in \mathcal{T}$. We say that φ is an *atomic formula* in $\mathbb{R}_{\exp,\Phi}$ if φ is either $t_1 = t_2$ or $t_1 < t_2$, for some $t_1, t_2 \in \mathcal{T}$. The set of *(first-order) formulas* $\mathcal{W}$ in (the language of) $\mathbb{R}_{\exp,\Phi}$ is the smallest set such that (i) every atomic formula is in $\mathcal{W}$, (ii) if $\varphi, \psi \in \mathcal{W}$, then $(\neg\varphi), (\varphi \vee \psi), (\varphi \wedge \psi) \in \mathcal{W}$, where the logic symbolds $\neg, \vee, \wedge$ stand for negation, disjunction and conjunction, and (iii) if $\varphi \in \mathcal{W}$ and $v \in \mathcal{V}$, then $\exists v\, \varphi$ and $\forall v\, \varphi$ are in $\mathcal{W}$. For a formula $\varphi \in \mathcal{W}$, we say that a variable v is *free* in φ if it is not inside a $\exists v$ or $\forall v$ quantifier; e.g. v_1 is free in the formula $(v_1 = 0) \vee (v_1 > 0)$, but it is not free in the formula $\exists v_1\ v_2 \cdot v_2 = v_1$. When a formula φ has free variables $v_1, \ldots, v_n$, we will often write $\varphi(v_1, \ldots, v_n)$ to make the free variables explicit. Now, letting $\varphi(v_1, \ldots, v_n) \in \mathcal{W}$ be a formula and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, we write $\mathbb{R}_{\exp,\Phi} \models \varphi(a)$ (we read '$\mathbb{R}_{\exp,\Phi}$ entails $\varphi(a)$' or '$\mathbb{R}_{\exp,\Phi}$ models $\varphi(a)$') if $\varphi(a)$ is true in $\mathbb{R}_{\exp,\Phi}$, and otherwise we write $\mathbb{R}_{\exp,\Phi} \not\models \varphi(a)$. For example, if $\varphi(v_1, v_2)$ is the formula $(v_1 = v_2) \vee (v_1 + v_2 = 0)$, then $\mathbb{R}_{\exp,\Phi} \models \varphi(1, 1)$ and $\mathbb{R}_{\exp,\Phi} \models \varphi(-1, 1)$, but $\mathbb{R}_{\exp,\Phi} \not\models \varphi(1, 2)$.

For $d \in \mathbb{N}$ and a set $A \subseteq \mathbb{R}^d$, we say that A is *definable* (in $\mathbb{R}_{\exp,\Phi}$) if there exist $m \in \mathbb{N}$, a formula $\varphi(v_1, \ldots, v_d, w_1, \ldots, w_m)$ in $\mathbb{R}_{\exp,\Phi}$ and $b \in \mathbb{R}^m$ such that $A = \{a \in \mathbb{R}^d : \mathbb{R}_{\exp,\Phi} \models \varphi(a, b)\}$; we will sometimes abbreviate the notation and write $A = \{a \in \mathbb{R}^d : \varphi(a, b)\}$ instead. For $d_1, d_2 \in \mathbb{N}$ and a function $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, we say that f is *definable* (in $\mathbb{R}_{\exp,\Phi}$) if its *graph* $\{(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} : y = f(x)\}$ is a definable set in $\mathbb{R}_{\exp,\Phi}$. Moreover, for a collection $\mathcal{C}$ of subsets of $\mathbb{R}^d$, we say that $\mathcal{C}$ is *uniformly definable* (in $\mathbb{R}_{\exp,\Phi}$) if there exists a formula $\varphi(v, w)$, where $v = (v_1, \ldots, v_d)$ and $w = (w_1, \ldots, w_m)$ for some $m \in \mathbb{N}$, such that for every $C \in \mathcal{C}$, there exists $b_C \in \mathbb{R}^m$ with $C = \{a \in \mathbb{R}^d : \varphi(a, b_C)\}$. Similarly, for a collection $\mathcal{F}$ of functions from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$, we say that $\mathcal{F}$ is *uniformly definable* (in $\mathbb{R}_{\exp,\Phi}$) if the collection of graphs of functions in $\mathcal{F}$ is uniformly definable.

Lemma C.11. *(a) Let $\mathcal{F}$ be a collection of uniformly definable functions from $\mathbb{R}^{d_1}$ to $\mathbb{R}$, and let $\mathcal{G}$ be a collection of uniformly definable functions from $\mathbb{R}^{d_2}$ to $\mathbb{R}$. Then $\big\{\{(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R} : y \leq f(x)\} : f \in \mathcal{F}\big\}$, $\{(x, z) \mapsto f(x) + g(z) : f \in \mathcal{F}, g \in \mathcal{G}\}$ and $\{(x, z) \mapsto f(x) \cdot g(z) : f \in \mathcal{F}, g \in \mathcal{G}\}$ are all uniformly definable.*

(b) Let $\mathcal{F}$ be a collection of uniformly definable functions from $\mathbb{R}^d$ to $\mathbb{R} \setminus \{0\}$. Then $\{x \mapsto 1/f(x) : f \in \mathcal{F}\}$ is uniformly definable.

(c) Let $p \in \mathbb{N}$ and for $j \in [p]$, let $\mathcal{G}_j$ be a collection of uniformly definable functions from $\mathbb{R}^{d_j}$ to $\mathbb{R}$ for some $d_j \in \mathbb{N}$. Further let $\mathcal{F}$ be a collection of uniformly definable functions from $\mathbb{R}^p$ to $\mathbb{R}$. Then $\big\{(x_1, \ldots, x_p) \mapsto f\big(g_1(x_1), \ldots, g_p(x_p)\big) : f \in \mathcal{F}, g_j \in \mathcal{G}_j\ \forall j \in [p]\big\}$ is uniformly definable.

Proof. (a) Suppose that for each $f \in \mathcal{F}$, the graph of f is $\{(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R} : \varphi(x, y, b_f)\}$ for some formula φ, some $b_f \in \mathbb{R}^m$ and some $m \in \mathbb{N}$; also suppose that for each $g \in \mathcal{G}$, the graph of g is $\{(z, y) \in \mathbb{R}^{d_2} \times \mathbb{R} : \psi(z, y, c_g)\}$ for some formula ψ, some $c_g \in \mathbb{R}^n$ and some $n \in \mathbb{N}$. The subgraph $\{(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R} : y \leq f(x)\}$ of $f \in \mathcal{F}$ can be written as $\big\{(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R} : \exists t\ \big(\varphi(x, t, b_f) \wedge (y < t \vee y = t)\big)\big\}$. Hence the collection of subgraphs of $f \in \mathcal{F}$ is uniformly definable. For $f \in \mathcal{F}$ and $g \in \mathcal{G}$, the graph of the function $(x, z) \mapsto f(x) + g(z)$ can be written as $\big\{(x, z, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R} : \exists u, v\ \big(\varphi(x, u, b_f) \wedge \psi(z, v, c_g) \wedge (y = u + v)\big)\big\}$. Finally, the graph of the function $(x, z) \mapsto f(x) \cdot g(z)$ can be written as $\big\{(x, z, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R} : \exists u, v\ \big(\varphi(x, u, b_f) \wedge \psi(z, v, c_g) \wedge (y = u \cdot v)\big)\big\}$.

(b) Again, suppose that for each $f \in \mathcal{F}$, the graph of f is $\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : \varphi(x, y, b_f)\}$ for some formula φ, some $b_f \in \mathbb{R}^m$ and some $m \in \mathbb{N}$. Then, for $f \in \mathcal{F}$, the graph of the function $x \mapsto 1/f(x)$ can be written as $\big\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : \exists u\ \big(\varphi(x, u, b_f) \wedge (y \cdot u = 1)\big)\big\}$. This proves the claim.

(c) Suppose that for each $f \in \mathcal{F}$, the graph of f is $\{(x, y) \in \mathbb{R}^p \times \mathbb{R} : \varphi(x, y, b_f)\}$ for some formula φ, some $b_f \in \mathbb{R}^m$ and some $m \in \mathbb{N}$; also suppose that for $j \in [p]$ and $g_j \in \mathcal{G}_j$, the graph of g_j is $\{(z, y) \in \mathbb{R}^{d_j} \times \mathbb{R} : \psi_j(z, y, c_{g_j})\}$ for some formula ψ_j, some $c_{g_j} \in \mathbb{R}^{n_j}$ and some $n_j \in \mathbb{N}$. Then

the graph of the function $(x_1, \dots, x_p) \mapsto f\big(g_1(x_1), \dots, g_p(x_p)\big)$ can be written as

$$\left\{(x_1, \dots, x_p, y) \in \mathbb{R}^{p+1} : \exists v_1, \dots, v_p \left(\bigwedge_{j=1}^{p} \psi_j(x_j, v_j, c_{g_j}) \wedge \varphi(v_1, \dots, v_p, y, b_f) \right) \right\}.$$

This proves the claim. □

Lemma C.12. *For any $L, H, d_{\text{model}}, d_{\text{ffn}} \in \mathbb{N}$, the class $\mathcal{F}' \coloneqq \mathsf{Read} \circ \mathcal{F}_{\text{TF}}(L, H, d_{\text{model}}, d_{\text{ffn}})$ has finite pseudo-dimension.*

Remark C.13. It is possible to obtain upper bounds on the pseudo-dimension of $\mathcal{F}'$ with polynomial dependence on L, H, d_{model} and d_{ffn} via a more detailed analysis, but a finite pseudo-dimension is already enough for our purpose.

Proof. Note that $\mathcal{F}'$ is defined on the space $\bigcup_{n=1}^{N} \mathbb{R}^{(n+1)\times d_{\text{model}}}$, equipped with the disjoint union topology. Thus, it suffices to show that, for each $n \in [N]$, the restriction $\mathcal{F}'_n$ of the functions in $\mathcal{F}'$ to $\mathbb{R}^{(n+1)\times d_{\text{model}}}$ has finite pseudo-dimension. Now let

$$\mathcal{C}_n \coloneqq \left\{ \big\{(Z, y) \in \mathbb{R}^{(n+1)\times d_{\text{model}}} \times \mathbb{R} : y \le f(Z)\big\} : f \in \mathcal{F}'_n \right\}$$

denote the collection of subgraphs of $f \in \mathcal{F}'_n$. Then by definition, the pseudo-dimension of $\mathcal{F}'_n$ is equal to the VC-dimension of $\mathcal{C}_n$. We will now show that $\mathcal{C}_n$ is uniformly definable in $\mathbb{R}_{\exp,\Phi}$. The graph $\{(x, y) \in \mathbb{R}^2 : y = \mathsf{ReLU}(x)\}$ of ReLU can be written as

$$\big\{(x, y) \in \mathbb{R}^2 : \big(\neg(x < 0) \wedge y = x\big) \vee (x < 0 \wedge y = 0)\big\},$$

so $\mathsf{ReLU} : \mathbb{R} \to \mathbb{R}$ is definable on $\mathbb{R}_{\exp,\Phi}$. By Lemma C.11, sums, products, reciprocals and compositions of definable functions are definable. Thus, for any $d, p \in \mathbb{N}$, the set of polynomials on $\mathbb{R}^d$ with degree at most p is uniformly definable in $\mathbb{R}_{\exp,\Phi}$. Moreover, since $\exp$ and Φ are definable in $\mathbb{R}_{\exp,\Phi}$ by definition, we deduce that GELU, SiLU, and $\mathsf{Softmax}$ are also definable in $\mathbb{R}_{\exp,\Phi}$. Each function $f \in \mathcal{F}'_n$ comprises compositions, sums and products of $\rho \in \{\mathsf{ReLU}, \mathsf{GELU}, \mathsf{SiLU}\}$, $\mathsf{Softmax}$ and polynomials of degree at most two, we deduce by Lemma C.11 that $\mathcal{F}'_n$ is uniformly definable in $\mathbb{R}_{\exp,\Phi}$. Further, by Lemma C.11*(a)*, we have that the subgraph class of $\mathcal{F}'_n$ is also uniformly definable on $\mathbb{R}_{\exp,\Phi}$ and

$$\mathcal{C}_n \subseteq \left\{ \big\{a \in \mathbb{R}^{(n+1)\times d_{\text{model}}} \times \mathbb{R} : \mathbb{R}_{\exp,\Phi} \models \varphi(a, b)\big\} : b \in \mathbb{R}^m \right\}.$$

By Marker (2002, Therorem 3.4.37), the structure $\mathbb{R}_{\exp} \coloneqq (\mathbb{R}; 0, 1, +, -, \cdot, <, \exp, \Phi)$ is *o-minimal*, i.e. every definable set is a finite union of open intervals and singleton points. Further, the derivative Φ' given by $\Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is definable in $\mathbb{R}_{\exp}$, so by Speissegger (1999), there exists an o-minimal extension $\widetilde{\mathbb{R}}$ of $\mathbb{R}_{\exp}$ in which Φ is definable. Now, every definable set in $\mathbb{R}_{\exp,\Phi}$ is definable in $\widetilde{\mathbb{R}}$, so $\mathbb{R}_{\exp,\Phi}$ is o-minimal. We may therefore apply Laskowski (1992, Theorem 3.2) (see also Karpinski and Macintyre, 1995, Theorem 2) to deduce that the VC-dimension of $\mathcal{C}_n$ is finite. □

Proof of Proposition 3.2. By Theorem 3.1, there exist $d^{\circ}_{\text{model}} \ge d + 2$ and $d^{\circ}_{\text{ffn}} \in \mathbb{N}$ such that if $d_{\text{model}} \ge d^{\circ}_{\text{model}}$, $d_{\text{ffn}} \ge d^{\circ}_{\text{ffn}}$ and $\mathcal{F} \coloneqq \mathsf{Read} \circ \mathcal{F}_{\text{TF}}(L, H, d_{\text{model}}, d_{\text{ffn}})$, then

$$\inf_{f \in \mathcal{F}} \mathbb{E}_{P\sim\pi,\,(\mathcal{D}_n, X, Y)|P\sim P^{\otimes(n+1)}} \left\{ \big(f(X, \mathcal{D}_n) - g_\pi(X, \mathcal{D}_n)\big)^2 \right\} \le \frac{\epsilon}{4},$$

Now by Lemma C.12, we have that $\text{Pdim}(\mathcal{F})$ is finite. Thus, for all T sufficiently large, we have $\frac{C \log^3(T) \cdot \text{Pdim}(\mathcal{F})}{T} \le \frac{\epsilon}{2}$, where $C > 0$ is taken from Lemma C.10. The claim then follows from Lemma C.10. □

D Function spaces and wavelets

D.1 Function spaces

In this section, we restrict the domain of the functions to $[0, 1]^d$. By Jensen's inequality, $\|f\|_{L^{p_1}} \le \|f\|_{L^{p_2}}$ for all $1 \le p_1 \le p_2 \le \infty$. For $f, g \in L^2([0, 1]^d)$, define the inner product

$$\langle f, g \rangle \coloneqq \int_{[0,1]^d} f(x) g(x) \, \mathrm{d}x.$$

Let $C([0,1]^d)$ denote the set of continuous functions from $[0,1]^d$ to $\mathbb{R}$. For $m = (m_1, \ldots, m_d)^\top \in \mathbb{N}_0^d$ with $k := \|m\|_1$, and a k-times differentiable $f : [0,1]^d \to \mathbb{R}$, we use the multi-index notation $\partial^m f := \partial^{m_1} \cdots \partial^{m_d} f$ for partial derivatives.

Definition D.1 (Hölder space)**.** For $\alpha > 0$, let $\alpha_0 := \lceil \alpha \rceil - 1$ be the largest integer strictly smaller than α and define the *Hölder space* $H^\alpha([0,1]^d)$ to be the set of α_0-times differentiable $f : [0,1]^d \to \mathbb{R}$ with

$$\sup_{x \neq y} \sum_{m \in \mathbb{N}_0^d : \|m\|_1 = \alpha_0} \frac{\partial^m f(x) - \partial^m f(y)}{\|x - y\|_2^{\alpha - \alpha_0}} < \infty.$$

For $r \in \mathbb{N}$, define the *rth order difference operator* $\Delta_h^r : L^1([0,1]^d) \to L^1([0,1]^d)$ by

$$\Delta_h^r(f)(x) = \mathbb{1}_{\{x + rh \in [0,1]^d\}} \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x + kh).$$

Note that $\Delta_h^1(f)(x) = \{f(x+h) - f(x)\}\mathbb{1}_{\{x+h \in [0,1]^d\}}$ and $\Delta_h^r(f)(x) = \Delta_h^1\big(\Delta_h^{r-1}(f)(x)\big)$ if $x + rh \in [0,1]^d$. Now, for $p \in [1, \infty]$, $f \in L^p([0,1]^d)$ and $t \geq 0$, define the *rth modulus of smoothness*

$$\omega_{r,p}(f, t) := \sup_{\|h\|_2 \leq t} \|\Delta_h^r(f)\|_{L^p}.$$

Definition D.2 (Besov space)**.** For $p, q \in [1, \infty]$ and $\alpha > 0$, let $r := \lfloor \alpha \rfloor + 1$ and define the *Besov space* $B_{p,q}^\alpha([0,1]^d)$ by

$$B_{p,q}^\alpha([0,1]^d) := \begin{cases} \left\{ f \in L^p([0,1]^d) : \left[\int_0^\infty \big(t^{-\alpha} \omega_{r,p}(f,t)\big)^q \frac{\mathrm{d}t}{t} \right]^{1/q} < \infty \right\} & \text{for } q \in [1, \infty) \\ \left\{ f \in L^p([0,1]^d) : \sup_{t > 0} t^{-\alpha} \omega_{r,p}(f,t) < \infty \right\} & \text{for } q = \infty. \end{cases}$$

For non-integer smoothness levels, Hölder spaces are special cases of Besov spaces in the sense that $H^\alpha([0,1]^d) = B_{\infty,\infty}^\alpha([0,1]^d)$ for all $\alpha \in (0, \infty) \setminus \mathbb{N}$, while for $\alpha \in \mathbb{N}$, we have $H^\alpha([0,1]^d) \subseteq B_{\infty,\infty}^\alpha([0,1]^d)$, and in fact, $B_{\infty,\infty}^\alpha([0,1]^d)$ coincides with Zygmund spaces when $\alpha \in \mathbb{N}$; see e.g. Giné and Nickl (2021, p. 351) or Triebel (1983, Chapter 2.5.7).

D.2 Wavelets

We start with the univariate setting where the notation is simpler. For $S \in \mathbb{N}$, denote the Cohen–Daubechies–Vial (CDV) wavelet basis on $[0,1]$ (Cohen, Daubechies and Vial, 1993), constructed from S-regular and S-times continuously differentiable father and mother Daubechies wavelets (Daubechies, 1988), by

$$\big\{\phi_k : k \in [0 : 2^{\ell_0} - 1]\big\} \bigcup \big\{\psi_{\ell,k} : \ell \in [\ell_0 : \infty),\, k \in [0 : 2^\ell - 1]\big\}, \tag{25}$$

where $\ell_0 \in \mathbb{N}$ depends only on S. The precise definition and construction of this wavelet basis is not important for us, but the functions in (25) are S-times continuously differentiable and form an orthonormal basis for $L^2([0,1])$ (Cohen, Daubechies and Vial, 1993, Theorem 4.4). Moreover, ϕ_k is supported on $\left[\frac{k}{2^{\ell_0}}, \frac{k+1}{2^{\ell_0}}\right]$ for $k \in [0 : 2^{\ell_0} - 1]$, and $\psi_{\ell,k}$ is supported on $\left[\frac{k}{2^\ell}, \frac{k+1}{2^\ell}\right]$ for $\ell \in [\ell_0 : \infty)$, $k \in [0 : 2^\ell - 1]$.

The one-dimensional CDV wavelets in (25) can be extended to an orthonormal basis for $L^2([0,1]^d)$ via tensor products. For $k = (k_1, \ldots, k_d)^\top \in [0 : 2^{\ell_0} - 1]^d$ and $x = (x_1, \ldots, x_d)^\top \in [0,1]^d$, define

$$\Phi_k(x) := \prod_{j=1}^d \phi_{k_j}(x_j). \tag{26}$$

Further, for $\tau = (\tau_1, \ldots, \tau_d)^\top \in \{0,1\}^d \setminus \{0_d\}$, $\ell \in [\ell_0 : \infty)$, $k = (k_1, \ldots, k_d)^\top \in [0 : 2^\ell - 1]^d$ and $x = (x_1, \ldots, x_d)^\top \in [0,1]^d$, define

$$\Psi_{\ell,k}^{(\tau)}(x) := \prod_{j=1}^d \psi_{\ell,k_j}^{(\tau_j)}(x_j), \tag{27}$$

where $\psi^{(1)}_{\ell,k'} = \psi_{\ell,k'}$ for all $k' \in [0 : 2^\ell - 1]$, and $(\psi^{(0)}_{\ell,k'})_{k'=0}^{2^\ell-1}$ are the boundary-corrected (S-regular and S-times continuously differentiable) father Daubechies wavelets at resolution ℓ. Again, the precise definition and construction are not important for us, but we will exploit the following properties. To simplify notation, define

$$K \coloneqq [0 : 2^{\ell_0} - 1]^d \quad \text{and} \quad \Gamma_\ell \coloneqq [0 : 2^\ell - 1]^d \times \big(\{0,1\}^d \setminus \{0_d\}\big) \text{ for } \ell \in [\ell_0 : \infty), \tag{28}$$

and for $\gamma = (k, \tau) \in \Gamma_\ell$, we write $\Psi_{\ell,\gamma} \coloneqq \Psi^{(\tau)}_{\ell,k}$. Note that

$$|K| = 2^{\ell_0 d} \quad \text{and} \quad |\Gamma_\ell| = 2^{\ell d}(2^d - 1) < 2^{(\ell+1)d} \text{ for } \ell \in [\ell_0 : \infty). \tag{29}$$

Each term in the products in (26) and (27) is S-times continuously differentiable, and the functions

$$\big\{\Phi_k : k \in K\big\} \cup \big\{\Psi_{\ell,\gamma} : \ell \in [\ell_0 : \infty),\, \gamma \in \Gamma_\ell\big\}$$

form an orthonormal basis for $L^2([0,1]^d)$; see e.g. Giné and Nickl (2021, Chapter 4.3.6) and Dahmen (1997, Section 4). By construction, there exist $C, C' > 0$ (depending only on S and d) such that

$$\|\Phi_k\|_\infty \le C 2^{\ell_0 d/2} \quad \text{and} \quad \|\Psi_{\ell,\gamma}\|_\infty \le C 2^{\ell d/2} \tag{30}$$

for $k \in K$, $\ell \in [\ell_0 : \infty)$ and $\gamma \in \Gamma_\ell$, and that

$$\sum_{k \in K} \mathbb{1}_{\{\Phi_k(x) \neq 0\}} \vee \sup_{\ell \in [\ell_0:\infty)} \sum_{\gamma \in \Gamma_\ell} \mathbb{1}_{\{\Psi_{\ell,\gamma}(x) \neq 0\}} \le C' \quad \text{for all } x \in [0,1]^d. \tag{31}$$

For $f \in L^2([0,1]^d)$, we have the wavelet decomposition

$$f = \sum_{k \in K} a_k \Phi_k + \sum_{\ell=\ell_0}^{\infty} \sum_{\gamma \in \Gamma_\ell} b_{\ell,\gamma} \Psi_{\ell,\gamma}, \tag{32}$$

where $a_k \coloneqq \langle f, \Phi_k \rangle$ and $b_{\ell,\gamma} \coloneqq \langle f, \Psi_{\ell,\gamma} \rangle$. Define

$$a \coloneqq (a_k)_{k \in K} \in \mathbb{R}^K \quad \text{and} \quad b_\ell \coloneqq (b_{\ell,\gamma})_{\gamma \in \Gamma_\ell} \in \mathbb{R}^{\Gamma_\ell} \text{ for } \ell \in [\ell_0 : \infty). \tag{33}$$

For $f \in L^2([0,1]^d)$, $p \in [2, \infty]$, $q \in [1, \infty]$ and $\alpha \in (0, S)$, define its Besov norm by

$$\|f\|_{B^\alpha_{p,q}} \coloneqq \begin{cases} 2^{\ell_0(d/2-d/p)}\|a\|_p + \Big\{\sum_{\ell=\ell_0}^{\infty} \big(2^{\ell(\alpha+d/2-d/p)}\|b_\ell\|_p\big)^q\Big\}^{1/q} & \text{for } q \in [1, \infty) \\ 2^{\ell_0(d/2-d/p)}\|a\|_p + \sup_{\ell \in [\ell_0:\infty)} 2^{\ell(\alpha+d/2-d/p)}\|b_\ell\|_p & \text{for } q = \infty. \end{cases}$$

For $p \in [2, \infty]$, $q \in [1, \infty]$ and $\alpha \in (0, S)$, the Besov space $B^\alpha_{p,q}([0,1]^d)$ in Definition D.2 can be equivalently characterised by wavelet coefficients in the sense that $f \in B^\alpha_{p,q}([0,1]^d)$ if and only if $f \in L^2([0,1]^d)$ and $\|f\|_{B^\alpha_{p,q}} < \infty$ (e.g. Giné and Nickl, 2021, Chapter 4.3.6). Therefore, we can define a closed ball $B^\alpha_{p,q}([0,1]^d, R)$ of radius $R > 0$ in the space $B^\alpha_{p,q}([0,1]^d)$ by

$$B^\alpha_{p,q}([0,1]^d, R) \coloneqq \{f \in L^2([0,1]^d) : \|f\|_{B^\alpha_{p,q}} \le R\}. \tag{34}$$

As a special case, we have $\|f\|_{B^\alpha_{\infty,\infty}} = 2^{\ell_0 d/2}\|a\|_\infty + \sup_{\ell \in [\ell_0:\infty)} 2^{\ell(\alpha+d/2)}\|b_\ell\|_\infty$.

E Posterior contraction and adaptation

E.1 General theory

This section is primarily based on material from Ghosal, Lember and van der Vaart (2008); see also Ghosal and van der Vaart (2017, Section 10.2). Here, we make some modifications in order to obtain finite-sample posterior contraction at an exponential rate.

Let $\mathcal{Z}$ be a measurable space and let $\mathcal{A}$ be a countable index set. For $\alpha \in \mathcal{A}$, let $\mathcal{P}_\alpha$ be a space of distributions on $\mathcal{Z}$ dominated by a σ-finite measure ν, and let $\mathcal{P} \coloneqq \bigcup_{\alpha \in \mathcal{A}} \mathcal{P}_\alpha$, which is a metric space when equipped with the Hellinger metric d_{H}. For example, $\alpha > 0$ might index smoothness,

in which case $\mathcal{P}_\alpha$ might denote the set of distributions of (X, Y) for which the regression function $x \mapsto \mathbb{E}(Y \mid X = x)$ is α-smooth. For $\alpha \in \mathcal{A}$, let π_α be a distribution on $\mathcal{P}_\alpha$ and let $\lambda = (\lambda_\alpha)_{\alpha\in\mathcal{A}}$ be such that $\lambda_\alpha > 0$ for $\alpha \in \mathcal{A}$ and $\sum_{\alpha\in\mathcal{A}} \lambda_\alpha = 1$. We consider a prior distribution π on $\mathcal{P}$ of the form

$$\pi \coloneqq \sum_{\alpha\in\mathcal{A}} \lambda_\alpha \pi_\alpha.$$

For $z_1, \ldots, z_n \in \mathcal{Z}$, we may define a distribution on $\mathcal{P}$ by

$$\pi(B \mid z_1, \ldots, z_n) \coloneqq \frac{\int_B \prod_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\nu}(z_i)\,\mathrm{d}\pi(P)}{\int_{\mathcal{P}} \prod_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\nu}(z_i)\,\mathrm{d}\pi(P)},$$

for any measurable $B \subseteq \mathcal{P}$. Thus, $\pi(\cdot \mid Z_1, \ldots, Z_n)$ would be the posterior distribution when P has prior distribution π and $Z_1, \ldots, Z_n \mid P \overset{\text{iid}}{\sim} P$. In the sequel, however, we will assume instead that $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P_0$ for some fixed distribution P_0 on $\mathcal{Z}$, and seek bounds on the expected mass assigned by $\pi(\cdot \mid Z_1, \ldots, Z_n)$ to (Hellinger) balls around P_0.

For $\alpha \in \mathcal{A}$, a distribution P_0 on $\mathcal{Z}$ dominated by ν, and for $\epsilon > 0$, define

$$B_\alpha(P_0, \epsilon) \coloneqq \Big\{P \in \mathcal{P}_\alpha : \mathrm{KL}(P_0, P) \le \epsilon^2,\, \mathrm{V}_2(P_0, P) \le \epsilon^2\Big\} \tag{35}$$

and

$$B'_\alpha(P_0, \epsilon) \coloneqq \Big\{P \in \mathcal{P}_\alpha : d_{\mathrm{H}}(P_0, P) \le \epsilon\Big\}. \tag{36}$$

Let $(\epsilon_{n,\alpha})_{\alpha\in\mathcal{A}}$ be a positive sequence, thought of as the rate of convergence in Hellinger distance on each $\mathcal{P}_\alpha$. Fix $\beta \in \mathcal{A}$ and $H \ge 1$, and define

$$\mathcal{A}_{\ge\beta} \coloneqq \big\{\alpha \in \mathcal{A} : \epsilon_{n,\alpha}^2 \le H\epsilon_{n,\beta}^2\big\}$$
$$\mathcal{A}_{<\beta} \coloneqq \big\{\alpha \in \mathcal{A} : \epsilon_{n,\alpha}^2 > H\epsilon_{n,\beta}^2\big\},$$

to be the sets of indices that are 'at least as regular as β' and 'less regular than β', respectively. Here β should be thought of as the index of the 'best model' for P_0.

We now state five conditions for our posterior contraction theory:

(C1) Covering number bound: There exists a positive sequence $(E_\alpha)_{\alpha\in\mathcal{A}}$ such that

$$\sup_{\epsilon\ge\epsilon_{n,\alpha}} \log \mathcal{N}\big(\epsilon/3,\, B'_\alpha(P_0, 2\epsilon),\, d_{\mathrm{H}}\big) \le E_\alpha n\epsilon_{n,\alpha}^2 \quad \text{for all } \alpha \in \mathcal{A}.$$

(C2) Prior mass lower bound for β: There exists $F > 0$ such that

$$\pi_\beta\big(B_\beta(P_0, \epsilon_{n,\beta})\big) \ge e^{-Fn\epsilon_{n,\beta}^2}.$$

(C3) Mixture weights: There exists a positive sequence $(\mu_{n,\alpha})_{\alpha\in\mathcal{A}}$ such that

$$\frac{\lambda_\alpha}{\lambda_\beta} \le \mu_{n,\alpha}\, e^{n(H^{-1}\epsilon_{n,\alpha}^2 \vee \epsilon_{n,\beta}^2)} \quad \text{for all } \alpha \in \mathcal{A}.$$

(C4) Prior mass upper bound for $\mathcal{A}_{<\beta}$: There exist $M, G > 0$ such that

$$\sum_{\alpha\in\mathcal{A}_{<\beta}} \frac{\lambda_\alpha}{\lambda_\beta} \cdot \pi_\alpha\big(B'_\alpha(P_0, M\epsilon_{n,\alpha})\big) \le Ge^{-(F+3)n\epsilon_{n,\beta}^2}.$$

(C5) Uniformity of covering number constants: There exists $E \in (0, \infty)$ such that

$$E \ge \sup_{\alpha\in\mathcal{A}_{<\beta}} E_\alpha \vee \sup_{\alpha\in\mathcal{A}_{\ge\beta}} \frac{E_\alpha\epsilon_{n,\alpha}^2}{\epsilon_{n,\beta}^2}.$$

The set $\mathcal{A}_{<\beta}$ may be empty; in that case, condition (C4) is automatically satisfied, and the supremum over the empty set in (C5) is defined as $-\infty$.

The following theorem is a modification of Ghosal, Lember and van der Vaart (2008, Theorem 2.1), where we use a slightly larger constant in the exponent in (C4) and impose an additional sub-Gamma assumption on the log-likelihood ratios in order to obtain exponential posterior contraction (as opposed to the weaker conclusion that the left-hand side of (37) converges to zero as $n \to \infty$). In our in-context learning problem, this is crucial to derive rates of convergence for the posterior regression function. The sub-Gamma assumption is verified in Lemma G.2*(c)* under a nonparametric regression setting with independent Gaussian noise.

Proposition E.1. *Let P_0 be a distribution on $\mathcal{Z}$, let $\beta \in \mathcal{A}$ and let $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} P_0$. Assume that the conditions (C1)–(C5) hold with $M^2 > 243(F+1)(HE+1)$ and $\sum_{\alpha \in \mathcal{A}} \sqrt{\mu_{n,\alpha}} \leq e^{n\epsilon_{n,\beta}^2}$. For $i \in [n]$ and $P \in \mathcal{P}_\beta$ with with $P \ll P_0$, let*

$$W(P, Z_i) \coloneqq \log \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i) - \mathbb{E} \log \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i).$$

Assume that there exist $a, c > 0$ such that for all $P \in B_\beta(P_0, \epsilon_{n,\beta})$, we have that $W(P, Z_1)$ is sub-Gamma in the left tail with variance parameter $a\epsilon_{n,\beta}^2$ and scale parameter c. Then there exist $C_1, c_2, c_3 > 0$, not depending on n, such that if $n\epsilon_{n,\beta}^2 \geq c_3$, then

$$\mathbb{E}_{(Z_1,\dots,Z_n) \sim P_0^{\otimes n}} \left[\pi\big(\{P \in \mathcal{P} : d_{\mathrm{H}}(P, P_0) \geq M\epsilon_{n,\beta}\} \,\big|\, Z_1, \dots, Z_n \big) \right] \leq C_1 e^{-c_2 n \epsilon_{n,\beta}^2}. \tag{37}$$

Proof. Our assumptions ensure that the conditions of Ghosal, Lember and van der Vaart (2008, Theorem 2.1) hold[7] with $K = 1/9$, $L = K/3 = 1/27$, $I = \sqrt{27(F+1)}$ and $B = M/I = M/\sqrt{27(F+1)}$ therein. Moreover, our conditions (C2) and (C4) ensure that

$$\sum_{\alpha \in \mathcal{A}_{<\beta}} \frac{\lambda_\alpha}{\lambda_\beta} \cdot \frac{\pi_\alpha\big(B'_\alpha(P_0, M\epsilon_{n,\alpha})\big)}{\pi_\beta\big(B_\beta(P_0, \epsilon_{n,\beta})\big)} \leq G e^{-3n\epsilon_{n,\beta}^2}. \tag{38}$$

Now, by following the proof of Ghosal, Lember and van der Vaart (2008, Theorem 2.1), replacing the use of their Eq. (2.4) by our (38), and the use of their Lemma 6.3 by our Lemma G.1 (which applies because of our sub-Gamma assumption on $W(P, Z_1)$), we deduce the desired conclusion. □

E.2 Bayesian nonparametric regression

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. In this section, we specialise to the setting of regression and take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let P_X be a Borel distribution on $\mathcal{X}$, let $X \sim P_X$ and let $\xi \sim N(0, \sigma^2)$ be independent of X. Let $R > 0$, and for $\alpha \in \mathcal{A}$, let $\mathcal{G}_\alpha$ be a space of measurable functions from $\mathcal{X}$ to $[-R, R]$. For any Borel measurable function $g : \mathcal{X} \to \mathbb{R}$, we write P_g for the distribution of (X, Y_g) where $Y_g = g(X) + \xi$. For $\alpha \in \mathcal{A}$, let $\widetilde{\pi}_\alpha$ be a distribution on $\mathcal{G}_\alpha$, let $\widetilde{g}_\alpha \sim \widetilde{\pi}_\alpha$ and let π_α denote the distribution of the random measure $P_{\widetilde{g}_\alpha}$, where the randomness comes from $\widetilde{g}_\alpha \sim \widetilde{\pi}_\alpha$. Finally, let $\widetilde{\pi} \coloneqq \sum_{\alpha \in \mathcal{A}} \lambda_\alpha \widetilde{\pi}_\alpha$ and $\pi \coloneqq \sum_{\alpha \in \mathcal{A}} \lambda_\alpha \pi_\alpha$, where $\lambda_\alpha \geq 0$ for all $\alpha \in \mathcal{A}$ and $\sum_{\alpha \in \mathcal{A}} \lambda_\alpha = 1$, let $\mathcal{P}$ be the support of π and $\mathcal{G} \coloneqq \bigcup_{\alpha \in \mathcal{A}} \mathcal{G}_\alpha$.

We first provide an alternative expression for the π-posterior regression function g_π, defined in (3). For $P \in \mathcal{P}$, let $g_P : \mathcal{X} \to \mathcal{Y}$ be the regression function under P, so that $g_P(x) \coloneqq \mathbb{E}_{(X,Y) \sim P}(Y \mid X = x)$ for $x \in \mathcal{X}$. Then the π-posterior regression function g_π satisfies

$$\begin{aligned} g_\pi(x, D_n) &= \mathbb{E}_{P \sim \pi,\, (\mathcal{D}_n, X, Y) \mid P \sim P^{\otimes(n+1)}} \big\{ \mathbb{E}(Y \mid X = x,\, P) \mid X = x,\, \mathcal{D}_n = D_n \big\} \\ &= \mathbb{E}_{P \sim \pi,\, (\mathcal{D}_n, X, Y) \mid P \sim P^{\otimes(n+1)}} \big(g_P(x) \mid \mathcal{D}_n = D_n \big) \\ &= \int_{\mathcal{P}} g_P(x)\, \mathrm{d}\pi(P \mid D_n) = \int_{\mathcal{G}} g(x)\, \mathrm{d}\widetilde{\pi}(g \mid D_n), \end{aligned} \tag{39}$$

so $g_\pi(\cdot, D_n)$ can be thought of as the posterior mean of the regression function.

[7] As we seek a finite-sample bound, we do not require their condition that $n\epsilon_{n,\beta}^2 \to \infty$ as $n \to \infty$.

Corollary E.2. *Let $\beta \in \mathcal{A}$, let $g^0 : \mathcal{X} \to [-R, R]$ be measurable and write $P_0 \equiv P_{g^0}$. Assume that the conditions (C1)–(C5) hold with $M^2 > 243(F+1)(HE+1)$ and $\sum_{\alpha\in\mathcal{A}} \sqrt{\mu_{n,\alpha}} \le e^{n\epsilon_{n,\beta}^2}$. There exist $C_1, C_2, c_3, c_4 > 0$, not depending on n, such that if $n\epsilon_{n,\beta}^2 \ge c_4$, then*

$$\mathbb{E}_{(\mathcal{D}_n,X,Y)\sim P_0^{\otimes(n+1)}}\left[\left\{g_\pi(X,\mathcal{D}_n) - g^0(X)\right\}^2\right] \le C_1\epsilon_{n,\beta}^2 + C_2 e^{-c_3 n\epsilon_{n,\beta}^2}.$$

Proof. We write $\mathbb{E}_{P_0}$ for $\mathbb{E}_{(\mathcal{D}_n,X,Y)\sim P_0^{\otimes(n+1)}}$ and $C_1' \coloneqq 2R^2/\big(1 - e^{-R^2/(2\sigma^2)}\big)$. We have by (39) that

$$\begin{aligned}
\mathbb{E}_{P_0}\left[\left\{g_\pi(X,\mathcal{D}_n) - g^0(X)\right\}^2\right] &= \mathbb{E}_{P_0}\left[\left\{\int_{\mathcal{P}} g_P(X)\,\mathrm{d}\pi(P\,|\,\mathcal{D}_n) - g^0(X)\right\}^2\right] \\
&\overset{(i)}{\le} \mathbb{E}_{P_0}\left[\int_{\mathcal{P}} \left\{g_P(X) - g^0(X)\right\}^2 \mathrm{d}\pi(P\,|\,\mathcal{D}_n)\right] \\
&\overset{(ii)}{=} \mathbb{E}_{\mathcal{D}_n\sim P_0^{\otimes n}}\left(\int_{\mathcal{P}} \mathbb{E}_{(X,Y)\sim P_0}\left[\left\{g_P(X) - g^0(X)\right\}^2\right] \mathrm{d}\pi(P\,|\,\mathcal{D}_n)\right) \\
&\overset{(iii)}{\le} C_1'\mathbb{E}_{\mathcal{D}_n\sim P_0^{\otimes n}}\left[\int_{\mathcal{P}} d_{\mathrm{H}}^2(P,P_0)\,\mathrm{d}\pi(P\,|\,\mathcal{D}_n)\right] \\
&\overset{(iv)}{\le} C_1'M^2\epsilon_{n,\beta}^2 + 2C_1'\mathbb{E}_{\mathcal{D}_n\sim P_0^{\otimes n}}\left[\pi\big(\{P\in\mathcal{P} : d_{\mathrm{H}}(P,P_0) \ge M\epsilon_{n,\beta}\}\,\big|\,\mathcal{D}_n\big)\right] \\
&\overset{(v)}{\le} C_1\epsilon_{n,\beta}^2 + C_2 e^{-c_3 n\epsilon_{n,\beta}^2}.
\end{aligned}$$

Here, (i) follows from Jensen's inequality, (ii) follows by Fubini's theorem, (iii) follows from Lemma G.2*(d)*, (iv) follows by splitting the integral into regions where $d_{\mathrm{H}}(P,P_0) < M\epsilon_{n,\beta}$ and $d_{\mathrm{H}}(P,P_0) \ge M\epsilon_{n,\beta}$ and (v) follows from Proposition E.1 where the sub-Gamma assumption is verified in Lemma G.2*(c)*. □

F Proofs for Section 4

F.1 Definition of our random Besov function

Since Besov functions can be characterised by their wavelet decomposition, we will consider regression functions with random wavelet coefficients. Let $L^2([0,1]^d)$ denote the set of square integrable functions on $[0,1]^d$, let $S \in \mathbb{N}$ and let

$$\big\{\Phi_k : k \in K\big\} \cup \big\{\Psi_{\ell,\gamma} : \ell \in [\ell_0 : \infty),\ \gamma \in \Gamma_\ell\big\} \tag{40}$$

denote the tensor product Cohen–Daubechies–Vial (CDV) wavelet basis for $L^2([0,1]^d)$, constructed from S-regular and S-times continuously differentiable wavelets. The precise definitions of these wavelet functions are not important for us, and we will only use some basic properties of this wavelet basis, summarised in Appendix D.2. Let $c_0 \in (0,1]$ and $C_0 \ge 1$. For $\alpha \in (0,S)$, define the random function

$$\widetilde{g}_\alpha \coloneqq \sum_{k\in K} C_0 2^{-\ell_0 d/2} a_k^{(\alpha)} \Phi_k + \sum_{\ell=\ell_0}^{\infty}\sum_{\gamma\in\Gamma_\ell} C_0 2^{-\ell(\alpha+d/2)} b_{\ell,\gamma}^{(\alpha)} \Psi_{\ell,\gamma}, \tag{41}$$

where $\big(a_k^{(\alpha)}, b_{\ell,\gamma}^{(\alpha)} : k \in K, \ell \in [\ell_0 : \infty), \gamma \in \Gamma_\ell\big)$ are independent random variables supported on $[-1,1]$ with Lebesgue density bounded between $c_0/2$ and $c_0^{-1}/2$. Further, define $\widetilde{\pi}_\alpha \coloneqq \mathsf{Law}(\widetilde{g}_\alpha)$ as the distribution of $\widetilde{g}_\alpha$. The scaling of the wavelet coefficients in (41) is chosen to ensure that $\widetilde{g}_\alpha$ belongs to the Besov space $B_{\infty,\infty}^\alpha([0,1]^d)$ with smoothness α. Indeed, a function $g \coloneqq \sum_{k\in K} a_k'\Phi_k + \sum_{\ell=\ell_0}^\infty \sum_{\gamma\in\Gamma_\ell} b_{\ell,\gamma}'\Psi_{\ell,\gamma}$, with deterministic wavelet coefficients, belongs to the Besov ball $B_{\infty,\infty}^\alpha([0,1]^d, C)$ of radius $C > 0$ if and only if $2^{\ell_0 d/2}\max_k |a_k'| + \sup_{\ell,\gamma} 2^{\ell(\alpha+d/2)}|b_{\ell,\gamma}'| \le C$; see Appendix D. Thus, writing $\mathcal{G}_\alpha \coloneqq \mathrm{supp}(\widetilde{\pi}_\alpha)$, we have

$$B_{\infty,\infty}^\alpha([0,1]^d, C_0) \subseteq \mathcal{G}_\alpha \subseteq B_{\infty,\infty}^\alpha([0,1]^d, 2C_0). \tag{42}$$

We remark that the Besov space $B^{\alpha}_{\infty,\infty}([0,1]^d)$ coincides with the Hölder space $H^{\alpha}([0,1]^d)$ for $\alpha \notin \mathbb{N}$, and it contains $H^{\alpha}([0,1]^d)$ for $\alpha \in \mathbb{N}$; see Appendix D.1 for more details. Moreover, there exists $C_{\alpha,S,d} > 0$ such that $\sup_{g \in \mathcal{G}_\alpha} \|g\|_\infty \leq C_{\alpha,S,d}$; see (56). See Figure 3 for realisations of random Besov functions with $\alpha = 2$ and 4.

Assumption F.1. Let P_X be a probability distribution on $[0,1]^d$ with the property that there exists a hypercube $A \subseteq [0,1]^d$ such that $P_X(A_0) \geq c_0 \mathrm{Vol}_d(A_0)$ for all measurable $A_0 \subseteq A$.

F.2 Definition of our random multi-index model

For $p \in [d]$, let $V_p(\mathbb{R}^d) := \{U \in \mathbb{R}^{d \times p} : U^\top U = I_p\}$ denote the set of all projection matrices from $\mathbb{R}^d$ to $\mathbb{R}^p$ (also called the *Stiefel manifold*). We will consider regression functions on $\mathbb{B}^d$ of the form $x \mapsto g\big((U^\top x + 1_p)/2\big)$, where $U \in V_p(\mathbb{R}^d)$ for some *effective dimension* $p \in [d]$ and Borel measurable $g : [0,1]^p \to \mathbb{R}$. Note that we translate and scale $U^\top x$ to ensure that it belongs to $[0,1]^p$.

For $\alpha \in (0, S)$ and $p \in [d]$, let $\widetilde{g}^{(p)}_\alpha : [0,1]^p \to \mathbb{R}$ be the random function (with smoothness α) defined as in (41), but with d replaced by p in the wavelet coefficients, and using the tensor product CDV wavelet basis for $L^2([0,1]^p)$ instead of $L^2([0,1]^d)$. Let $\widetilde{\pi}_{\alpha,p}$ be the distribution of the random function $\widetilde{g}_{\alpha,p} : \mathbb{B}^d \to \mathbb{R}$ defined by

$$\widetilde{g}_{\alpha,p}(x) := \widetilde{g}^{(p)}_\alpha\left(\frac{(U^{(p)})^\top x + 1_p}{2}\right), \tag{43}$$

where $U^{(p)}$ is uniformly distributed[8] on the Stiefel manifold $V_p(\mathbb{R}^d)$, independently of $\widetilde{g}^{(p)}_\alpha$. We write $\mathcal{G}_{\alpha,p}$ for the support of $\widetilde{\pi}_{\alpha,p}$.

Let P_X be a distribution on $\mathbb{B}^d$ with the property that there exist $c_0 > 0$ and a non-empty Euclidean ball $A \subseteq \mathbb{B}^d$ such that $P_X(A_0) \geq c_0 \mathrm{Vol}_d(A_0)$ for all measurable $A_0 \subseteq A$. Given $g \in \mathcal{G}_{\alpha,p}$, let P_g denote the distribution of (X, Y_g) where $X \sim P_X$,

$$Y_g = g(X) + \xi, \quad \xi \sim N(0, \sigma^2) \quad \text{and} \quad \xi \perp\!\!\!\perp X.$$

We then define $\pi_{\alpha,p}$ to be the distribution of the random measure $P_{\widetilde{g}_{\alpha,p}}$, where the randomness comes from the random regression function $\widetilde{g}_{\alpha,p} \sim \widetilde{\pi}_{\alpha,p}$, and let $\mathcal{P}_{\alpha,p} := \mathrm{supp}(\pi_{\alpha,p})$.

F.3 Proof of Proposition 4.1

We begin with two preliminary lemmas in the setting of Section 4.1. Suppose that $g^\circ \in \mathcal{G}_\beta$ has wavelet decomposition

$$g^\circ = \sum_{k \in K} C_0 2^{-\ell_0 d/2} a^\circ_k \Phi_k + \sum_{\ell=\ell_0}^{\infty} \sum_{\gamma \in \Gamma_\ell} C_0 2^{-\ell(\beta + d/2)} b^\circ_{\ell,\gamma} \Psi_{\ell,\gamma}, \tag{44}$$

where $(a^\circ_k)_{k \in K}$ and $(b^\circ_{\ell,\gamma})_{\ell \in [\ell_0 : \infty), \gamma \in \Gamma_\ell}$ are deterministic and satisfy $|a^\circ_k| \leq 1$ and $|b^\circ_{\ell,\gamma}| \leq 1$ for $k \in K$, $\ell \in [\ell_0 : \infty)$ and $\gamma \in \Gamma_\ell$. Lemma F.2 below provides a lower bound on the prior mass; similar results for the one-dimensional case exist in the literature, e.g. Giné and Nickl (2011, Proposition 1) and Reiß and Schmidt-Hieber (2020, Lemma 4.1).

Lemma F.2. *Let $n \in \mathbb{N}$ and assume that (44) holds. Then there exists $C > 0$, depending only on β, d, c_0, C_0 and S, such that*

$$\widetilde{\pi}_\beta\big(\{g \in \mathcal{G}_\beta : \|g - g^\circ\|^2_\infty \leq C n^{-2\beta/(2\beta+d)}\}\big) \geq \exp\big(-C n^{d/(2\beta+d)}\big). \tag{45}$$

Moreover, there exists $C' > 0$, depending only on β, d, c_0, C_0, S and σ, such that

$$\pi_\beta\big(B_\beta(P_{g^\circ}, \epsilon_{n,\beta})\big) \geq \exp\big(-C' n \epsilon^2_{n,\beta}\big), \tag{46}$$

where $\epsilon_{n,\beta} := C' n^{-\beta/(2\beta+d)}$ and $B_\beta(P_{g^\circ}, \epsilon_{n,\beta})$ is defined by (35).

[8] One way to generate such a random matrix is to set $U^{(p)} := Z(Z^\top Z)^{-1/2}$, where Z is a $d \times p$ random matrix with independent $N(0,1)$ entries (e.g. Chikuse, 2003, Theorem 2.2.1(iii)).

Proof. Throughout the proof, $C_1, C_2, \ldots$ will denote positive quantities that may depend only on β, d, c_0, C_0 and S, while $C_1', C_2', \ldots$ may in addition depend on σ. Let $\widetilde{g}_\beta$ be given by (41) so that $\widetilde{g}_\beta \sim \widetilde{\pi}_\beta$, and write $a_k \equiv a_k^{(\beta)}$ and $b_{\ell,\gamma} \equiv b_{\ell,\gamma}^{(\beta)}$ where the random variables $(a_k^{(\beta)})$ and $(b_{\ell,\gamma}^{(\beta)})$ satisfy the conditions below (41). By (30) and (31), writing $L := \lceil \frac{\log_2 n}{2\beta+d} \rceil$ (and assuming for now that $L \geq \ell_0$), we have

$$\begin{aligned}\|\widetilde{g}_\beta - g^\circ\|_\infty &\leq C_1 \max_{k\in K} |a_k - a_k^\circ| + C_1 \sum_{\ell=\ell_0}^{L} 2^{-\ell\beta} \max_{\gamma\in\Gamma_\ell} |b_{\ell,\gamma} - b_{\ell,\gamma}^\circ| + 2C_0C_1 \sum_{\ell=L+1}^{\infty} 2^{-\ell\beta} \\ &\leq C_1 \max_{k\in K} |a_k - a_k^\circ| + C_1 \sum_{\ell=\ell_0}^{L} 2^{-\ell\beta} \max_{\gamma\in\Gamma_\ell} |b_{\ell,\gamma} - b_{\ell,\gamma}^\circ| + C_3 n^{-\beta/(2\beta+d)}. \qquad (47)\end{aligned}$$

Define the events

$$\begin{aligned}\mathcal{E}_1 &:= \bigcap_{k\in K} \Big\{ |a_k - a_k^\circ| \leq 2 \cdot n^{-\beta/(2\beta+d)} \Big\}, \\ \mathcal{E}_2 &:= \bigcap_{\ell=\ell_0}^{L} \bigcap_{\gamma\in\Gamma_\ell} \Big\{ |b_{\ell,\gamma} - b_{\ell,\gamma}^\circ| \leq 2 \cdot 2^{-2(L-\ell)\beta} \Big\}.\end{aligned}$$

Then, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have by (47) that

$$\|\widetilde{g}_\beta - g^\circ\|_\infty \leq 2C_1 n^{-\beta/(2\beta+d)} + 2C_1 2^{-L\beta} \sum_{\ell=\ell_0}^{L} 2^{-(L-\ell)\beta} + C_3 n^{-\beta/(2\beta+d)} \leq C_4 n^{-\beta/(2\beta+d)}.$$

Therefore,

$$\mathbb{P}\big(\|\widetilde{g}_\beta - g^\circ\|_\infty \leq C_4 n^{-\beta/(2\beta+d)}\big) \geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_1)\mathbb{P}(\mathcal{E}_2). \qquad (48)$$

Now, by (29), we have

$$\mathbb{P}(\mathcal{E}_1) \geq \big(c_0 n^{-\beta/(2\beta+d)}\big)^{2^{\ell_0 d}} = \exp(-C_5 \log n). \qquad (49)$$

Similarly,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2) > \prod_{\ell=\ell_0}^{L} \big(c_0 2^{-2(L-\ell)\beta}\big)^{2^{(\ell+1)d}} &= \exp\bigg((\log c_0) \sum_{\ell=\ell_0}^{L} 2^{(\ell+1)d} - (2^{d+1}\log 2)\beta \sum_{\ell=\ell_0}^{L} (L-\ell) 2^{\ell d} \bigg) \\ &\geq \exp(-C_6 n^{d/(2\beta+d)}), \qquad (50)\end{aligned}$$

where the final inequality follows since if we set $A := \sum_{\ell=\ell_0}^{L} (L-\ell) 2^{\ell d}$, then $2^d A - A \leq \sum_{\ell=0}^{L} 2^{\ell d} \leq 2^{(L+1)d}/(2^d-1)$, so $A \leq 2^{(L+1)d}/(2^d-1)^2$. The conclusion (45) for n large enough that $L \geq \ell_0$ follows from (48), (49) and (50). By increasing $C > 0$, depending only on β, d, c_0, C_0 and S, we can then ensure that (45) holds for all $n \in \mathbb{N}$.

For (46), note that by Lemma G.2*(a)* and *(b)*, we have that if $g \in \mathcal{G}_\beta$ satisfies $\|g - g^\circ\|_\infty^2 \leq C n^{-2\beta/(2\beta+d)}$, then

$$\mathrm{KL}(P_{g^\circ}, P_g) \leq C_1' C n^{-2\beta/(2\beta+d)} \quad \text{and} \quad \mathrm{V}_2(P_{g^\circ}, P_g) \leq C_1' C n^{-2\beta/(2\beta+d)}.$$

Writing $\epsilon_{n,\beta} := \sqrt{C_1' C} \cdot n^{-\beta/(2\beta+d)}$, we deduce from (45) that

$$\pi_\beta\big(B_\beta(P_{g^\circ}, \epsilon_{n,\beta})\big) \geq \exp\big(-C n^{d/(2\beta+d)}\big) \geq \exp(-C_2' n \epsilon_{n,\beta}^2).$$

Thus (46) holds with $C' := \sqrt{C_1' C} \vee C_2'$. □

Lemma F.3 below provides an upper bound on the prior mass.

Lemma F.3. *Let $\alpha \in \mathcal{A}$, $m > 0$ and assume that (44) holds. Suppose that P_X is a Borel distribution on $[0,1]^d$ with the property that there exist $c_0 > 0$ and a hypercube $A \subseteq [0,1]^d$ of side length $\tau > 0$ such that $P_X(A_0) \geq c_0 \mathrm{Vol}_d(A_0)$ for all measurable $A_0 \subseteq A$. Then there exist $c \equiv c(\alpha, d, c_0, C_0, \tau) > 0$ and $n_0 \equiv n_0(\alpha, d, c_0, C_0, S, m, \tau) \in \mathbb{N}$ such that*

$$\widetilde{\pi}_\alpha\Big(\Big\{g \in \mathcal{G}_\alpha : \|g - g^\circ\|^2_{L^2(P_X)} \leq \frac{c_0}{4} m^2 n^{-2\alpha/(2\alpha+d)}\Big\}\Big) \leq \exp\big(-cm^{-d/\alpha} n^{d/(2\alpha+d)}\big) \tag{51}$$

for $n \geq n_0$. Moreover, there exists $c' > 0$, depending only on $\alpha, \beta, d, c_0, C_0, S, \tau$ and σ, such that for $n \geq n_0$,

$$\pi_\alpha\big(B'_\alpha(P_{g^\circ}, c' m n^{-\alpha/(2\alpha+d)})\big) \leq \exp\big(-cm^{-d/\alpha} n^{d/(2\alpha+d)}\big) \tag{52}$$

where $B'_\alpha(P_{g^\circ}, c' m n^{-\alpha/(2\alpha+d)})$ is defined as in (36).

Proof. Throughout the proof, $C_1, C_2, \ldots$ will denote positive quantities that may depend only on α, d, c_0, C_0 and τ, while $C'_1, C'_2, \ldots$ may depend in addition on β, S and σ. Let $\epsilon := n^{-\alpha/(2\alpha+d)}$. Fix $g' \in \mathcal{G}_\alpha$ such that $\|g' - g^\circ\|^2_{L^2(P_X)} \leq \frac{c_0}{4} m^2 \epsilon^2$; we may assume without loss of generality that such a g' exists, since otherwise the left-hand side of (51) would be zero. Suppose that g' has wavelet decomposition

$$g' = \sum_{k \in K} C_0 2^{-\ell_0 d/2} a'_k \Phi_k + \sum_{\ell=\ell_0}^\infty \sum_{\gamma \in \Gamma_\ell} C_0 2^{-\ell(\alpha+d/2)} b'_{\ell,\gamma} \Psi_{\ell,\gamma},$$

where $|a'_k| \leq 1$ and $|b'_{\ell,\gamma}| \leq 1$. Note that if $g \in \mathcal{G}_\alpha$ satisfies $\|g - g^\circ\|^2_{L^2(P_X)} \leq \frac{c_0}{4} m^2 \epsilon^2$, then $\|g - g'\|^2_{L^2(P_X)} \leq 2\|g - g^\circ\|^2_{L^2(P_X)} + 2\|g' - g^\circ\|^2_{L^2(P_X)} \leq c_0 m^2 \epsilon^2$. Thus,

$$\widetilde{\pi}_\alpha\Big(\Big\{g \in \mathcal{G}_\alpha : \|g - g^\circ\|^2_{L^2(P_X)} \leq \frac{c_0}{4} m^2 \epsilon^2\Big\}\Big) \leq \widetilde{\pi}_\alpha\Big(\Big\{g \in \mathcal{G}_\alpha : \|g - g'\|^2_{L^2(P_X)} \leq c_0 m^2 \epsilon^2\Big\}\Big). \tag{53}$$

Let $\widetilde{g}_\alpha \sim \widetilde{\pi}_\alpha$, and write $a_k \equiv a_k^{(\alpha)}$ and $b_{\ell,\gamma} \equiv b_{\ell,\gamma}^{(\alpha)}$ with representation (41), where the random variables $(a_k^{(\alpha)})$ and $(b_{\ell,\gamma}^{(\alpha)})$ satisfy the conditions below (41). Further, let $C_1 := \lceil \log_2(2/\tau) \rceil$. Then, by the assumption on P_X, there exists $C_2 \in [0 : 2^{C_1} - 1]$ such that $\big[\frac{C_2}{2^{C_1}}, \frac{C_2+1}{2^{C_1}}\big]^d \subseteq A$. Choose $n_0 \equiv n_0(\alpha, d, c_0, C_0, S, m, \tau) \in \mathbb{N}$ large enough that $L := \big\lfloor \frac{1}{2\alpha} \log_2\big(\frac{c_0 C_0^2}{6m^2\epsilon^2 2^{C_1 d}}\big)\big\rfloor$ is at least $\ell_0 \vee C_1$ for $n \geq n_0$. Assume henceforth that $n \geq n_0$. Let $\Gamma'_L := \big\{\gamma \in \Gamma_L : \mathrm{supp}(\Psi_{L,\gamma}) \subseteq \big[\frac{C_2}{2^{C_1}}, \frac{C_2+1}{2^{C_1}}\big]^d\big\}$, $b_L := (b_{L,\gamma})_{\gamma \in \Gamma'_L}$ and $b'_L := (b'_{L,\gamma})_{\gamma \in \Gamma'_L}$. Since $\mathrm{supp}(\Psi_{L,\gamma})$ is contained in a dyadic cube of side length 2^{-L} (see Section D.2), we deduce that $|\Gamma'_L| \geq 2^{(L-C_1)d}$ and that

$$\begin{aligned} C_0^2 2^{-L(2\alpha+d)} \mathbb{E}\big(\|b_L - b'_L\|_2^2\big) &\geq C_0^2 2^{-L(2\alpha+d)} \mathrm{tr}\big(\mathrm{Cov}(b_L)\big) \\ &\geq |\Gamma'_L| C_0^2 2^{-L(2\alpha+d)} \cdot \frac{c_0}{3} \geq \frac{c_0 C_0^2 2^{-2\alpha L - C_1 d}}{3} \geq 2m^2\epsilon^2. \end{aligned}$$

Therefore, by the assumption on P_X and Lemma G.3, we have

$$\begin{aligned} \mathbb{P}\big(\|\widetilde{g}_\alpha - g'\|^2_{L^2(P_X)} \leq c_0 m^2 \epsilon^2\big) &\leq \mathbb{P}\big(c_0 \|\widetilde{g}_\alpha - g'\|^2_{L^2([C_2 2^{-C_1}, (C_2+1)2^{-C_1}]^d)} \leq c_0 m^2 \epsilon^2\big) \\ &\leq \mathbb{P}\bigg\{\|b_L - b'_L\|_2^2 - \mathbb{E}\big(\|b_L - b'_L\|_2^2\big) \leq \frac{m^2\epsilon^2}{C_0^2 2^{-L(2\alpha+d)}} - \mathbb{E}\big(\|b_L - b'_L\|_2^2\big)\bigg\} \\ &\leq \mathbb{P}\bigg\{\|b_L - b'_L\|_2^2 - \mathbb{E}\big(\|b_L - b'_L\|_2^2\big) \leq -\frac{m^2\epsilon^2}{C_0^2 2^{-L(2\alpha+d)}}\bigg\}. \end{aligned} \tag{54}$$

Since $(b_{L,\gamma} - b'_{L,\gamma})^2 \in [0,4]$ for $\gamma \in \Gamma'_L$, we have by Hoeffding's inequality and (29) that

$$\begin{aligned} \mathbb{P}\bigg\{\|b_L - b'_L\|_2^2 - \mathbb{E}\big(\|b_L - b'_L\|_2^2\big) \leq -\frac{m^2\epsilon^2}{C_0^2 2^{-L(2\alpha+d)}}\bigg\} &\leq \exp\bigg(-\frac{m^4\epsilon^4}{8C_0^4 2^{-(4\alpha+d)L - C_1 d}}\bigg) \\ &\leq \exp\bigg\{-\frac{c_0^2}{288 \cdot 2^{4\alpha + (C_1+1)d}} \bigg(\frac{c_0 C_0^2}{6m^2\epsilon^2 2^{C_1 d}}\bigg)^{d/(2\alpha)}\bigg\} =: \exp\big(-cm^{-d/\alpha} n^{d/(2\alpha+d)}\big), \end{aligned} \tag{55}$$

where $c > 0$ depends only on α, d, c_0, C_0 and τ. Combining (53), (54) and (55) yields (51).

For the second claim, note that by (30) and (31),

$$\|g^\circ\|_\infty \le C_3 \max_{k\in K} |a_k^\circ| + C_3 \sum_{\ell=\ell_0}^{\infty} 2^{-\ell\beta} \max_{\gamma\in\Gamma_\ell} |b_{\ell,\gamma}^\circ| \le C_1', \tag{56}$$

and similarly, $\|\widetilde{g}_\alpha\|_\infty \le C_2'$ almost surely. Thus, if $d_{\mathrm{H}}(P_{g^\circ}, P_g) \le c' m n^{-\alpha/(2\alpha+d)}$ for some $c' > 0$, then by Lemma G.2*(d)*,

$$\|g - g^\circ\|_{L^2(P_X)}^2 \le C_3' d_{\mathrm{H}}^2(P_{g^\circ}, P_g) \le C_3'(c')^2 m^2 n^{-2\alpha/(2\alpha+d)}.$$

Therefore, taking $c' := \sqrt{c_0/(4C_3')}$ and applying (51) proves (52). □

We are now in a position to prove Proposition 4.1.

Proof of Proposition 4.1. Throughout the proof, $C_1, C_2, \ldots$ will denote positive quantities that do not depend on n. By an argument similar to (56), there exists $C_1 > 0$ such that $\|g^\circ\|_\infty \le C_1$ and $\|g\|_\infty \le C_1$ for all $g \in \mathcal{G}$. We now verify the assumptions of Corollary E.2 with $H = 1$. For each $\alpha \in \mathcal{A}$, let $C' > 0$ (which does not depend on n) be the quantity in Lemma F.2 such that (46) holds, and let $\epsilon_{n,\alpha} := C' n^{-\alpha/(2\alpha+d)}$ for $\alpha \in \mathcal{A}$. Then by Lemma F.2, condition (C2) holds with $F = C'$. By Lemma G.2*(d)*, (42) and Giné and Nickl (2021, Eq. (4.184)), for each $\alpha \in \mathcal{A}$, we have

$$\begin{aligned}
\sup_{\epsilon \ge \epsilon_{n,\alpha}} \log \mathcal{N}\big(\epsilon/3,\, B'_\alpha(P_0, 2\epsilon),\, d_{\mathrm{H}}\big) &\le \log \mathcal{N}\big(\epsilon_{n,\alpha}/3,\, \mathcal{P}_\alpha,\, d_{\mathrm{H}}\big) \\
&\le \log \mathcal{N}\big(2\sigma\epsilon_{n,\alpha}/3,\, \mathcal{G}_\alpha,\, \|\cdot\|_\infty\big) \\
&\le \log \mathcal{N}\big(2\sigma\epsilon_{n,\alpha}/3,\, B^\alpha_{\infty,\infty}([0,1]^d, 2C_0),\, \|\cdot\|_\infty\big) \\
&\le E_\alpha n \epsilon_{n,\alpha}^2,
\end{aligned}$$

for some $E_\alpha > 0$ not depending on n; this verifies condition (C1). Moreover, since $\mathcal{A}$ is finite, condition (C3) is satisfied with $\mu_{n,\alpha} = 1$ for all $\alpha \in \mathcal{A}$ when n is sufficiently large. Similarly, condition (C5) is satisfied for some $E > 0$ not depending on n. Finally, take $M > 0$ such that $M^2 > 243(F+1)(HE+1)$. Letting $c, c' > 0$ be the quantities in Lemma F.3 and applying Lemma F.3 with $m := MC'/c'$, we deduce that for n sufficiently large,

$$\sum_{\alpha\in\mathcal{A}_{<\beta}} \frac{\lambda_\alpha}{\lambda_\beta} \cdot \pi_\alpha\big(B'_\alpha(P_0, M\epsilon_{n,\alpha})\big) \le \sum_{\alpha\in\mathcal{A}:\alpha<\beta} \frac{\lambda_\alpha}{\lambda_\beta} \cdot \exp\big(-cm^{-d/\alpha} n^{d/(2\alpha+d)}\big) \le e^{-(F+3)n\epsilon_{n,\beta}^2},$$

which verifies condition (C4). Therefore, by Corollary E.2, the claim follows for sufficiently large n, since $e^{-c_3 n\epsilon_{n,\beta}^2} \le \epsilon_{n,\beta}^2$ for n large enough. Moreover, since g_π and g° are bounded, the conclusion now follows for all $n \in \mathbb{N}$ by increasing $C > 0$ in the statement if necessary. □

F.4 Statement and proof of Proposition F.7

We work in the setting of Section 4.2. For $(\alpha, p) \in \mathcal{A}'$, we define $\mathcal{G}_\alpha^{(p)}$ similarly to the definition of $\mathcal{G}_\alpha$ in Section 4.1 but replacing the dimensionality d by p, i.e. we let $\mathcal{G}_\alpha^{(p)}$ be the set of all functions $h_\alpha^{(p)} : [0,1]^p \to \mathbb{R}$ such that h has wavelet decomposition

$$h_\alpha^{(p)} = \sum_{k\in K} C_0 2^{-\ell_0 r/2} a_k \Phi_k^{(p)} + \sum_{\ell=\ell_0}^{\infty} \sum_{\gamma\in\Gamma_\ell} C_0 2^{-\ell(\beta+r/2)} b_{\ell,\gamma} \Psi_{\ell,\gamma}^{(p)},$$

where $(\Phi_k^{(p)})$ and $(\Psi_{\ell,\gamma}^{(p)})$ is the tensor product CDV wavelet basis for $L^2([0,1]^p)$, $|a_k| \le 1$ and $|b_{\ell,\gamma}| \le 1$. Thus, similarly to (42),

$$B^\alpha_{\infty,\infty}([0,1]^p, C_0) \subseteq \mathcal{G}_\alpha^{(p)} \subseteq B^\alpha_{\infty,\infty}([0,1]^p, 2C_0).$$

The proof of Proposition F.7 relies on three preliminary lemmas that we now formulate. We may write

$$g^\circ(x) = h^\circ\left(\frac{(U^\circ)^\top x + 1_r}{2}\right), \tag{57}$$

where $U^\circ \in V_r(\mathbb{R}^d)$ and $h^\circ \in \mathcal{G}_\beta^{(r)}$ has wavelet decomposition

$$h^\circ = \sum_{k\in K} C_0 2^{-\ell_0 r/2} a_k^\circ \Phi_k^{(r)} + \sum_{\ell=\ell_0}^{\infty} \sum_{\gamma\in\Gamma_\ell} C_0 2^{-\ell(\beta+r/2)} b_{\ell,\gamma}^\circ \Psi_{\ell,\gamma}^{(r)}, \tag{58}$$

where $|a_k^\circ| \le 1$ and $|b_{\ell,\gamma}^\circ| \le 1$.

Lemma F.4. *Let $n \in \mathbb{N}$ and assume that (57) and (58) hold. Then there exists $C > 0$, depending only on β, d, r, c_0, C_0 and S, such that*

$$\widetilde{\pi}_{\beta,r}\big(\{g \in \mathcal{G}_{\beta,r} : \|g - g^\circ\|_\infty^2 \le Cn^{-2\beta/(2\beta+r)}\}\big) \ge \exp\big(-Cn^{r/(2\beta+r)}\big). \tag{59}$$

Moreover, there exists $C' > 0$, depending only on β, d, r, c_0, C_0, S and σ, such that if $\epsilon_{n,\beta,r} := C' n^{-\beta/(2\beta+r)}$, then

$$\pi_{\beta,r}\big(B_{\beta,r}(P_{g^\circ}, \epsilon_{n,\beta,r})\big) \ge \exp\big(-C' n\epsilon_{n,\beta,r}^2\big), \tag{60}$$

where $B_{\beta,r}(P_{g^\circ}, \epsilon_{n,\beta,r}) := \big\{P \in \mathcal{P}_{\beta,r} : \mathrm{KL}(P_{g^\circ}, P) \le \epsilon_{n,\beta,r}^2, \mathrm{V}_2(P_{g^\circ}, P) \le \epsilon_{n,\beta,r}^2\big\}$, similarly to (35).

Proof. Throughout the proof, $C_1, C_2, \ldots$ will denote positive quantities that may depend only on β, d, r, c_0, C_0 and S, while $C_1', C_2', \ldots$ may in addition depend on σ. Let $\widetilde{g}_{\beta,r}(x) = \widetilde{g}_\beta^{(r)}\big(\frac{(U^{(r)})^\top x + 1_r}{2}\big)$ be defined as in (43) so that $\widetilde{g}_{\beta,r} \sim \widetilde{\pi}_{\beta,r}$. Observe that $B_{\infty,\infty}^\beta([0,1]^r, 2C_0) \subseteq B_{\infty,\infty}^{\beta\wedge 1}([0,1]^r, 2C_0)$ (e.g. Giné and Nickl, 2021, Proposition 4.3.10(ii)). When $\beta < 1$, the space $B_{\infty,\infty}^{\beta\wedge 1}$ is equal to the Hölder space H^β with equivalent norms (e.g. Giné and Nickl, 2021, Proposition 4.3.23), so for all $x, y \in [0,1]^r$ and $g \in B_{\infty,\infty}^{\beta\wedge 1}([0,1]^r, 2C_0)$, we have $|g(x) - g(y)| \le C_1\|x - y\|_2^\beta$; when $\beta \ge 1$, $B_{\infty,\infty}^{\beta\wedge 1}$ is equal to the Zygmund space of order 1 with equivalent norms (Triebel, 1983, p. 113), so by Anderson and Pitt (1989, Proposition 2.4) (see also Zygmund, 2002, Chapter II, Theorem 3.4), for all $x, y \in [0,1]^r$ and $g \in B_{\infty,\infty}^{\beta\wedge 1}([0,1]^r, 2C_0)$, we have $|g(x) - g(y)| \le C_2\|x - y\|_2\big\{\log(1/\|x - y\|_2) \vee 1\big\}$. Therefore, for all $x, y \in [0,1]^r$ and $g \in B_{\infty,\infty}^\beta([0,1]^r, 2C_0)$,

$$|g(x) - g(y)| \le C_3\|x - y\|_2^{\beta\wedge 1}\big\{\log(1/\|x - y\|_2) \vee 1\big\}. \tag{61}$$

On the event $E := \big\{\|U^{(r)} - U^\circ\|_{\mathrm{op}} \le \big(n^{-\beta/(2\beta+r)}/\log(en)\big)^{1/(\beta\wedge 1)}\big\}$, we have for all $x \in \mathbb{B}^d$ that

$$\begin{aligned}
&|\widetilde{g}_{\beta,r}(x) - g^\circ(x)| \\
&\le \left|\widetilde{g}_\beta^{(r)}\left(\frac{(U^{(r)})^\top x + 1_r}{2}\right) - \widetilde{g}_\beta^{(r)}\left(\frac{(U^\circ)^\top x + 1_r}{2}\right)\right| + \left|\widetilde{g}_\beta^{(r)}\left(\frac{(U^\circ)^\top x + 1_r}{2}\right) - h^\circ\left(\frac{(U^\circ)^\top x + 1_r}{2}\right)\right| \\
&\le C_4 n^{-\beta/\{(2\beta+r)\}} + \|\widetilde{g}_\beta^{(r)} - h^\circ\|_\infty,
\end{aligned} \tag{62}$$

where the second inequality follows from (61), since $\widetilde{g}_\beta^{(r)} \in \mathcal{G}_\beta^{(r)} \subseteq B_{\infty,\infty}^\beta([0,1]^r, 2C_0)$ and $\big\|\frac{(U^{(r)})^\top x + 1_r}{2} - \frac{(U^\circ)^\top x + 1_r}{2}\big\|_2 \le \frac{1}{2}\big(n^{-\beta/(2\beta+r)}/\log(en)\big)^{1/(\beta\wedge 1)}$ on E. Now by Lemma G.4, we have

$$\mathbb{P}(E) \ge C_5\left(\frac{n^{-\beta/(2\beta+r)}}{\log(en)} \wedge \frac{1}{2}\right)^{dr/(\beta\wedge 1)} \ge \exp(-C_6 n^{r/(2\beta+r)}),$$

and by (45) (with d replaced by r therein), we have

$$\mathbb{P}\big(\|\widetilde{g}_\beta^{(r)} - h^\circ\|_\infty \le C_7 n^{-\beta/(2\beta+r)}\big) \ge \exp(-C_7 n^{r/(2\beta+r)}).$$

Hence,

$$\begin{aligned}
\mathbb{P}\big(\|\widetilde{g}_{\beta,r} - g^\circ\|_\infty^2 &\le (C_4 + C_7)^2 n^{-2\beta/(2\beta+r)}\big) \\
&\ge \mathbb{P}\big(\|\widetilde{g}_{\beta,r} - g^\circ\|_\infty \le (C_4 + C_7) n^{-\beta/(2\beta+r)} \,\big|\, E\big)\mathbb{P}(E) \\
&\ge \mathbb{P}\big(\|\widetilde{g}_\beta^{(r)} - h^\circ\|_\infty \le C_7 n^{-\beta/(2\beta+r)}\big)\mathbb{P}(E) \ge \exp\big\{-(C_6 + C_7) n^{r/(2\beta+r)}\big\},
\end{aligned}$$

where the second inequality follows from (62) and the fact that $\widetilde{g}_\beta^{(r)}$ and $U^{(r)}$ are independent. This proves (59) with $C := (C_4 + C_7)^2 \vee (C_6 + C_7)$.

The proof of (60) follows the same argument as the proof of (46). By Lemma G.2*(a)* and *(b)*, if $g \in \mathcal{G}_{\beta,r}$ satisfies $\|g - g^\circ\|_\infty^2 \le Cn^{-2\beta/(2\beta+r)}$, then

$$\mathrm{KL}(P_{g^\circ}, P_g) \le C_1' C n^{-2\beta/(2\beta+r)} \quad \text{and} \quad \mathrm{V}_2(P_{g^\circ}, P_g) \le C_2' C n^{-2\beta/(2\beta+r)}.$$

Writing $C' := C^{1/3} \vee \sqrt{(C_1' \vee C_2')C}$ and $\epsilon_{n,\beta,r} := C' n^{-\beta/(2\beta+r)}$, we deduce from (59) that

$$\pi_{\beta,r}\big(B_{\beta,r}(P_{g^\circ}, \epsilon_{n,\beta,r})\big) \ge \exp\big(-Cn^{r/(2\beta+r)}\big) \ge \exp(-C' n \epsilon_{n,\beta,r}^2).$$

This proves the claim. □

Lemma F.5. *Let $(\alpha, p) \in \mathcal{A}$, $m > 0$ and assume that (57) and (58) hold. Suppose that P_X is a Borel distribution on $\mathbb{B}^d$ with the property that there exist $c_0 > 0$ and a closed Euclidean ball $A \subseteq \mathbb{B}^d$ of radius $\tau > 0$ such that $P_X(A_0) \ge c_0 \mathrm{Vol}_d(A_0)$ for all measurable $A_0 \subseteq A$. Then there exist $c_0' \equiv c_0'(c_0, \tau, d, p)$, $c \equiv c(\alpha, d, p, c_0, C_0, \tau) > 0$ and $n_0 \equiv n_0(\alpha, d, p, c_0, C_0, S, m, \tau) \in \mathbb{N}$, such that*

$$\widetilde{\pi}_{\alpha,p}\Big(\Big\{g \in \mathcal{G}_{\alpha,p} : \|g - g^\circ\|_{L^2(P_X)}^2 \le c_0' m^2 n^{-2\alpha/(2\alpha+p)}\Big\}\Big) \le \exp\big(-cm^{-p/\alpha} n^{p/(2\alpha+p)}\big) \tag{63}$$

for $n \ge n_0$. Moreover, there exists $c' > 0$, depending only on $\alpha, \beta, d, p, c_0, C_0, S, \tau$ and σ, such that for $n \ge n_0$,

$$\pi_{\alpha,p}\big(B'_{\alpha,p}(P_{g^\circ}, c' m n^{-\alpha/(2\alpha+p)})\big) \le \exp\big(-cm^{-p/\alpha} n^{p/(2\alpha+p)}\big) \tag{64}$$

where $B'_{\alpha,p}(P_{g^\circ}, c' m n^{-\alpha/(2\alpha+p)}) := \big\{P \in \mathcal{P}_{\alpha,p} : d_{\mathrm{H}}(P_{g^\circ}, P) \le c' m n^{-\alpha/(2\alpha+p)}\big\}$ is defined as in (36).

Proof. Throughout the proof, $C_1, C_2, \ldots$ will denote positive quantities that may depend only on α, d, p, c_0, C_0 and τ. Let $c_0' \equiv c_0'(c_0, \tau, d, p)$ be the quantity in Lemma G.5. Let $\widetilde{g}_{\alpha,p}$ be defined as in (43) so that $\widetilde{g}_{\alpha,p} \sim \widetilde{\pi}_{\alpha,p}$. Then

$$\begin{aligned}
&\widetilde{\pi}_{\alpha,p}\Big(\Big\{g \in \mathcal{G}_{\alpha,p} : \|g - g^\circ\|_{L^2(P_X)}^2 \le c_0' m^2 n^{-2\alpha/(2\alpha+p)}\Big\}\Big) \\
&= \mathbb{P}\bigg\{\bigg\|\widetilde{g}_\alpha^{(p)}\bigg(\frac{(U^{(p)})^\top x + 1_p}{2}\bigg) - h^\circ\bigg(\frac{(U^\circ)^\top x + 1_p}{2}\bigg)\bigg\|_{L^2(P_X)}^2 \le c_0' m^2 n^{-2\alpha/(2\alpha+p)}\bigg\} \\
&= \mathbb{E}\bigg[\mathbb{P}\bigg\{\bigg\|\widetilde{g}_\alpha^{(p)}\bigg(\frac{(U^{(p)})^\top x + 1_p}{2}\bigg) - h^\circ\bigg(\frac{(U^\circ)^\top x + 1_p}{2}\bigg)\bigg\|_{L^2(P_X)}^2 \le c_0' m^2 n^{-2\alpha/(2\alpha+p)} \,\bigg|\, U^{(p)}\bigg\}\bigg].
\end{aligned} \tag{65}$$

Now conditional on $U^{(p)}$, we fix $g' \in \mathcal{G}_\alpha^{(p)}$ such that $\big\|g'\big(\frac{(U^{(p)})^\top x + 1_p}{2}\big) - h^\circ\big(\frac{(U^\circ)^\top x + 1_p}{2}\big)\big\|_{L^2(P_X)}^2 \le c_0' m^2 n^{-2\alpha/(2\alpha+p)}$. We may assume without loss of generality that such g' exists, since otherwise the conditional probability in the expectation in (65) would be zero. Using the same argument as that in (53), we deduce that

$$\begin{aligned}
&\mathbb{P}\bigg\{\bigg\|\widetilde{g}_\alpha^{(p)}\bigg(\frac{(U^{(p)})^\top x + 1}{2}\bigg) - h^\circ\bigg(\frac{(U^\circ)^\top x + 1_p}{2}\bigg)\bigg\|_{L^2(P_X)}^2 \le c_0' m^2 n^{-2\alpha/(2\alpha+p)} \,\bigg|\, U^{(p)}\bigg\} \\
&\le \mathbb{P}\bigg\{\bigg\|\widetilde{g}_\alpha^{(p)}\bigg(\frac{(U^{(p)})^\top x + 1_p}{2}\bigg) - g'\bigg(\frac{(U^{(p)})^\top x + 1_p}{2}\bigg)\bigg\|_{L^2(P_X)}^2 \le 4c_0' m^2 n^{-2\alpha/(2\alpha+p)} \,\bigg|\, U^{(p)}\bigg\} \\
&= \mathbb{P}\big(\|\widetilde{g}_\alpha^{(p)} - g'\|_{L^2(Q(U^{(p)}))}^2 \le 4c_0' m^2 n^{-2\alpha/(2\alpha+p)} \,\big|\, U^{(p)}\big),
\end{aligned} \tag{66}$$

where $Q(U^{(p)})$ denotes the conditional distribution of $\frac{(U^{(p)})^\top X + 1_p}{2}$ given $U^{(p)}$, with $X \sim P_X$. By Lemma G.5, for every $U \in V_p(\mathbb{R}^d)$, there exists a hypercube $A' \subseteq [0,1]^p$ of side length $\frac{\tau}{2\sqrt{p}}$ such that $Q(U)(A_0') \ge c_0' \mathrm{Vol}_p(A_0')$ for all measurable $A_0' \subseteq A'$. We may thus apply (51), with d replaced by p, m replaced by $4m$ and c_0 replaced by c_0' therein, to deduce that

$$\mathbb{P}\big\{\|\widetilde{g}_\alpha^{(p)} - g'\|_{L^2(Q(U^{(p)}))}^2 \le 4c_0' m^2 n^{-2\alpha/(2\alpha+p)} \,\big|\, U^{(p)}\big\} \le \exp\big(-cm^{-p/\alpha} n^{p/(2\alpha+p)}\big),$$

for $n \geq n_0$. Combining this with (65) and (66) proves (63).

Finally, the proof of (64) follows a very similar argument to that in (52). □

Lemma F.6. *Let $(\alpha, p) \in \mathcal{A}$, $C > 0$, $P_0 \in \mathcal{P}_{\alpha,p}$ and $\epsilon_{n,\alpha,p} := Cn^{-\alpha/(2\alpha+p)}$. Then there exists C' depending only on α, p, d, σ, C and C_0 such that*

$$\sup_{\epsilon \geq \epsilon_{n,\alpha,p}} \log \mathcal{N}\big(\epsilon/3, B'_{\alpha,p}(P_0, 2\epsilon), d_{\mathrm{H}}\big) \leq C' n \epsilon_{n,\alpha,p}^2.$$

Proof. Throughout this proof, $C_1, C_2, \ldots$ will denote positive quantities depending only on α, p, d, σ, C and C_0. By Lemma G.2*(d)*,

$$\begin{aligned}\sup_{\epsilon \geq \epsilon_{n,\alpha,p}} \log \mathcal{N}\big(\epsilon/3,\, B'_{\alpha,p}(P_0, 2\epsilon),\, d_{\mathrm{H}}\big) &\leq \log \mathcal{N}\big(\epsilon_{n,\alpha,p}/3,\, \mathcal{P}_{\alpha,p},\, d_{\mathrm{H}}\big) \\ &\leq \log \mathcal{N}\big(2\sigma\epsilon_{n,\alpha,p}/3,\, \mathcal{G}_{\alpha,p},\, \|\cdot\|_\infty\big).\end{aligned}$$

Recall that $\mathcal{G}_\alpha^{(p)} \subseteq B^\alpha_{\infty,\infty}([0,1]^p, 2C_0)$ and $\mathcal{G}_{\alpha,p} = \mathcal{G}_\alpha^{(p)} \circ \big\{x \mapsto \frac{U^\top x + 1_p}{2} : U \in V_p(\mathbb{R}^d)\big\}$. Now let $\mathcal{N}_1$ be a $(\sigma\epsilon_{n,\alpha,p}/3)$-cover of $B^\alpha_{\infty,\infty}([0,1]^p, 2C_0)$ with respect to $\|\cdot\|_\infty$ such that $|\mathcal{N}_1| \leq \exp(C_1 n \epsilon_{n,\alpha,p}^2)$ for some $C_1 > 0$, which exists by, e.g. Giné and Nickl (2021, Eq. (4.184)). Recall from (61) that for all $x, y \in [0,1]^p$,

$$|g(x) - g(y)| \leq C_2 \|x - y\|_2^{\alpha \wedge 1} \big\{\log(1/\|x-y\|_2) \vee 1\big\}. \tag{67}$$

The function $z \mapsto z^{\alpha\wedge 1}\log(1/z)$ is increasing on $(0, e^{-1/(\alpha\wedge 1)}]$, so let $\delta := \frac{1}{\sqrt{pd}}\big(\frac{\sigma\epsilon_{n,\alpha,p}}{3C_3\log(en)}\big)^{1/(\alpha\wedge 1)} \wedge \frac{1}{\sqrt{pd}} e^{-1/(\alpha\wedge 1)}$ for $C_3 > 0$ large enough such that

$$C_2(\delta\sqrt{pd})^{\alpha\wedge 1}\left\{\log\left(\frac{1}{\delta\sqrt{pd}}\right) \vee 1\right\} \leq \frac{\sigma\epsilon_{n,\alpha,p}}{3}. \tag{68}$$

Let $\Delta := \{i\delta : i \in \mathbb{Z}, |i\delta| \leq 1\}$ and $\mathcal{N}_2 := \Delta^{d\times p} \subseteq \mathbb{R}^{d\times p}$. Then for any $g \in \mathcal{G}_\alpha^{(p)} \subseteq B^\alpha_{\infty,\infty}([0,1]^p, 2C_0)$ and $U \in V_p(\mathbb{R}^d)$, there exist $g' \in \mathcal{N}_1$ and $U' \in \mathcal{N}_2$ such that $\|g' - g\|_\infty \leq \sigma\epsilon_{n,\alpha,p}/3$ and $\|U' - U\|_{\mathrm{op}} \leq \sqrt{pd}\|U' - U\|_\infty \leq \delta\sqrt{pd}$. Thus, for all $x \in \mathbb{B}^d$, we have $\|U^\top x - (U')^\top x\|_2 \leq \delta\sqrt{pd}$, so by (67) and (68), we deduce that

$$\begin{aligned}&\left|g\left(\frac{U^\top x + 1_p}{2}\right) - g'\left(\frac{(U')^\top x + 1_p}{2}\right)\right| \\ &\quad\leq \left|g\left(\frac{U^\top x + 1_p}{2}\right) - g\left(\frac{(U')^\top x + 1_p}{2}\right)\right| + \left|g\left(\frac{(U')^\top x + 1_p}{2}\right) - g'\left(\frac{(U')^\top x + 1_p}{2}\right)\right| \\ &\quad\leq \frac{2\sigma\epsilon_{n,\alpha,p}}{3}.\end{aligned}$$

Therefore, $\mathcal{N}_1 \circ \mathcal{N}_2$ is a $(2\sigma\epsilon_{n,\alpha,p}/3)$-cover of $\mathcal{G}_{\alpha,p}$ with respect to $\|\cdot\|_\infty$. Hence, we have

$$\log \mathcal{N}\big(2\sigma\epsilon_{n,\alpha,p}/3,\, \mathcal{G}_{\alpha,p},\, \|\cdot\|_\infty\big) \leq \log|\mathcal{N}_1| + \log|\mathcal{N}_2| \leq C' n \epsilon_{n,\alpha,p}^2, \tag{69}$$

which proves the claim. □

Proposition F.7. *For any $(\beta, r) \in \mathcal{A}'$ and $g^\circ \in \mathcal{G}_{\beta,r}$, write $P_0 \equiv P_{g^\circ}$, let π be the prior distribution defined in (7) and let g_π denote the posterior regression function with respect to π. Then there exists $C > 0$, not depending on n, such that*

$$\mathbb{E}_{P_0}\big[\big\{g_\pi(X, \mathcal{D}_n) - g^\circ(X)\big\}^2\big] \leq Cn^{-2\beta/(2\beta+r)}$$

for all $n \in \mathbb{N}$.

Proof of Proposition F.7. The proof of Proposition F.7 follows the same arguments as the proof of Proposition 4.1, but instead we apply Lemmas F.4, F.5 and F.6 to verify the assumptions of Corollary E.2. □

F.5 Proofs of Theorems 4.2 and 4.3

Proof of Theorem 4.2. For any test distribution μ on $\mathcal{P}_\beta$ such that $\chi^2(\mu, \pi_\beta) \leq \kappa$, we have

$$\chi^2(\mu, \pi) = \mathbb{E}_\pi\left\{\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2\right\} - 1 \leq \max_{\alpha \in \mathcal{A}} \frac{1}{\lambda_\alpha} \cdot \mathbb{E}_{\pi_\beta}\left\{\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi_\beta}\right)^2\right\} - 1 \leq (\kappa + 1) \max_{\alpha \in \mathcal{A}} \frac{1}{\lambda_\alpha} - 1 =: \kappa'.$$

Now, by Proposition 2.1, then by taking $d^\circ_{\mathrm{model}} \geq d + 2$, $d^\circ_{\mathrm{ffn}} \in \mathbb{N}$ and $T \in \mathbb{N}$ large enough that Proposition 3.2 holds with $\epsilon = 1/\{16R^2n^2(\kappa' + 1)\}$ and applying Proposition 4.1, we have

$$\begin{aligned}\mathcal{R}^{\mathrm{ICL}}_\mu(\widetilde{f}_T) &\leq 4R\sqrt{\{\chi^2(\mu, \pi) + 1\}\mathcal{E}(\widetilde{f}_T)} + 2\mathbb{E}_\mu\mathbb{E}_P\big[\{g_\pi(X, \mathcal{D}_n) - g^\circ(X)\}^2\big] \\ &\leq \frac{1}{n} + C'n^{-2\beta/(2\beta+d)} \leq Cn^{-2\beta/(2\beta+d)},\end{aligned}$$

where $C', C > 0$ depend on neither n nor κ. □

Proof of Theorem 4.3. As in the proof of Theorem 4.2, for any test distribution μ on $\mathcal{P}_{\beta,r}$ such that $\chi^2(\mu, \pi_{\beta,r}) \leq \kappa$, we have $\chi^2(\mu, \pi) \leq (\kappa + 1)\max_{(\alpha,p)\in\mathcal{A}'} \lambda^{-1}_{\alpha,p} - 1 =: \kappa'$. The result therefore follows by the same argument as in the proof of Theorem 4.2, except that we apply Proposition F.7 in place of Proposition 4.1 to obtain the improved bound $\mathcal{R}^{\mathrm{ICL}}_\mu(\widetilde{f}_T) \leq Cn^{-2\beta/(2\beta+r)}$ for sufficiently large T, where $C > 0$ depends on neither n nor κ. □

F.6 Proof of Theorem 4.4

Proof of Theorem 4.4. We use the Bayes risk lower bound techniques developed by Chen, Guntuboyina and Zhang (2016). Let $(\mathcal{P}_*, \mathcal{G}_*)$ be either $(\mathcal{P}_\beta, \mathcal{G}_\beta)$ defined as in Section 4.1 or $(\mathcal{P}_{\beta,r}, \mathcal{G}_{\beta,r})$ defined as in Section 4.2, and let $\widetilde{\mu}$ denote the distribution on $\mathcal{G}_*$ induced by μ on $\mathcal{P}_*$, so that $\widetilde{\mu}$ is the distribution of the random function $x \mapsto \mathbb{E}_P(Y \mid X = x)$, where the randomness comes from $P \sim \mu$. By Chen, Guntuboyina and Zhang (2016, Corollary 12*(i)* and Eq. (48)), we have

$$\begin{aligned}\inf_{\widehat{g}_n \in \widehat{\mathcal{G}}_n} \mathcal{R}^{\mathrm{ICL}}_\mu(\widehat{g}_n) &= \inf_{\widehat{g}_n \in \widehat{\mathcal{G}}_n} \mathbb{E}_{P\sim\mu}\mathbb{E}_{\mathcal{D}_n|P\sim P^{\otimes n}}\big\{\|\widehat{g}_n(\cdot, \mathcal{D}_n) - g_P\|^2_{L^2(P_X)}\big\} \\ &= \inf_{\widehat{g}_n \in \widehat{\mathcal{G}}_n} \mathbb{E}_{\widetilde{g}\sim\widetilde{\mu}}\mathbb{E}_{\mathcal{D}_n|\widetilde{g}\sim P^{\otimes n}_{\widetilde{g}}}\big\{\|\widehat{g}_n(\cdot, \mathcal{D}_n) - \widetilde{g}\|^2_{L^2(P_X)}\big\} \\ &\geq \frac{1}{2}\sup\bigg\{t > 0 : \sup_{g^\circ \in \mathcal{G}_*} \widetilde{\mu}\Big(\big\{g \in \mathcal{G}_* : \|g - g^\circ\|^2_{L^2(P_X)} < t\big\}\Big) < \frac{1}{4}e^{-2I^{\mathrm{up}}_{\mathrm{KL}}}\bigg\},\end{aligned} \tag{70}$$

where

$$I^{\mathrm{up}}_{\mathrm{KL}} := \inf_{\eta > 0}\big\{\log \mathcal{N}(\eta^2/n, \mathcal{P}_*, \mathrm{KL}) + \eta^2\big\}$$

and we recall that $\mathcal{N}(\eta^2/n, \mathcal{P}_*, \mathrm{KL})$ is the (η^2/n)-covering number of $\mathcal{P}_*$ with respect to the Kullback–Leibler divergence. By Lemma G.2*(a)*, we deduce that

$$I^{\mathrm{up}}_{\mathrm{KL}} \leq \inf_{\eta > 0}\big\{\log \mathcal{N}\big(\sigma\eta\sqrt{2/n}, \mathcal{G}_*, \|\cdot\|_\infty\big) + \eta^2\big\}.$$

From now on, $C_1, C_2, \ldots$ will denote positive quantities that do not depend on n.

(a) By choosing $\eta := n^{d/(2\beta+d)}$ and Giné and Nickl (2021, Eq. (4.184)), we deduce that

$$I^{\mathrm{up}}_{\mathrm{KL}} \leq C_1 n^{d/(2\beta+d)}.$$

By the measurable bijection between $\mathcal{G}_\beta$ and $\mathcal{P}_\beta$ and our assumption on μ, we have that $\sup_{g\in\mathcal{G}_\beta} \frac{\mathrm{d}\widetilde{\mu}}{\mathrm{d}\widetilde{\pi}_\beta}(g) < \infty$. Thus, by Lemma F.3, there exists $C_2 > 0$ small enough that for all $g^\circ \in \mathcal{G}_\beta$,

$$\begin{aligned}&\widetilde{\mu}\Big(\big\{g \in \mathcal{G}_\beta : \|g - g^\circ\|^2_{L^2(P_X)} < C_2 n^{-2\beta/(2\beta+d)}\big\}\Big) \\ &\qquad \leq \bigg\{\sup_{g\in\mathcal{G}_\beta} \frac{\mathrm{d}\widetilde{\mu}}{\mathrm{d}\widetilde{\pi}_\beta}(g)\bigg\}\widetilde{\pi}_\beta\Big(\big\{g \in \mathcal{G}_\beta : \|g - g^\circ\|^2_{L^2(P_X)} < C_2 n^{-2\beta/(2\beta+d)}\big\}\Big) < \frac{1}{4}e^{-2I^{\mathrm{up}}_{\mathrm{KL}}}.\end{aligned}$$

The claim then follows from (70).

(b) By choosing $\eta := n^{d/(2\beta+r)}$ and applying (69) with (β, r) replacing (α, p), we deduce that

$$I_{\mathrm{KL}}^{\mathrm{up}} \leq C_3 n^{r/(2\beta+r)}.$$

Similarly, by Lemma F.5, there exists $C_4 > 0$ small enough that for all $g^\circ \in \mathcal{G}_{\beta,r}$,

$$\begin{aligned}
&\widetilde{\mu}\Big(\big\{g \in \mathcal{G}_{\beta,r} : \|g - g^\circ\|_{L^2(P_X)}^2 < C_4 n^{-2\beta/(2\beta+r)}\big\}\Big) \\
&\quad \leq \Big\{\sup_{g \in \mathcal{G}_{\beta,r}} \frac{\mathrm{d}\widetilde{\mu}}{\mathrm{d}\widetilde{\pi}_{\beta,r}}(g)\Big\} \widetilde{\pi}_{\beta,r}\Big(\big\{g \in \mathcal{G}_{\beta,r} : \|g - g^\circ\|_{L^2(P_X)}^2 < C_4 n^{-2\beta/(2\beta+r)}\big\}\Big) < \frac{1}{4} e^{-2I_{\mathrm{KL}}^{\mathrm{up}}}.
\end{aligned}$$

The claim then follows again from (70). □

G Auxiliary lemmas

The lemma below is a modification of Ghosal, Lember and van der Vaart (2008, Lemma 6.3), where we assume that the log-likelihood ratio is sub-Gamma in the left tail with uniform sub-Gamma parameters in order to obtain an exponential concentration.

Lemma G.1. *Let $\mathcal{P}$ be a measurable space of distributions on a measurable space $\mathcal{Z}$, let P_0 be a distribution on $\mathcal{Z}$ and let $Z_1, \ldots, Z_n \overset{\mathrm{iid}}{\sim} P_0$. For $\epsilon > 0$, define*

$$B(P_0, \epsilon) := \Big\{P \in \mathcal{P} : \mathrm{KL}(P_0, P) \leq \epsilon^2,\ \mathrm{V}_2(P_0, P) \leq \epsilon^2\Big\}.$$

For $i \in [n]$ and $P \in \mathcal{P}$ with $P \ll P_0$, let

$$W(P, Z_i) := \log \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i) - \mathbb{E} \log \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i).$$

Assume that there exist $a, c > 0$ such that for all $P \in B(P_0, \epsilon)$, we have that $W(P, Z_1)$ is sub-Gamma in the left tail with variance parameter $a\epsilon^2$ and scale parameter c. Then, for any probability measure Π on $\mathcal{P}$ and any $\epsilon, D > 0$, we have with $P_0^{\otimes n}$-probability at most $\exp\big(-\frac{D^2 n\epsilon^2}{2(a+cD)}\big)$ that

$$\int_{B(P_0,\epsilon)} \prod_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i)\, \mathrm{d}\Pi(P) < \Pi\big(B(P_0, \epsilon)\big) e^{-(1+D)n\epsilon^2}. \tag{71}$$

Proof. Writing $B := B(P_0, \epsilon)$, we observe that the conclusion is clear if $\Pi(B) = 0$ because the right-hand side of (71) is zero. Moreover, by replacing Π with its conditional distribution on B if necessary, we may assume without loss of generality that $\Pi(B) = 1$. Now, by Jensen's inequality,

$$\begin{aligned}
\frac{1}{n} \log \int_B \prod_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i)\, \mathrm{d}\Pi(P) &\geq \frac{1}{n} \sum_{i=1}^n \int_B \log \frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i)\, \mathrm{d}\Pi(P) \\
&= \frac{1}{n} \sum_{i=1}^n \int_B W(P, Z_i)\, \mathrm{d}\Pi(P) - \int_B \mathrm{KL}(P_0, P)\, \mathrm{d}\Pi(P) \\
&\geq \frac{1}{n} \sum_{i=1}^n \int_B W(P, Z_i)\, \mathrm{d}\Pi(P) - \epsilon^2,
\end{aligned} \tag{72}$$

where the final inequality follows from the definition of B. Moreover, for $\lambda \in (-1/c, 0]$, by Jensen's inequality and Fubini's theorem,

$$\begin{aligned}
\mathbb{E}_{Z_1} \exp\Big\{\lambda \int_B W(P, Z_1)\, \mathrm{d}\Pi(P)\Big\} &\leq \mathbb{E}_{Z_1}\Big[\int_B \exp\big\{\lambda W(P, Z_1)\big\}\, \mathrm{d}\Pi(P)\Big] \\
&= \int_B \mathbb{E}_{Z_1} \exp\big\{\lambda W(P, Z_1)\big\}\, \mathrm{d}\Pi(P) \leq \exp\Big(\frac{a\epsilon^2\lambda^2}{2(1 - c\lambda)}\Big),
\end{aligned}$$

where the final inequality follows from the sub-Gamma assumption on $W(P, Z_1)$. This shows that $\int_B W(P, Z_1)\,\mathrm{d}\Pi(P)$ is sub-Gamma in the left tail with variance parameter $a\epsilon^2$ and scale parameter c, so $n^{-1}\sum_{i=1}^n \int_B W(P, Z_i)\,\mathrm{d}\Pi(P)$ is also sub-Gamma in the left tail with variance parameter $a\epsilon^2/n$ and scale parameter c/n. Hence, by (72),

$$\mathbb{P}\bigg(\frac{1}{n}\log\int_B\prod_{i=1}^n\frac{\mathrm{d}P}{\mathrm{d}P_0}(Z_i)\,\mathrm{d}\Pi(P) < -(1+D)\epsilon^2\bigg) \\ \leq \mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n\int_B W(P, Z_i)\,\mathrm{d}\Pi(P) < -D\epsilon^2\bigg) \leq \exp\bigg(-\frac{D^2 n\epsilon^2}{2(a+cD)}\bigg),$$

where the final inequality follows from Bernstein's inequality (Boucheron, Lugosi and Massart, 2013, Corollary 2.11). □

We next derive some properties about the distributions of nonparametric regression models with Gaussian noise.

Lemma G.2. *Let P_X be a Borel distribution on $\mathcal{X} \subseteq \mathbb{R}^d$, let $X \sim P_X$ and let $\xi \sim N(0, \sigma^2)$ be independent of X. Further, let $R > 0$, $g_1, g_2 : \mathcal{X} \to [-R, R]$ be Borel measurable, let $Y_1 = g_1(X) + \xi$ and $Y_2 = g_2(X) + \xi$. Finally, let P_1 and P_2 denote the distributions of (X, Y_1) and (X, Y_2) respectively. Then*

(a) $\mathrm{KL}(P_1, P_2) = \frac{1}{2\sigma^2}\mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\}$.

(b) $\mathrm{V}_2(P_1, P_2) = \frac{1}{2\sigma^2}\mathrm{Var}\big\{\big(g_1(X) - g_2(X)\big)^2\big\} + 2\mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\}$.

(c) If $\mathrm{KL}(P_1, P_2) \leq \epsilon^2$ and $\mathrm{V}_2(P_1, P_2) \leq \epsilon^2$, then $W := \log\frac{\mathrm{d}P_2}{\mathrm{d}P_1}(X, Y_1) - \mathbb{E}\log\frac{\mathrm{d}P_2}{\mathrm{d}P_1}(X, Y_1)$ is sub-Gamma in both tails with variance parameter $\big(8 \vee \frac{2}{\sigma^2}\big)\epsilon^2$ and scale parameter $c := \frac{2R}{\sigma} \vee \frac{4R^2}{3\sigma^2}$.

(d) $\frac{1-e^{-R^2/(2\sigma^2)}}{2R^2} \cdot \mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\} \leq d_{\mathrm{H}}^2(P_1, P_2) \leq \frac{1}{4\sigma^2} \cdot \mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\}$.

Proof. *(a)* Let $P_{Y_1|X}$ and $P_{Y_2|X}$ denote the conditional distributions of Y_1 given X and Y_2 given X respectively. Then, using the fact that $Y_1 = g_1(X) + \xi$,

$$\log\frac{\mathrm{d}P_2}{\mathrm{d}P_1}(X, Y_1) = \log\frac{\mathrm{d}P_{Y_2|X}}{\mathrm{d}P_{Y_1|X}}(X, Y_1) = -\frac{\big(Y_1 - g_2(X)\big)^2}{2\sigma^2} + \frac{\big(Y_1 - g_1(X)\big)^2}{2\sigma^2} \\ = \frac{-\big(g_1(X) - g_2(X)\big)^2 - 2\big(g_1(X) - g_2(X)\big)\xi}{2\sigma^2}. \tag{73}$$

Hence,

$$\mathrm{KL}(P_1, P_2) = -\mathbb{E}\log\frac{\mathrm{d}P_2}{\mathrm{d}P_1}(X, Y_1) = \frac{1}{2\sigma^2}\mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\}. \tag{74}$$

(b) By (73) and (74), we have

$$\mathrm{V}_2(P_1, P_2) = \frac{1}{2\sigma^2}\mathbb{E}\Big[\Big\{\big(g_1(X) - g_2(X)\big)^2 + 2\big(g_1(X) - g_2(X)\big)\xi - \mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\}\Big\}^2\Big] \\ = \frac{1}{2\sigma^2}\mathrm{Var}\big\{\big(g_1(X) - g_2(X)\big)^2\big\} + 2\mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\},$$

where in the final equality, we used the fact that $\xi \perp\!\!\!\perp X$, $\mathbb{E}(\xi) = 0$ and $\mathbb{E}(\xi^2) = \sigma^2$.

(c) From (73) and (74), we deduce that

$$W = \log\frac{\mathrm{d}P_2}{\mathrm{d}P_1}(X, Y_1) - \mathbb{E}\log\frac{\mathrm{d}P_2}{\mathrm{d}P_1}(X, Y_1) \\ = \frac{\big(g_2(X) - g_1(X)\big)\xi}{\sigma^2} - \frac{\big(g_1(X) - g_2(X)\big)^2 - \mathbb{E}\big\{\big(g_1(X) - g_2(X)\big)^2\big\}}{2\sigma^2} =: W_1 + W_2.$$

For $\lambda \in [0, \sigma/R)$, using the fact that $\log(1+x) \le x$ for $x \ge 0$ and Fubini's theorem,

$$\begin{aligned}
\log \mathbb{E} e^{\lambda W_1} &= \log \mathbb{E} \exp\biggl(\frac{\lambda\bigl(g_2(X) - g_1(X)\bigr)\xi}{\sigma^2}\biggr) = \log \mathbb{E} \exp\biggl(\frac{\lambda^2\bigl(g_2(X) - g_1(X)\bigr)^2}{2\sigma^2}\biggr) \\
&\le \sum_{r=1}^{\infty} \frac{\lambda^{2r} \mathbb{E}\bigl\{\bigl(g_2(X) - g_1(X)\bigr)^{2r}\bigr\}}{(2\sigma^2)^r \cdot r!} \le \mathbb{E}\bigl\{\bigl(g_2(X) - g_1(X)\bigr)^2\bigr\} \sum_{r=1}^{\infty} \frac{\lambda^{2r}(2R)^{2r-2}}{(2\sigma^2)^r \cdot r!} \\
&\le \lambda^2\epsilon^2 \sum_{r=1}^{\infty} \frac{2^{r-1}}{r!}\biggl(\frac{\lambda R}{\sigma}\biggr)^{2r-2} \le \lambda^2\epsilon^2 \sum_{r=1}^{\infty}\biggl(\frac{\lambda R}{\sigma}\biggr)^{r-1} = \frac{2\lambda^2\epsilon^2}{2\bigl(1 - \frac{\lambda R}{\sigma}\bigr)}.
\end{aligned}$$

The same holds true if W_1 is replaced by $-W_1$. Thus W_1 is sub-Gamma in both tails with variance parameter $2\epsilon^2$ and scale parameter R/σ. By part *(b)* and the assumption that $\mathrm{V}_2(P_1, P_2) \le \epsilon^2$, we have $\mathbb{E}(W_2^2) \le \epsilon^2/(2\sigma^2)$. Moreover, since W_2 is bounded by $2R^2/\sigma^2$, we have for $r \ge 3$ that

$$\mathbb{E}(|W_2|^r) \le \mathbb{E}(W_2^2) \cdot \biggl(\frac{2R^2}{\sigma^2}\biggr)^{r-2} \le \frac{r!\epsilon^2}{2 \cdot 2\sigma^2}\biggl(\frac{2R^2}{3\sigma^2}\biggr)^{r-2}.$$

The same bound holds for $-W_2$, so by Boucheron, Lugosi and Massart (2013, Theorem 2.10), W_2 is sub-Gamma in both tails with variance parameter $\frac{\epsilon^2}{2\sigma^2}$ and scale parameter $\frac{2R^2}{3\sigma^2}$. Finally, it follows that that $W = W_1 + W_2$ is sub-Gamma in both tails with variance parameter $\bigl(8 \vee \frac{2}{\sigma^2}\bigr)\epsilon^2$ and scale parameter $c = \frac{2R}{\sigma} \vee \frac{4R^2}{3\sigma^2}$.

(d) We have

$$\begin{aligned}
d_{\mathrm{H}}^2(P_1, P_2) = \mathbb{E}\, d_{\mathrm{H}}^2(P_{Y_1|X}, P_{Y_2|X}) &= 2\,\mathbb{E}\biggl\{1 - \exp\biggl(-\frac{\bigl(g_1(X) - g_2(X)\bigr)^2}{8\sigma^2}\biggr)\biggr\} \\
&\ge \frac{1 - e^{-R^2/(2\sigma^2)}}{2R^2} \cdot \mathbb{E}\bigl\{\bigl(g_1(X) - g_2(X)\bigr)^2\bigr\},
\end{aligned}$$

where the inequality follows since $x \mapsto (1 - e^{-x})/x$ is decreasing on $(0, \infty)$ and moreover $\frac{(g_1(X) - g_2(X))^2}{8\sigma^2} \le \frac{R^2}{2\sigma^2}$. On the other hand, using the fact that $1 - e^{-x} \le x$ for $x \ge 0$, we also have

$$2\mathbb{E}\biggl\{1 - \exp\biggl(-\frac{\bigl(g_1(X) - g_2(X)\bigr)^2}{8\sigma^2}\biggr)\biggr\} \le \frac{1}{4\sigma^2} \cdot \mathbb{E}\bigl\{\bigl(g_1(X) - g_2(X)\bigr)^2\bigr\},$$

as required. □

Lemma G.3. *Suppose $f \in L^2([0,1]^d)$ has wavelet decomposition*

$$f = \sum_{k \in K} a_k \Phi_k + \sum_{\ell=\ell_0}^{\infty} \sum_{\gamma \in \Gamma_\ell} b_{\ell,\gamma} \Psi_{\ell,\gamma}.$$

Let $0 \le c_1 < c_2 \le 1$ and let $\Lambda := \bigl\{(\ell, \gamma) : \mathrm{supp}(\Psi_{\ell,\gamma}) \subseteq [c_1, c_2]^d\bigr\}$. Then

$$\|f\|_{L^2([c_1,c_2]^d)}^2 \ge \sum_{(\ell,\gamma) \in \Lambda} b_{\ell,\gamma}^2.$$

Proof. Note that $\mathrm{supp}(\Psi_{\ell,\gamma}) \subseteq [c_1, c_2]^d$ for all $(\ell, \gamma) \in \Lambda$, so by the orthonormality of the wavelet basis, we have that $(\Psi_{\ell,\gamma})_{(\ell,\gamma)\in\Lambda}$ is also a sequence of orthonormal functions in the Hilbert space $L^2([c_1, c_2]^d)$ equipped with the inner product

$$\langle f_1, f_2 \rangle_{[c_1,c_2]^d} := \int_{[c_1,c_2]^d} f_1(x) f_2(x)\,\mathrm{d}x.$$

Moreover, for $(\ell, \gamma) \in \Lambda$, we have

$$\langle f, \Psi_{\ell,\gamma} \rangle_{[c_1,c_2]^d} = \int_{[c_1,c_2]^d} f(x)\Psi_{\ell,\gamma}(x)\,\mathrm{d}x = \int_{[0,1]^d} f(x)\Psi_{\ell,\gamma}(x)\,\mathrm{d}x = b_{\ell,\gamma}.$$

Thus, an application of Bessel's inequality (e.g. Rudin, 1987, Theorem 4.17) yields the desired result. □

Lemma G.4. *Let $U^{(r)}$ be uniformly distributed on the Stiefel manifold $V_r(\mathbb{R}^d)$ and let $U \in V_r(\mathbb{R}^d)$ be deterministic. Then for any $\epsilon > 0$, there exists $c > 0$, depending only on d and r, such that*

$$\mathbb{P}\big(\|U^{(r)} - U\|_{\mathrm{op}} \le \epsilon\big) \ge c\bigg(\epsilon \wedge \frac{1}{2}\bigg)^{dr}.$$

Proof. We may assume that $\epsilon \le 1/2$. By Chikuse (2003, Theorem 2.2.1(iii)), we may assume without loss of generality that $U^{(r)} = Z(Z^\top Z)^{-1/2} =: F(Z)$, where Z is a $d \times r$ random matrix with independent $N(0,1)$ entries. Let $\Delta \in \mathbb{R}^{d\times r}$ be such that $\delta := \|\Delta\|_{\mathrm{op}} \le 1/6$, and define

$$A := U^\top \Delta + \Delta^\top U + \Delta^\top \Delta \in \mathbb{R}^{r\times r}.$$

Then $\|A\|_{\mathrm{op}} \le 3\delta \le 1/2$, so $(U+\Delta)^\top(U+\Delta) = I_r + A$ is positive definite. Thus

$$\begin{aligned}\|F(U+\Delta) - U\|_{\mathrm{op}} &= \|(U+\Delta)(I_r+A)^{-1/2} - U\|_{\mathrm{op}} \\ &\le \|U\|_{\mathrm{op}}\|(I_r+A)^{-1/2} - I_r\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}}\|(I_r+A)^{-1/2}\|_{\mathrm{op}}.\end{aligned}$$

Moreover, we can write $I_r + A = VDV^\top$ where $V \in \mathbb{R}^{r\times r}$ is orthogonal and $D \in \mathbb{R}^{r\times r}$ is diagonal with diagonal elements bounded between $1-3\delta$ and $1+3\delta$. Thus, $\|(I_r+A)^{-1/2} - I_r\|_{\mathrm{op}} \le 3\delta$ and $\|(I_r+A)^{-1/2}\|_{\mathrm{op}} \le 2$, so

$$\|F(U+\Delta) - U\|_{\mathrm{op}} \le 3\delta + 2\delta = 5\delta.$$

Hence, if $\|Z - U\|_{\mathrm{op}} \le \epsilon/5 < 1/6$, then $\|F(Z) - U\|_{\mathrm{op}} \le \epsilon$, so

$$\begin{aligned}\mathbb{P}\big(\|U^{(r)} - U\|_{\mathrm{op}} \le \epsilon\big) &= \mathbb{P}\big(\|F(Z) - U\|_{\mathrm{op}} \le \epsilon\big) \ge \mathbb{P}\bigg(\|Z - U\|_{\mathrm{op}} \le \frac{\epsilon}{5}\bigg) \\ &\ge \mathbb{P}\bigg(\|Z - U\|_{\infty} \le \frac{\epsilon}{5\sqrt{dr}}\bigg) \ge \bigg\{\mathbb{P}\bigg(|Z_{11} - 1| \le \frac{\epsilon}{5\sqrt{dr}}\bigg)\bigg\}^{dr} \ge \bigg(\frac{2\epsilon}{25\sqrt{dr}}\bigg)^{dr},\end{aligned}$$

where the third inequality follows since $\|U\|_\infty \le 1$ and Z has independent $N(0,1)$ entries, and the final inequality follows since $\frac{\epsilon}{5\sqrt{dr}} \le \frac{1}{10}$ and the standard normal density is bounded below by $1/5$ on the interval $[0.9, 1.1]$. □

Lemma G.5. *Let P_X be a distribution on $\mathbb{B}^d$ that satisfies the assumptions of Lemma F.5. Let $p \in [d]$, $U \in V_p(\mathbb{R}^d)$, $X \sim P_X$ and let Q denote the distribution of $\frac{U^\top X + 1_p}{2}$. Then there exist $c_0' \equiv c_0'(c_0, \tau, d, p)$ and a hypercube $A' \subseteq [0,1]^p$ of side length $\frac{\tau}{2\sqrt{p}}$ with the property that $Q(A_0') \ge c_0' \mathrm{Vol}_p(A_0')$ for all measurable $A_0' \subseteq A'$.*

Proof. If $p = d$, then $U \in \mathbb{R}^{d\times d}$ is orthogonal, so for all measurable $A_0 \subseteq A$,

$$Q\bigg(\frac{U^\top A_0 + 1_d}{2}\bigg) = \mathbb{P}(X \in A_0) \ge c_0 \mathrm{Vol}_d(A_0).$$

Moreover, $(U^\top A + 1_d)/2$ is a Euclidean ball of radius $\tau/2$, so it contains a hypercube of side length $\frac{\tau}{2\sqrt{p}}$. Now suppose that $p \in [d-1]$. For $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$ and $1 \le j_1 \le j_2 \le d$, we write $v_{j_1:j_2} := (v_{j_1}, \ldots, v_{j_2})^\top \in \mathbb{R}^{j_1 - j_2 + 1}$. Let $o \in \mathbb{R}^d$ be the centre of A and let $K \subseteq A$ be the Euclidean ball of radius $\tau/2$ and centred at o. We may write $U = VP$ where $V \in \mathbb{R}^{d\times d}$ is orthogonal and $P \in \mathbb{R}^{d\times p}$ is the first p columns of I_d. Let $S \subseteq K$ be Borel measurable, and define $T := \{x \in V^\top A : x_{1:p} \in P^\top V^\top S\}$. Then $VT \subseteq A$; moreover, if $y = (V^\top x)_{1:p}$ for some $x \in S \subseteq K$ and $z \in \mathbb{R}^{d-p}$ satisfies $\|z - (V^\top o)_{(p+1):d}\|_2 \le \tau/2$, then by the choice of K,

$$\bigg\|\begin{pmatrix} y \\ z\end{pmatrix} - V^\top o\bigg\|_2^2 = \|y - (V^\top o)_{1:p}\|_2^2 + \|z - (V^\top o)_{(p+1):d}\|_2^2 \le (\tau/2)^2 + (\tau/2)^2 < \tau^2,$$

so $(y^\top, z^\top)^\top \in T$. Thus

$$T' := \big\{(y^\top, z^\top)^\top \in \mathbb{R}^{p+(d-p)} : y \in P^\top V^\top S,\ \|z - (V^\top o)_{(p+1):d}\|_2 \le \tau/2\big\} \subseteq T. \tag{75}$$

Therefore, writing $v(\tau, d, p)$ for the Lebesgue measure of a Euclidean ball in $\mathbb{R}^{d-p}$ of radius $\tau/2$, we deduce that

$$\begin{aligned} Q\left(\frac{U^\top S + 1_p}{2}\right) &= \mathbb{P}\left(\frac{P^\top V^\top X + 1_p}{2} \in \frac{P^\top V^\top S + 1_p}{2}\right) \\ &\geq \mathbb{P}(V^\top X \in T) = \mathbb{P}(X \in VT) \geq c_0 \mathrm{Vol}_d(VT) = c_0 \mathrm{Vol}_d(T) \geq c_0 \mathrm{Vol}_d(T') \\ &= c_0 \int_{y \in P^\top V^\top S} \int_{z \in \mathbb{R}^{d-p} : \|z - (V^\top o)_{(p+1):d}\|_2 \leq \tau/2} \mathrm{d}z \, \mathrm{d}y \\ &= c_0 v(\tau, d, p) \mathrm{Vol}_p(P^\top V^\top S) =: c_0' \mathrm{Vol}_p\left(\frac{U^\top S + 1_p}{2}\right). \end{aligned}$$

Since this holds for all measurable $S \subseteq K$, we deduce that $Q(A_0') \geq c_0' \mathrm{Vol}_p(A_0')$ for all $A_0' \subseteq \frac{U^\top K + 1_p}{2}$. Now $\frac{U^\top K + 1_p}{2}$ is a ball of radius $\tau/4$ in $\mathbb{R}^p$, so it contains a hypercube A' of side length $\frac{\tau}{2\sqrt{p}}$. □

H ICL Example

When we provide ChatGPT with the first two in-context examples and the query from the right column of Figure 1 (i.e. the prompt is 'Cat → 10, Fish → 25, Bird → ?'), we obtain the following output (though this may vary due to uncertainty in the sampling procedure):

> There isn't a unique answer from just Cat → 10 and Fish → 25.
> With only two example pairs, infinitely many rules can fit them, and those rules would give different values for Bird. So Bird → ? is underdetermined (cannot be determined uniquely).
> If you give one more example mapping (any other word → number), I can pin down the rule and compute Bird.

When ChatGPT is provided with all four examples and the query from the right column of Figure 1 (i.e. now the prompt is 'Cat → 10, Fish → 25, Deer → 17, Dog → 13, Bird → ?'), we obtain the following:

> The pattern is:
> Number = (number of letters in the word) × (alphabet position of the first letter) + 1
> Check:
> • Cat → 3 × C(3) + 1 = 10
> • Fish → 4 × F(6) + 1 = 25
> • Deer → 4 × D(4) + 1 = 17
> • Dog → 3 × D(4) + 1 = 13
> So:
> • Bird → 4 × B(2) + 1 = 9

I Further details for simulations in Section 5

I.1 Data generating mechanism

The generating mechanism of our pretraining data is as described in Section 4.2, with $C_0 = 2$, $(a_k^{(\alpha)}), (b_{\ell,\gamma}^{(\alpha)}) \overset{\text{iid}}{\sim} \mathrm{Unif}[-1, 1]$ in (41) and P_X being the uniform distribution on $\mathbb{B}^d$.

At test time, we randomly generate new single-index models with fixed smoothness levels under two settings:

(i) No distribution shift: the random test functions are generated from the same distribution as the pretraining data of that particular smoothness,

(ii) With distribution shift: similarly as (i), except that the random variables $(a_k^{(\alpha)})$ in (41) are sampled from $\mathrm{Unif}[0,1]$ instead of $\mathrm{Unif}[-1,1]$.

See Figure 3 for random Besov functions with and without distribution shift (blue: no distribution shift; orange: with distribution shift).

I.2 Training details

For training, we used AdamW (Loshchilov and Hutter, 2019), with weight decay 0.001 and learning rate 0.0003 (with a cosine annealing schedule). We also employed pre-layer normalisation (Xiong et al., 2020) and gradient clipping to stabilise the optimisation. Our Transformer was trained for 200,000 optimisation steps. At each step, we sampled 36 independent single-index regression functions, and for each function, we randomly generated 16 prompts and queries (i.e. the covariate vectors and queries are random while keeping the regression function fixed), yielding a total of 576 pretraining sequences per optimisation step. The training was done on 12 NVIDIA A100 GPUs (80GB) over approximately 17 hours. To facilitate training, the difficulty of the tasks was progressively increased via a curriculum learning strategy. In the initial warm-up phase (the first 10,000 training steps), the ambient dimension d was gradually increased from 1 to 5, while keeping $p = 1$ and $\alpha = 4$ fixed. Subsequently, for the next 50,000 steps, we introduced a smoothness curriculum, where regression functions were drawn from increasingly diverse mixtures whose smoothness parameters were gradually decreased over time. Finally, for the remaining 140,000 steps, training proceeded with a uniform mixture over tasks, where the smoothness parameter was sampled uniformly from $\{2,\ 2.5,\ 3,\ 4\}$.

I.3 Further simulation results

Our theoretical guarantees in Section 4 assume that the support of the test distribution is contained in the support of the pretraining prior. Our simulations in Section 5 also consider a within-support distribution shift. In this subsection, we further examine the empirical effect of different within-support and out-of-support distribution shifts on the performance of our pretrained Transformer. To this end, we consider the following six types of distribution shifts:

(i) No distribution shift: the random test functions are generated from the same distribution as the pretraining data of that particular smoothness,

(ii) Within-support shift $[0,1]$: similarly as (i), except that the random variables $(a_k^{(\alpha)})$ in (41) are sampled from $\mathrm{Unif}[0,1]$ instead of $\mathrm{Unif}[-1,1]$.

(iii) Within-support shift $[0.5,1]$: similarly as (i), except that the random variables $(a_k^{(\alpha)})$ in (41) are sampled from $\mathrm{Unif}[0.5,1]$ instead of $\mathrm{Unif}[-1,1]$.

(iv) Within-support shift $[0.75,1]$: similarly as (i), except that the random variables $(a_k^{(\alpha)})$ in (41) are sampled from $\mathrm{Unif}[0.75,1]$ instead of $\mathrm{Unif}[-1,1]$.

(v) Out-of-support shift $[0,2]$: similarly as (i), except that the random variables $(a_k^{(\alpha)})$ in (41) are sampled from $\mathrm{Unif}[0,2]$.

(vi) Out-of-support shift $[1,2]$: similarly as (i), except that the random variables $(a_k^{(\alpha)})$ in (41) are sampled from $\mathrm{Unif}[1,2]$.

Settings (i) and (ii) above are the same as those in Appendix I.1, while settings (iii) and (iv) impose more severe within-support distribution shifts and settings (v) and (vi) are out-of-support distribution shifts. The MSEs under these settings are given in Figure 5.

To understand the effect of the pretraining data sample size, in Figure 6 we plot the logarithm of the MSE against the logarithm of the number of pretraining sequences, obtained from the checkpoints of our simulation study. We see that there is an approximately linear trend, suggesting a polynomial decay of the MSE with respect to the number of pretraining sequences.

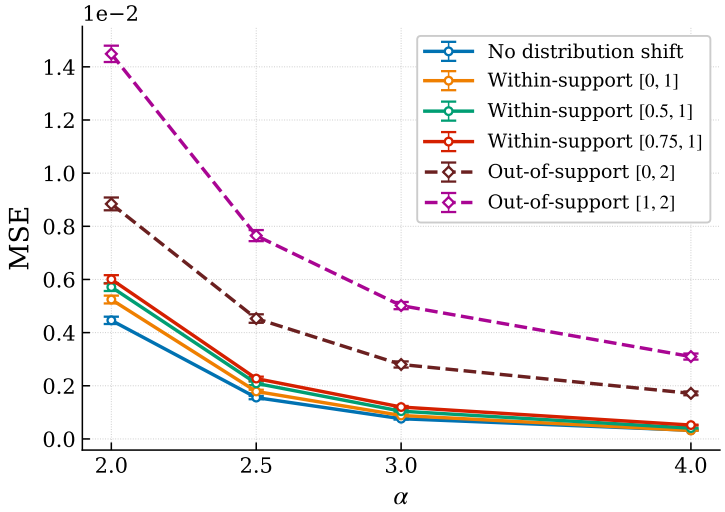


Figure 5: MSE at test time based on 100,000 random test functions. Error bars represent 90% confidence intervals.

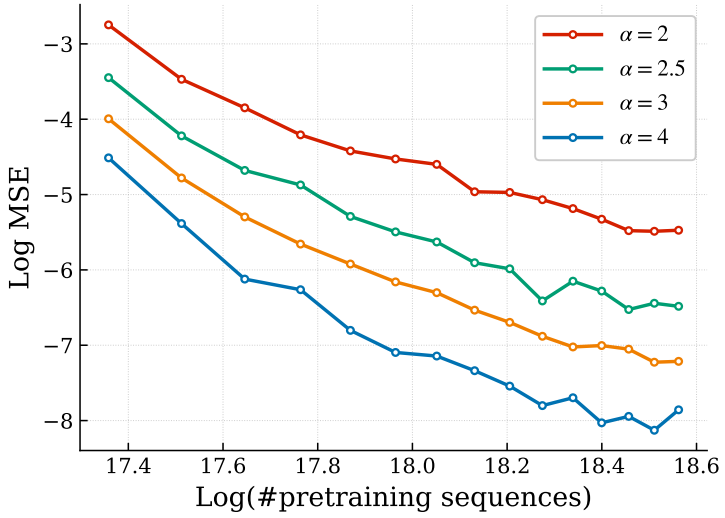


Figure 6: The logarithm of the MSE against the logarithm of the number of pretraining sequences.

NeurIPS Paper Checklist

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

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

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

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

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

IMPORTANT, please:

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

1. **Claims**

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

   Answer: [Yes]

   Justification: We state our main theorems in the main text, and provide rigorous proofs in the appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

Justification: We discuss possible extensions of our work in Section 6.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: We state all assumptions needed clearly and provide complete proofs in the appendix.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide experimental details in Appendix I. The code for our simulations will be made publicly available upon acceptance.

   Guidelines:

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

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

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

   Answer: [Yes]

   Justification: The code for our simulations will be made publicly available upon acceptance.

   Guidelines:

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

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide experimental details in Appendix I. The code for our simulations will be made publicly available upon acceptance.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report error bars for all our experiments in Section 5.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Computer resources used in our experiments are stated in Appendix I.2.

   Guidelines:

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

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our paper conform with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: Our paper is a theoretical work, studying optimal in-context adaptivity and distributional robustness of Transformers, and it does not have any societal impact concerns.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We use open-source packages in our experiments, and we have cited the original paper.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: Our paper is a theoretical work, and we do not release new assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in our paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.