IN-CONTEXT LEARNING IS PROVABLY BAYESIAN INFERENCE: A GENERALIZATION THEORY FOR META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper develops a finite-sample statistical theory for in-context learning (ICL), analyzed within a meta-learning framework that accommodates mixtures of diverse task types. We introduce a principled risk decomposition that separates the total ICL risk into two orthogonal components: Bayes Gap and Posterior Variance. The Bayes Gap quantifies how well the trained model approximates the Bayes-optimal in-context predictor. For a uniform-attention Transformer, we derive a non-asymptotic upper bound on this gap, which explicitly clarifies the dependence on the number of pretraining prompts and their context length. The Posterior Variance is a model-independent risk representing the intrinsic task uncertainty. Our key finding is that this term is determined solely by the difficulty of the true underlying task, while the uncertainty arising from the task mixture vanishes exponentially fast with only a few in-context examples. Together, these results provide a unified view of ICL: the Transformer selects the optimal meta-algorithm during pretraining and rapidly converges to the optimal algorithm for the true task at test time.

1 INTRODUCTION

Large language models (LLMs) have moved far beyond classic NLP benchmarks into complex, real-world workflows (Naveed et al., 2024; Zhao et al., 2025) such as code assistants and generators (GitHub, 2025; Team et al., 2025) in software engineering, Med-PaLM 2 (Singhal et al., 2025) in healthcare, text-to-SQL systems (Gao et al., 2024; Shi et al., 2024) in business intelligence, and vision-language-action models (Kim et al., 2024b; Zitkovich et al., 2023) in robotics. In particular, since GPT-3, modern LLMs have demonstrated a striking ability to adapt to new tasks from only a handful of input-output exemplars, without parameter updates (Brown et al., 2020). This phenomenon, known as in-context learning (ICL), appears across diverse datasets and task formats and is at the heart of these workflows (Min et al., 2022; Dong et al., 2024). These deployments share common constraints: inference-time (test-time) prompts are short, and upstream pretraining covers heterogeneous task types. A concrete, finite-sample account of predictive error under these constraints is therefore of key importance to practitioners.

Numerous studies aim to elucidate the behavior of ICL. Wang et al. (2023); Akyürek et al. (2023); von Oswald et al. (2023); Li et al. (2023); Bai et al. (2023); Garg et al. (2022); Mahankali et al. (2024) have empirically or theoretically shown that Transformers can implement canonical estimators and learning procedures in context (e.g., least squares, ridge, and Lasso, gradient-descent steps, model selection), sometimes achieving near-Bayes-optimal performance on linear tasks. Concurrently, Jeon et al. (2024) provide information-theoretic analysis and Kim et al. (2024a) present nonparametric rates for particular architectures and settings, with subsequent progress (Wang et al., 2024; Oko et al., 2024; Nishikawa et al., 2025). A compelling perspective frames ICL as a form of implicit Bayesian inference (Xie et al., 2022; Wang et al., 2023; Panwar et al., 2024; Arora et al., 2025; Reuter et al., 2025; Zhang et al., 2025). Although this viewpoint provides an explanatory framework for ICL's capabilities, the aforementioned theories have not fully leveraged the theoretical relationship between ICL and Bayes. Hence, they lack a statistical theory that can (i) jointly couple pretraining size N and prompt length p and (ii) accommodate heterogeneous mixtures of task types, the regime in which modern LLMs operate.

We develop a Bayes-centric framework that offers a concrete account of the sources of error and clarifies how they shrink with p and N. Specifically, viewing ICL risk as the Bayes risk (e.g., §5.3.1.2 of Murphy, 2022), we treat the Bayes-optimal predictor as the optimal in-context predictor and derive the following orthogonal decomposition under squared loss (Theorem 1):

$$\text{ICL risk} = \text{Bayes Gap} + \text{Posterior Variance},$$

where the *Bayes Gap* measures the discrepancy between a pretrained model and the optimal in-context (Bayes) predictor, and the *Posterior Variance* is independent of the model and shrinks as the observed context length grows. Conceptually, performance limits at inference time are governed by Bayesian uncertainty about the test task (i.e., the task at inference time), not by pretraining alone. We summarize the further contributions below:

1. **Provide non-asymptotic upper bounds that couple the number of pretraining prompts N and their context length p (Theorem 2).** For uniform-attention Transformers, we leverage sequential learning theory (Rakhlin et al., 2010), develop optimal transport-based approximation theory, and then obtain

$$\mathbb{E}R_{\mathrm{BG}}(M_{\hat{\theta}}) \;\lesssim\; \underbrace{m^{-2\alpha/d_{\mathrm{eff}}}}_{\text{approximation}} + \underbrace{m(pN)^{-1} + N^{-1}}_{\text{pretraining generalization}} \quad \text{(ignoring logarithmic factors)}$$

   Here m is the number of learned features in the Transformer, d_{eff} is the effective dimension, and α is a Hölder exponent. The rate $\propto m/(pN)$ clarifies the dependence on both p and N, which earlier theories on ICL (Kim et al., 2024a; Wu et al., 2024; Zhang et al., 2024) have not fully captured. Importantly, the result suggests that **Transformers select the optimal meta-algorithm during pretraining**.

2. **Explain in-context error via the test-task difficulty (Theorem 3).** In a mixture of task types, the posterior over the task index concentrates exponentially fast with respect to the observed context length, and the irreducible term R_{PV} is upper bounded by the minimax risk of the test (true) task family. Without assuming specific algorithms (Akyürek et al., 2023; Bai et al., 2023; Zhang et al., 2024), our result implies that even in mixed-task settings **the optimal meta-algorithm rapidly converges to the optimal algorithm for the true task at inference time**. This finding is consistent with empirical reports (Panwar et al., 2024; Arora et al., 2025), which show that ICL often behaves like Bayesian inference, particularly in task-mixture settings.

3. **Characterize stability under input-distribution shift (Theorem 4).** We demonstrate that under input-distribution shift from pretraining data to inference-time prompt, the Bayes Gap incurs an out-of-distribution (OOD) penalty proportional to the Wasserstein distance between the distributions, while the Posterior Variance is intrinsic to the target domain. Zhang et al. (2024) have noted that ICL is vulnerable to input-distribution shift in some settings, whereas our results specifically show that only the Bayes Gap increases in proportion to the magnitude of the shift.

The paper is organized as follows. Section 2 formalizes the meta-learning prompt model, introduces the Transformer architecture, and states assumptions, followed by a primer on the Bayes-optimal in-context predictor. Section 3 presents the risk decomposition and then analyzes (i) the Bayes Gap (Section 3.1), (ii) the Posterior Variance (Section 3.2), and (iii) OOD stability under input-distribution shift (Section 3.3). Section 4 concludes with limitations and future work. The Appendix contains a list of notation, numerical experiments, all technical proofs, auxiliary lemmas, and extended discussions.

Related Work

(A) ICL as Bayesian inference. ICL has been framed as (implicit) Bayesian inference under structured pretraining. Xie et al. (2022) show that mixtures of hidden Markov model-style documents enable Transformers to perform posterior prediction; Panwar et al. (2024) show that Transformers mimic Bayes across task mixtures. Lin & Lee (2024) reconcile task retrieval versus task learning with a probabilistic pretraining model. Wang et al. (2023) view LLMs as latent-variable predictors enabling principled exemplar selection. Reuter et al. (2025) empirically show full Bayesian posterior inference in-context and Arora et al. (2025) demonstrate Bayesian scaling laws predicting many-shot reemergence of suppressed behaviors. Our results explicitly use Bayesian properties for

ICL theory and provide a concrete non-asymptotic validation both in pretraining and at inference time.

(B) ICL as Meta-Learning. ICL is widely understood as meta-learning (Brown et al., 2020). Transformers implement gradient-descent-style updates within their forward pass, acting as meta-optimizers that perform implicit fine-tuning (von Oswald et al., 2023; Dai et al., 2023). Models can be meta-trained to execute general-purpose in-context algorithms across tasks (Kirsch et al., 2022). From a learning-to-learn perspective, ICL's expressivity explains few-shot strength while exposing generalization limits (Wu et al., 2025). Beyond single tasks, meta-in-context learning shows recursive adaptation of ICL strategies without parameter updates (Coda-Forno et al., 2023). From this perspective, we theoretically clarify how ICL identifies the task at inference time and solves the true task.

2 Problem Setup

2.1 Meta-Learning: Mixture of Multiple Regression Types

We consider a meta-learning framework that accommodates a finite number of distinct task types (task families).

Definition 1 (Prompt-Generating Process)**.** The data generating process for prompts proceeds as follows:

1. Sample a task type: $I \sim \mathcal{P}_I = \mathrm{Categorical}(\boldsymbol{\alpha})$, i.e., $\Pr(I = i) = \alpha_i > 0$ for $i = 1, \dots, T$.
2. Given $I = i$, sample a task function: $f \sim \mathcal{P}_{F_i}$ where $\mathcal{P}_{F_i}$ is a distribution on the i-th function space $F_i = \{f : \mathbb{R}^{d_{\text{feat}}} \to \mathbb{R}\}$.
3. For $k = 1, \dots, p+1$:
   - Sample an $\mathbb{R}^{d_{\text{feat}}}$-dimensional input: $\boldsymbol{x}_k \overset{\text{i.i.d.}}{\sim} \mathcal{P}_X$
   - Generate output: $y_k = f(\boldsymbol{x}_k) + \varepsilon_k$, where $\varepsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{P}_\varepsilon$ is sub-Gaussian random noise with $\mathbb{E}[\varepsilon_k] = 0$, $\mathrm{Var}(\varepsilon_k) = \sigma_\varepsilon^2$, and $\varepsilon_k \perp (f, \boldsymbol{x}_k)$.
4. Form the length-p (complete) prompt: $P = (\underbrace{\boldsymbol{x}_1, y_1, \dots, \boldsymbol{x}_p, y_p}_{\text{context } D^p}, \underbrace{\boldsymbol{x}_{p+1}}_{\text{query}})$.

This setting allows for a mixture of $T\,(< \infty)$ different task types (task families), such as linear regression type $F = \{\boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} + b\}$, sparse regression type $F = \{\boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} + b : \|\boldsymbol{w}\|_0 \le s\}$, and basis-function regression type $F = \{\boldsymbol{x} \mapsto \sum_{j=0}^{R} a_j g_j(\boldsymbol{x})\}$, where g_j are, for example, Hermite polynomials or Fourier basis functions. Note that Step 1 of Definition 1 selects the task family F_i (via I), and Step 2 samples a particular function f from that family; in the linear-regression case, this corresponds to choosing coefficients such as $\boldsymbol{w}$ and b.

A length-k partial prompt is denoted by $P^k = (\boldsymbol{x}_1, y_1, \dots, \boldsymbol{x}_k, y_k, \boldsymbol{x}_{k+1})$ and its context dataset by $D^k = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{k} \in \mathbb{R}^{k d_{\text{eff}}}$, where $d_{\text{eff}} := d_{\text{feat}} + 1$. We fix a maximum context length p. At inference time, after observing $k \le p$ examples, we sequentially evaluate the risk of predicting y_{k+1} from P^k.

2.2 Transformer Architecture

We begin by briefly reviewing the standard Transformer architecture. The standard Transformer (Vaswani et al., 2017) processes sequences through self-attention mechanisms: $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\big(\frac{QK^\top}{\sqrt{d_k}}\big)V$, where queries Q, keys K, and values V are linear projections of the input embeddings. Each Transformer layer consists of self-attention and a position-wise feed-forward network.

In this work, we adopt a specialized uniform-attention ($Q = K = 0$) Transformer architecture. The components of our prompts are generated independently conditional on the task function (Definition 1). Therefore, a permutation-invariant mechanism like uniform attention is sufficient, which motivates our choice of the following architecture. Further justification is provided in Appendix C.

Definition 2 (Uniform-attention Transformer Architecture)**.** We study a uniform-attention (mean-pooling) Transformer of the form:

$$M_\theta(P^k) := \rho_\theta\Big(\frac{1}{k}\sum_{i=1}^{k}\phi_\theta(\boldsymbol{x}_i, y_i), \boldsymbol{x}_{k+1}\Big).$$

Here, the feature encoder $\phi_\theta : \mathcal{U} \to \Delta^{m-1}$ and the decoder $\rho_\theta : \Delta^{m-1} \times \mathcal{C} \to \mathbb{R}$, where Δ^{m-1} denotes the $(m-1)$-dimensional probability simplex, $\mathcal{U}$ denotes the example domain (the space of $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d_{\text{eff}}}$) and $\mathcal{C}$ denotes the query domain (the space of $\boldsymbol{x}_{k+1}$), have the following structures:

Feature Encoder Network ϕ_θ: The feature encoder consists of a depth-D_ϕ feedforward ReLU network followed by a renormalization layer:

$$\begin{aligned}\phi_\theta(\boldsymbol{x}, y) &:= \text{Renorm}_\tau \circ g_\theta(\boldsymbol{x}, y),\\ g_\theta(\boldsymbol{u}) &:= W^{(D_\phi)}\sigma\big(W^{(D_\phi-1)}\sigma\big(\cdots\sigma\big(W^{(1)}\boldsymbol{u} + \boldsymbol{b}^{(1)}\big)\cdots\big) + \boldsymbol{b}^{(D_\phi-1)}\big) + \boldsymbol{b}^{(D_\phi)},\end{aligned}$$

where $\boldsymbol{u} = [\boldsymbol{x}^\top, y]^\top \in \mathbb{R}^{d_{\text{eff}}}$, $\sigma(\cdot) = \max\{0, \cdot\}$ is the ReLU activation applied element-wise, $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ are weight matrices with $n_0 = d_{\text{eff}}$ and $n_{D_\phi} = m$, and the renormalization layer is defined as $\text{Renorm}_\tau(\boldsymbol{s}) = \frac{\sigma(\boldsymbol{s}) + \frac{\tau}{m}\mathbf{1}}{\mathbf{1}^\top\sigma(\boldsymbol{s}) + \tau}$ $(\tau \in (0, 1])$. This ensures $\phi_\theta(\boldsymbol{x}, y) \in \Delta^{m-1}$.

Decoder Network ρ_θ: The decoder is a depth-D_ρ feedforward ReLU network that jointly processes the aggregated features and query:

$$\begin{aligned}\rho_\theta(\boldsymbol{z}, \boldsymbol{c}) &:= \text{clip}_{[-B_M, B_M]}\big(h_\theta(\boldsymbol{z}, \boldsymbol{c})\big),\\ h_\theta(\boldsymbol{v}) &:= W^{(D_\rho)}\sigma\big(W^{(D_\rho-1)}\sigma\big(\cdots\sigma\big(W^{(1)}\boldsymbol{v} + \boldsymbol{b}^{(1)}\big)\cdots\big) + \boldsymbol{b}^{(D_\rho-1)}\big) + b^{(D_\rho)},\end{aligned}$$

where $\boldsymbol{v} = [\boldsymbol{z}^\top, \boldsymbol{c}^\top]^\top \in \mathbb{R}^{m+d_{\text{feat}}}$, $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ with $n_0 = m + d_{\text{feat}}$ and $n_{D_\rho} = 1$, and the clipping operation ensures $|M_\theta(P^k)| \le B_M$.

Size of the Networks: Throughout, $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices. For depth-D ReLU network $\mathcal{T}_\theta$, define the spectral product $S(\mathcal{T}_\theta) := \prod_{d=1}^{D}\big\|W^{(d)}\big\|_2$. There exist fixed constants $C_\phi,\ C_\rho > 0$ (independent of p, N) such that $S(\phi_\theta) \le C_\phi m^{1/d_{\text{eff}}}$ and $S(\rho_\theta) \le C_\rho m^{1/2}$. For the feature encoder ϕ_θ, we assume a depth of $D_\phi = O(\log m)$ and the number of trainable parameters is $O(m \log m)$. Also, we assume the decoder ρ_θ is uniformly Lipschitz in both arguments: $\big|\rho_\theta(\boldsymbol{z}, \boldsymbol{c}) - \rho_\theta(\boldsymbol{z}', \boldsymbol{c}')\big| \le L_s\|\boldsymbol{z} - \boldsymbol{z}'\|_2 + L_c\|\boldsymbol{c} - \boldsymbol{c}'\|_2$, with $L_s, L_c \le S(\rho_\theta)$. Finally, let Θ denote the parameter space that satisfies these conditions.

Averaging simplex-valued features produces a summary statistic that is permutation-invariant and has a fixed total mass of 1 irrespective of k. Thus, the summary carries only distributional information about the context, rather than scale information due to sequence length.

2.3 Risk, Training, and Assumptions

Throughout, we use the squared loss $\ell(u, v) = (u - v)^2$. The *ICL risk* of a predictor M averages the mean-squared error over $k = 1, \ldots, p$ and the aforementioned generative process:

$$R(M) = \frac{1}{p}\sum_{k=1}^{p}\mathbb{E}_{I\sim\mathcal{P}_I, f\sim\mathcal{P}_{F_I}, D^k\sim\mathcal{P}_{X,Y|f}^{\otimes k}, \boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[(f(\boldsymbol{x}_{k+1}) - M(P^k))^2\right],$$

where $\mathcal{P}_{X,Y|f}$ is the joint distribution of (X, Y) conditional on the task function f. Pretraining is performed with N i.i.d. length-p prompts; the empirical risk minimizer (ERM) is

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\ \frac{1}{pN}\sum_{j=1}^{N}\sum_{k=1}^{p}\big(y_{j,k+1} - M_\theta(P_j^k)\big)^2. \tag{1}$$

Remark 1 (Meta-train/test protocol)**.** The pretraining dataset consists of N i.i.d. prompts $\{P_j\}_{j=1}^N$, each generated by first sampling I_j, next drawing f_j, and then sampling context examples and a query from the same $\mathcal{P}_X$. At inference time, I^{test} and f^{test} are drawn from the same mixture, and the risk $R(M)$ is averaged over new prompts from the same meta-distribution.

For the subsequent discussion and analysis, we make the following assumptions about the task function and the input data.

Assumption 1 (Bounded task functions)**.** There exists $B_f > 0$ such that for any i and $f \in F_i$, $|f(\boldsymbol{x})| \leq B_f$ for all $\boldsymbol{x}$ in the support of $\mathcal{P}_X$.

Assumption 2 (Bounded inputs and conditional independence)**.** There exists $B_X < \infty$ such that $\|\boldsymbol{x}\|_2 \leq B_X$, $\mathcal{P}_X$-almost surely. $\{\boldsymbol{x}_k\}_{k=1}^{p}$ are i.i.d. samples from $\mathcal{P}_X$ and, conditional on a sampled task function f, the pairs $\{(\boldsymbol{x}_k, y_k)\}_k$ are conditionally independent across k.

2.4 Primer on the Bayes-Optimal In-Context Predictor

In this section, we characterize the optimal predictor that minimizes the ICL risk. Since the ICL risk is equivalent to the Bayes risk (e.g., §5.3.1.2 of Murphy, 2022), the theoretically optimal in-context predictor is the Bayes predictor, i.e., the posterior mean of the function value given the context in this setting. We explain this point below.

The ICL risk minimization problem is to find a predictor M that solves:

$$\min_M R(M) = \min_M \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{I\sim\mathcal{P}_I}\mathbb{E}_{f\sim\mathcal{P}_{F_I}}\mathbb{E}_{D^k\sim\mathcal{P}_{X,Y|f}^{\otimes k}}\mathbb{E}_{\boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[\ell\left(f(\boldsymbol{x}_{k+1}), M(P^k)\right)\right].$$

Using the law of total expectation, we can rewrite the risk as an expectation over the context D^k. For each context, we aim to minimize the conditional expectation of the loss:

$$\min_M \mathbb{E}_{D^k\sim\mathcal{P}_{X,Y}^{\otimes k}}\mathbb{E}_{I\sim\mathcal{P}_{I|D^k}}\mathbb{E}_{f\sim\mathcal{P}_{F_I|D^k}}\mathbb{E}_{\boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[\ell\left(f(\boldsymbol{x}_{k+1}), M(P^k)\right)\right].$$

To minimize the outer expectation, it suffices to minimize the inner conditional expectation for each fixed context D^k. The minimizer is exactly the definition of the *Bayes estimator* (Bernardo & Smith, 1994; Robert, 2007) because the inner conditional expectation is the Bayes risk, which is the expected predictive loss $\mathbb{E}_{\boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[\ell\left(f(\boldsymbol{x}_{k+1}), M(P^k)\right)\right]$ with respect to the *Bayes posterior distribution* $\mathbb{E}_{I\sim\mathcal{P}_{I|D^k}}[\mathcal{P}_{F_I|D^k}]$. Specifically, for the squared error loss, the value $M(P^k)$ that minimizes the conditional mean squared error, $\mathbb{E}_{I\sim\mathcal{P}_{I|D^k}}\mathbb{E}_{f\sim\mathcal{P}_{F_I|D^k}}\mathbb{E}_{\boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[\ell\left(f(\boldsymbol{x}_{k+1}), M(P^k)\right)\right]$, is the Bayes posterior mean (e.g., Murphy, 2022; Lehmann & Casella, 1998). Thus, the optimal predictor M_{Bayes} that minimizes the ICL risk is the posterior mean:

$$M_{\text{Bayes}}(P^k) := \mathbb{E}_{I\sim\mathcal{P}_{I|D^k}}\mathbb{E}_{f\sim\mathcal{P}_{F_I|D^k}}[f(\boldsymbol{x}_{k+1})] \equiv \arg\min_M R(M).$$

This Bayes predictor serves as the theoretical target during pretraining. (See Figure 1.) Intuitively, the optimal ICL can be viewed as performing implicit prompt learning (Li & Liang, 2021; Lester et al., 2021): given a context D^k, it infers a task-specific representation ($\mathbb{E}_{I\sim\mathcal{P}_{I|D^k}}[\mathcal{P}_{F_I|D^k}]$) which plays a similar role as a learned prompt in prompt learning. The **Bayes Gap**, which we introduce next, measures how well the pretrained model $M_{\hat{\theta}}$ emulates this predictor.

Posterior notation.
Let $\pi_i(D^k) := \Pr(I = i \mid D^k)$ and $\mathcal{P}(f \mid D^k) = \sum_{i=1}^{T} \pi_i(D^k)\mathcal{P}_{F_i}(f \mid D^k, I = i)$. We write the Bayes predictor as $M_{\text{Bayes}}(P^k) = \mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$[1]. Note that, throughout, we work on standard Borel spaces so that regular conditional distributions exist. Accordingly, $\Pr(f \in \cdot \mid D^k)$ and the quantities $\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid D^k]$ and $\mathrm{Var}(f(\boldsymbol{x}_{k+1}) \mid D^k)$ are well-defined.

Permutation invariance of the Bayes predictor.
For each k, we write $\boldsymbol{u}_k = (\boldsymbol{x}_k, y_k) \in \mathcal{U}$ and $\boldsymbol{c} = \boldsymbol{x}_{k+1} \in \mathcal{C}$ and view the Bayes predictor $M_{\text{Bayes}}(P^k)$ as $M_{\text{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c})$ here. Since the posterior $\mathcal{P}(f \mid D^k)$ depends on D^k only through the multiset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k}$, for any permutation π of $\{1, \ldots, k\}$, $M_{\text{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) = M_{\text{Bayes}}\left(\boldsymbol{u}_{\pi(1)}, \ldots, \boldsymbol{u}_{\pi(k)}, \boldsymbol{c}\right)$. Thus, the Bayes predictor is a symmetric set functional, which justifies using the uniform-attention Transformer to emulate it. See Appendix C for more details.

[1] As the query $\boldsymbol{x}_{k+1}$ is drawn independently of f and D^k, $\mathcal{P}(f \mid P^k) = \mathcal{P}(f \mid D^k)$.

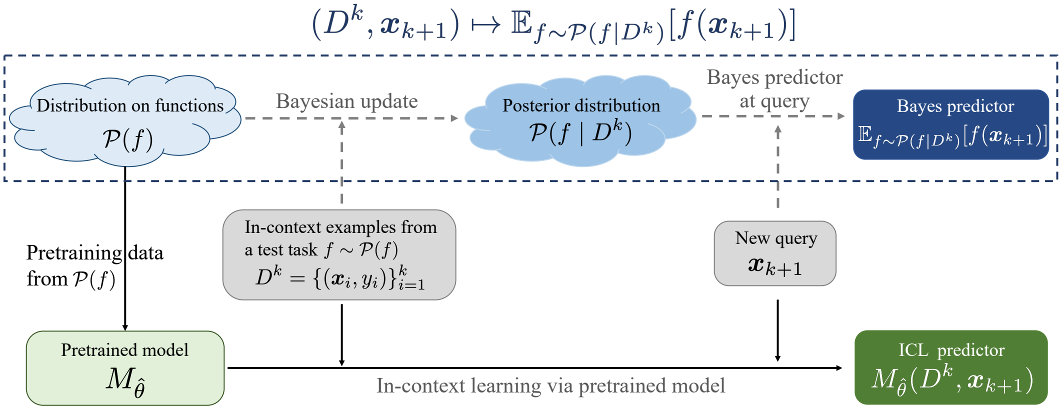
$(D^k, \boldsymbol{x}_{k+1}) \mapsto \mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$
Distribution on functions
$\mathcal{P}(f)$
Bayesian update
Posterior distribution
$\mathcal{P}(f \mid D^k)$
Bayes predictor at query
Bayes predictor
$\mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$
In-context examples from a test task $f \sim \mathcal{P}(f)$
$D^k = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^k$
New query
$\boldsymbol{x}_{k+1}$
Pretraining data from $\mathcal{P}(f)$
Pretrained model
$M_{\hat{\theta}}$
In-context learning via pretrained model
ICL predictor
$M_{\hat{\theta}}(D^k, \boldsymbol{x}_{k+1})$

Figure 1: Bayesian view of in-context learning (ICL). The upper path: the process of computing the optimal prediction is $(D^k, \boldsymbol{x}_{k+1}) \mapsto \mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$ given $\mathcal{P}(f)$. The lower path: since $\mathcal{P}(f)$ is unknown, the model $M_{\hat{\theta}}$, pretrained on data from $\mathcal{P}(f)$, aims to emulate this process via $(D^k, \boldsymbol{x}_{k+1}) \mapsto M_{\hat{\theta}}(D^k, \boldsymbol{x}_{k+1})$.

3 Risk Analysis of In-Context Learning

In this section, we first present a risk identity (Theorem 1), then control each term separately: Section 3.1 bounds the Bayes Gap (pretraining approximation and generalization), while Section 3.2 analyzes the Posterior Variance (inference-time uncertainty in mixtures).

The following identity, using the Bayes predictor, decomposes the ICL risk into a model-dependent term and a model-independent term.

Theorem 1 (Risk decomposition for in-context learning)**.** *Consider the prompt-generating process from Definition 1 and assume that Assumption 1 holds. For a measurable, bounded map M, the ICL risk decomposes as*

$$R(M) = \underbrace{R_{\mathrm{BG}}(M)}_{\textit{Bayes Gap}} + \underbrace{R_{\mathrm{PV}}}_{\textit{Posterior Variance}}$$

where:

1. ***Bayes Gap**: $R_{\mathrm{BG}}(M) := \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{P^k}\left[\left(M(P^k) - M_{\mathrm{Bayes}}(P^k)\right)^2\right]$. This measures how closely the model M approximates the optimal Bayes predictor. In other words, this is the excess risk to the Bayes predictor.*

2. ***Posterior Variance**: $R_{\mathrm{PV}} := \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{P^k}\left[\mathrm{Var}_{f\sim\mathcal{P}(f|D^k)}(f(\boldsymbol{x}_{k+1}))\right]$, which is independent of M and irreducible. This represents the behavior of the Bayes estimator given the context.*

This decomposition reveals how each term can be reduced. The Bayes Gap is the aggregate imperfections of pretraining, controlled through architecture design and pretraining scale (N, p). In contrast, the Posterior Variance stems from the inference-time uncertainty of the test task and can be reduced only by increasing the context length k at inference time because $\mathbb{E}[\mathrm{Var}_{f\sim\mathcal{P}(f|D^{k+1})}(f)] \leq \mathbb{E}[\mathrm{Var}_{f\sim\mathcal{P}(f|D^k)}(f)]$ follows from the law of total variance. Therefore, under sufficiently large pretraining, the final error bottleneck is the latter. Also, relative to information-theoretic decompositions under log-loss (Jeon et al., 2024), our identity is exact under squared loss and directly interprets the irreducible term as the Posterior Variance. Note that this decomposition holds for any measurable and bounded predictor, independent of the specific form of the model. Also, the analysis can be extended to a broader class of losses that admit an analogous decomposition with the Bayes estimator, such as Bregman-type losses (Adlam et al., 2022; Pfau, 2025).

3.1 Bayes Gap: Pretraining Generalization Error and Approximation Error

This section answers "*Can M_θ emulate the hypothetical map $P^k \mapsto \mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$?*"

For the uniform-attention Transformers, the following theorem decomposes the Bayes Gap into an approximation term and a pretraining generalization term, and provides a non-asymptotic upper bound that depends jointly on both p and N.

Theorem 2 (Bayes Gap upper bound)**.** *Consider the prompt-generating process defined in Definition 1 under Assumptions 1 and 2. For $k = 1, \dots, p$, assume the Bayes predictor $M_{\text{Bayes}} : (\mathbb{R}^{d_{\text{eff}}})^k \times \mathbb{R}^{d_{\text{feat}}} \to \mathbb{R}$ satisfies the Hölder condition: $\left|M_{\text{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) - M_{\text{Bayes}}(\boldsymbol{u}'_{1:k}, \boldsymbol{c}')\right| \leq L \frac{1}{k} \sum_{i=1}^{k} \left\|(\boldsymbol{u}_i, \boldsymbol{c}) - (\boldsymbol{u}'_i, \boldsymbol{c}')\right\|_2^{\alpha}$ for bounded $\boldsymbol{u}_i, \boldsymbol{u}'_i \in \mathcal{U}$, $\boldsymbol{c}, \boldsymbol{c}' \in \mathcal{C}$, and $\alpha \in (0, 1]$. Let $\hat{\theta}$ be the ERM* (1) *with $\mathcal{D}_{\text{train}} = \{\{(P_j^k, y_{j,k+1})\}_{k=1}^{p}\}_{j=1}^{N}$. Then, for any $p \geq 2$,*

$$\mathbb{E}R_{\text{BG}}(M_{\hat{\theta}}) \lesssim \underbrace{m^{-\frac{2\alpha}{d_{\text{eff}}}}}_{\textit{Approximation error}} + \underbrace{\frac{m}{pN}\operatorname{polylog}(pN) + \frac{1}{N}\operatorname{polylog}(pN)}_{\textit{Pretraining generalization error}},$$

where the expectation is taken over $\mathcal{D}_{\text{train}}$ and $\operatorname{polylog}(\cdot) \asymp \log^r(\cdot)$ *with some $r \in \mathbb{N}$. Choosing $m^\star \asymp (pN)^{\frac{d_{\text{eff}}}{d_{\text{eff}}+2\alpha}}$ and ignoring* $\operatorname{polylog}(pN)$ *yield* $\mathbb{E}R_{\text{BG}}(M_{\hat{\theta}}) \lesssim \left((pN)^{-\frac{2\alpha}{d_{\text{eff}}+2\alpha}} + N^{-1}\right)$.

Proof Idea: Regarding the pretraining generalization error, we handle the N meta-training prompts via conventional learning theory across j (van der Vaart & Wellner, 2023; Shalev-Shwartz & Ben-David, 2014), and the p context examples per prompt via a sequential learning theory across k (Rakhlin et al., 2015; Block et al., 2021). Concerning the approximation error, we build a mollified partition-of-unity ("soft histogram") over the example domain $\mathcal{U}$ and mean-pool it to encode prompts. Then the Bayes predictor on empirical measures is approximated by a decoder defined via a McShane extension over a discrete 1-Wasserstein metric between histograms (Peyré & Cuturi, 2019), yielding a Lipschitz, piecewise-linear target. Both encoder and decoder are then realized by moderate-size ReLU networks. As these proof ideas do not depend on a specific Bayesian formulation, the result holds under milder data assumptions compared to prior Bayesian analyses (Xie et al., 2022; Zhang et al., 2025).

The key point of Theorem 2 is that R_{BG} decomposes into (i) an approximation error $m^{-2\alpha/d_{\text{eff}}}$ stemming from the expressiveness of the Transformer, and (ii) a generalization error $\tilde{O}\big(m/(pN) + 1/N\big)$ coming from a finite dataset. This decomposition clarifies the respective roles of the two terms. The feature dimension m governs the expressive power of the Transformer, and increasing m allows a smoother approximation of the Bayesian predictor. (The depth and number of parameters of the Transformer affect Theorem 2 only implicitly, through the feature dimension m and the Lipschitz constants of the encoder and decoder (Lemmas 4-5).) On the other hand, p represents the amount of information within one task, while N represents the coverage of the meta-distribution. The rate $\propto m/(pN)$ makes explicit the joint effect of pN, which earlier non-asymptotic theories on ICL (Kim et al., 2024a; Wu et al., 2024; Zhang et al., 2024) have not fully captured, as they typically considered the effect of p and N separately or focused on only one of them. Many works (e.g., Akyürek et al., 2023; Bai et al., 2023; Zhang et al., 2024) have theoretically and empirically shown that Transformers approximate ridge regression and gradient descent in linear settings. By contrast, we non-asymptotically demonstrate that in more general settings (nonparametric, nonlinear, meta-learning), the optimal meta-algorithm is selected.

We also highlight its ability to avoid *the curse of dimensionality* with respect to context length p. Since the Bayes predictor is unchanged no matter the order in which the context arrives, we can compress a long input sequence into a single mean vector without losing information, and the network only needs to handle that fixed-length vector of dimension d_{eff} rather than $pd_{\text{eff}} + d_{\text{feat}}$.

The Hölder condition holds, intuitively, if (i) each task function is smooth (e.g., Hölder) with respect to the input, (ii) inputs and responses are effectively bounded (e.g., sub-Gaussian noise), and (iii) Bayesian updates are stable (e.g., distributions of parameters are light-tailed or log-concave), so perturbing any single context point by $O(\delta)$ changes the posterior mean by at most $O(\delta^\alpha/k)$ under the prompt metric. These conditions are typically met for mixtures of common task families (e.g., linear regression, basis-function regression, finite convex-dictionary regression). Further discussion is deferred to Appendix E. Moreover, the rate $(pN)^{-\frac{2\alpha}{d_{\text{eff}}+2\alpha}}$ matches the minimax lower bound for estimating, for example, the density of the joint distribution of $(\boldsymbol{x}_i, y_i) \in \mathcal{U}$ under the standard Hölder smoothness assumption (Tsybakov, 2009).

In practice, as the token budget used for pretraining LLMs is enormous (say, infinite), the only risk that essentially remains is the inference-time risk (R_{PV}) analyzed in the following section.

3.2 Posterior Variance: Inference-time Error

Having established bounds on the Bayes Gap, we now turn to the other component of the ICL risk: the Posterior Variance, R_{PV}. This term represents the irreducible error of the Bayes predictor itself. A key question is: *How does this Posterior Variance, arising from a mixture of T task types, relate to the intrinsic difficulty of the true task at inference time?*

The following theorem shows that, under some assumptions on the data (discussed later), the Bayes predictor quickly identifies the true task type at inference time.

Theorem 3 (Gap between Posterior Variance and minimax risk of the true task type)**.** *Suppose Assumption 1 holds. Let $i^\star$ be the true task index. For each wrong task $j \neq i^\star$ and each $t \geq 1$, define the predictive log-likelihood ratio increment $Z_{j,t} := \log \frac{p_j(y_t|\boldsymbol{x}_t, D^{t-1})}{p_{i^\star}(y_t|\boldsymbol{x}_t, D^{t-1})}$. Under the true task, there exist a task type divergence $D_j > 0$ and constants (ν_j, b_j) such that, for all $t \geq 1$ and the filtration $\mathcal{G}_{t-1}$, $\mathbb{E}\big[Z_{j,t} \mid \mathcal{G}_{t-1}, I = i^\star\big] \leq -D_j$ and $\mathbb{E}\left[\exp\{\lambda(Z_{j,t} + D_j)\} \mid \mathcal{G}_{t-1}, I = i^\star\right] \leq \exp\left(\lambda^2\nu_j^2/2\right)$ hold for all $|\lambda| \leq 1/b_j$. Let $D_{\min} := \min_{j \neq i^\star} D_j > 0$ and $C := \min_{j \neq i^\star} \frac{D_j^2}{8(\nu_j^2 + b_j D_j/2)} > 0$. Then, for all $k \geq 1$,*

$$\mathbb{E}_{D^k,\boldsymbol{x}|I=i^\star}\Big[\underbrace{\mathrm{Var}_{f|D^k}\{f(\boldsymbol{x})\}}_{\textit{mixture Posterior Variance}}\Big] \leq \underbrace{\inf_{M} \sup_{f \in F_{i^\star}} \mathbb{E}_{P^k}\left[\left(f(\boldsymbol{x}_{k+1}) - M(P^k)\right)^2 \Big| f\right]}_{\textit{the true task type's minimax risk}}$$
$$+ \underbrace{5B_f^2\left(\frac{1-\alpha_{i^\star}}{\alpha_{i^\star}} e^{-D_{\min}k/2} + (T-1)e^{-Ck}\right)}_{\textit{task type identification error}}.$$

This theorem quantitatively justifies the empirical observation that ICL can quickly adapt to the specific task at hand, even when pretrained on a diverse mixture. Concretely, the posterior distribution over the task index, $\mathcal{P}_{I|D^k}$, concentrates exponentially fast on the true index $i^\star$ as k grows. This result is consistent with empirical demonstrations. Panwar et al. (2024) show that in hierarchical mixtures, Transformers mimic the Bayes predictor based on the true task distribution. Also, the above theorem explains the "Bayesian scaling laws" of Arora et al. (2025), which model ICL's error curves as repeated Bayesian updates, and under an ideal Bayesian learner, the task posterior converges to the true task as context grows.

Compared to prior ICL theories, Theorem 3 can be seen as the general form of the result in Kim et al. (2024a), which showed that even when the function class used at pretraining is wider than the one at inference, the inference error depends only on the hardness of the latter class. Although Jeon et al. (2024) also mentions an irreducible error of ICL, our addition is to show that it manifests as Posterior Variance and that it approaches the minimax risk for the "true family" up to a small gap. Moreover, this phenomenon of ICL "selecting algorithms on the fly" is consistent with the theoretical results on in-context algorithm selection in generalized linear models and the Lasso (Akyürek et al., 2023; Bai et al., 2023; Zhang et al., 2024). Our result proves that even without assuming a specific algorithmic form, behavior close to optimal algorithm selection emerges through posterior concentration in mixture settings.

The assumptions are fairly standard in the theory of sequential data and ensure that the in-context examples provide sufficient signal to rapidly rule out incorrect task types: (i) the supermartingale condition $\mathbb{E}[Z_{j,t} \mid \mathcal{G}_{t-1}, I = i^\star] \leq -D_j < 0$ (Williams, 1991) means each new observation, on average, decreases the predictive log-likelihood ratio of any wrong type j against the true type; (ii) the Bernstein-type condition $\mathbb{E}\left[\exp\{\lambda(Z_{j,t} + D_j)\} \mid \mathcal{G}_{t-1}, I = i^\star\right] \leq \exp\left(\lambda^2\nu_j^2/2\right)$ (Bercu et al., 2015) yields the concentration of the cumulative log-likelihood ratio, so occasional misleading samples cannot outweigh the overall trend. Note that D_j is the per-step information gap that favors the true task over the wrong type j, ν_j is the sub-exponential scale of the log-likelihood ratio, b_j bounds the tail via the moment-generating-function (smaller means heavier tails), and $\min_{j \neq i^\star} D_j^2/8(\nu_j^2 + b_j D_j/2)$ sets the uniform exponential rate at which posterior mass on wrong types decays with more context.

In Appendix F, we consider a concrete regression problem (linear vs. series regression) and specify ν_j, b_j, D_j and C that appear in Theorem 3. The results show that to make the task type identification error at most η, one requires the context length: $k \asymp \frac{\text{error variance+within-task variance}}{\text{true-task signal}} \log \frac{\text{\# of task types}}{\eta}$.

Remark that if the likelihood does not have a density function (with respect to Lebesgue measure), assume that all predictive distributions $P_i(y_t \mid \boldsymbol{x}_t, D^{t-1})$ are dominated by a common reference measure so that the Radon-Nikodym derivative exists. Then $Z_{j,t}$ can be rigorously defined as a log-likelihood ratio.

3.3 OOD Stability of the ICL risk

This section investigates how the ICL risk changes under a distributional shift in the input between pretraining data and inference-time prompt. Note that the task distribution and the noise distribution are unchanged. Since R_{PV} represents the uncertainty of the task at inference time, it depends only on the prompt distribution at inference time. In contrast, the Bayes Gap $R_{\mathrm{BG}}(M_{\hat{\theta}})$, which measures the performance of the pretrained model $M_{\hat{\theta}}$, is directly affected by the discrepancy between the pretraining (source domain) and inference-time (target domain) distributions.

To formalize the problem, let P denote the prompt distribution based on the source input distribution $\mathcal{P}_X$ used during pretraining, and Q be the prompt distribution based on a target input distribution $\mathcal{Q}_X$ at inference time. Denote the Bayes Gap evaluated under a distribution R by

$$R_{\mathrm{BG}}^{(\mathsf{R})}(M_\theta) := \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{P^k \sim \mathsf{R}}\big[\big\{M_\theta(P^k) - M_{\mathrm{Bayes}}(P^k)\big\}^2\big].$$

We measure the shift at the prompt level. For $0 < \alpha \le 1$ and $k \in \{1, \ldots, p\}$, define the ground metric $\overline{d}_{k,\alpha}\big((\boldsymbol{u}_{1:k}, \boldsymbol{c}), (\boldsymbol{u}'_{1:k}, \boldsymbol{c}')\big) := \frac{1}{k}\sum_{i=1}^{k} \|\boldsymbol{u}_i - \boldsymbol{u}'_i\|_2^\alpha + \|\boldsymbol{c} - \boldsymbol{c}'\|_2^\alpha$, and the associated 1-Wasserstein distance $\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_{\mathsf{P}}(P^k), \mathcal{L}_{\mathsf{Q}}(P^k)\big) := W_1\big(\mathcal{L}_{\mathsf{P}}(P^k), \mathcal{L}_{\mathsf{Q}}(P^k); \overline{d}_{k,\alpha}\big)$. Assume $\mathcal{U}$ and $\mathcal{C}$ have finite diameters (e.g., by truncating on a high-probability event under the sub-Gaussian noise model) for brevity, and recall that the decoder is uniformly Lipschitz in its two arguments with constants (L_s, L_c), while $\mathrm{Lip}(\phi_\theta)$ denotes the encoder's Lipschitz constant.

Theorem 4 (Wasserstein stability of the Bayes Gap)**.** *Consider the prompt-generating process defined in Definition 1 under Assumptions 1 and 2. Suppose that the Bayes predictor satisfies the same α-Hölder condition as in Theorem 2 with exponent $\alpha \in (0, 1]$ and constant L. Then, for every parameter θ,*

$$\big|R_{\mathrm{BG}}^{(\mathsf{Q})}(M_\theta) - R_{\mathrm{BG}}^{(\mathsf{P})}(M_\theta)\big| \le \frac{2(B_M + B_f)}{p} \sum_{k=1}^{p} \big(L + \Lambda_\alpha\big)\, \mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_{\mathsf{P}}(P^k), \mathcal{L}_{\mathsf{Q}}(P^k)\big),$$

where $\Lambda_\alpha := \big(L_s\,\mathrm{Lip}(\phi_\theta) + L_c\big)\,\big(\mathrm{diam}(\mathcal{U}) + \mathrm{diam}(\mathcal{C})\big)^{1-\alpha}$.

This result implies that the Bayes Gap is distributionally Lipschitz: its change across domains is controlled by (i) the smoothness L of the Bayes predictor and (ii) the architectural regularity of M_θ through L_s, L_c, and $\mathrm{Lip}(\phi_\theta)$. The penalty scales with the prompt-level Wasserstein shift and does not depend on the number of pretraining prompts N. In line with the findings of Zhang et al. (2024) that ICL is susceptible to input-distribution shifts, we find that only the Bayes Gap is affected by such shifts and quantify the magnitude of this effect.

For additional theories and detailed discussions, please see Appendix D.

4 Conclusion

In this work, we introduced a Bayesian-centric framework to dissect the ICL phenomenon. Our central contribution is an orthogonal decomposition of the ICL risk into two conceptually distinct components: a model-dependent *Bayes Gap* and a model-independent *Posterior Variance*. This decomposition provides a principled lens through which to understand the sources of error in ICL and how they are reduced by pretraining and in-context examples.

Our analysis of the Bayes Gap (Theorem 2) yielded non-asymptotic bounds that jointly couple the number of pretraining prompts N and their context length p. This result clarifies their synergistic role in learning an optimal meta-algorithm, showing that the model's ability to emulate the ideal Bayes predictor improves as the total number of pretraining examples (pN) grows. The analysis of the Posterior Variance (Theorem 3) revealed that in a heterogeneous mixture of tasks, ICL rapidly identifies the true underlying task family at inference time. The irreducible error converges exponentially fast to the minimax risk of the true task, explaining ICL's adaptability without explicit algorithm selection. Finally, we characterized the model's stability under distribution shift (Theorem 4), demonstrating that the Bayes Gap increases moderately in proportion to the Wasserstein distance between the pretraining and inference input distributions, while the Posterior Variance remains intrinsic to the target domain.

Practical implications for model design and training. For practitioners and model designers, our theory provides concrete guidance for optimizing in-context learning performance:

- **Architecture**: Choosing the model's feature dimension m (hidden size) according to the scaling $m^\star \asymp (pN)^{\frac{d_{\text{eff}}}{d_{\text{eff}}+2\alpha}}$ is beneficial (Theorem 2). Moreover, when within-task examples are i.i.d., using permutation-invariant architectures such as uniform-attention (mean-pooling) Transformers, which match the symmetry of the Bayes-optimal predictor and avoid the curse of dimensionality in context length, is sufficient (cf. Appendix C).
- **Inference-time prompts**: At inference time, an excessively long prompt may not be necessary, but the model does need at least a few examples to identify the task (about three in our experiments, cf. Figure 3). For more complex problems, the model needs enough examples to solve them, such that the minimax risk becomes sufficiently small (Theorem 3).
- **Distributional shift**: One should match the input distribution at inference to that used in pretraining; our Wasserstein stability bound shows that distribution shifts primarily degrade the Bayes Gap (Theorem 4). When moderate domain shift is unavoidable, it is preferable to control the Lipschitz constants of the Transformer (e.g., via spectral normalization and output clipping) and then increase p and N and adjust m according to the above scaling, rather than arbitrarily growing model size.
- **Benchmark design**: Since ICL performance transitions from being dominated by task-type identification (small k) to within-family generalization (large k), few-shot benchmarks primarily measure task discrimination, while longer-context evaluations assess learning within the identified family.

Limitations and Future Work. Our analysis focuses on a uniform-attention Transformer, motivated by the permutation invariance of the Bayes predictor in our setup. If inputs in a prompt are chosen adaptively (active learning, bandit-style data acquisition), non-uniform attention or explicitly sequential models would be beneficial. Future work could explore how these results extend to more complex architectures with non-uniform attention under dependent data settings. Our generalization analysis already uses tools from sequential learning, so the technical machinery can be compatible with order-dependent settings; what would change are the target Bayes predictor and its symmetry.

References

Ben Adlam, Neha Gupta, Zelda Mariet, and Jamie Smith. Understanding the bias-variance tradeoff of Bregman divergences. *arXiv preprint arXiv:2202.04167*, 2022. doi: 10.48550/arXiv.2202.04167.

J. Aitchison and S. M. Shen. Logistic-Normal Distributions: Some Properties and Uses. *Biometrika*, 67(2):261–272, 1980. doi: 10.2307/2335470. URL https://doi.org/10.2307/2335470.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah Goodman. Bayesian scaling laws for in-context learning. 2025. URL `https://openreview.net/forum?id=I4YU0oECtK`.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL `http://jmlr.org/papers/v20/17-612.html`.

Bernard Bercu, Bernard Delyon, and Emmanuel Rio. *Concentration Inequalities for Sums and Martingales*. SpringerBriefs in Mathematics. Springer, Cham, 2015. doi: 10.1007/978-3-319-22099-4.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., May 1994. ISBN 9780471494645. doi: 10.1002/9780470316870.

Adam Block, Yuval Dagan, and Alexander Rakhlin. Majorizing measures, sequential complexities, and online learning. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 587–590. PMLR, 15–19 Aug 2021. URL `https://proceedings.mlr.press/v134/block21a.html`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X Wang, and Eric Schulz. Meta-in-context learning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=sx0xpaO0za`.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247. URL `https://aclanthology.org/2023.findings-acl.247/`.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A Survey on In-context Learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL `https://aclanthology.org/2024.emnlp-main.64/`.

Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. ISBN 978-0-8218-4974-3. doi: 10.1090/gsm/019. URL `https://doi.org/10.1090/gsm/019`.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1–22, 2015. doi: 10.1214/EJP.v20-3496. URL https://doi.org/10.1214/EJP.v20-3496.

Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. *Proceedings of the VLDB Endowment*, 17(5):1132–1145, 2024. doi: 10.14778/3641204.3641221. URL https://doi.org/10.14778/3641204.3641221.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, 2013. ISBN 9781439840955. doi: 10.1201/b16018. URL https://sites.stat.columbia.edu/gelman/book/.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017.

GitHub. GitHub Copilot. https://github.com/features/copilot, 2025. Accessed: 2025-09-18.

Hong Jun Jeon, Jason D. Lee, Qi Lei, and Benjamin Van Roy. An Information-Theoretic Analysis of In-Context Learning. In *Forty-first International Conference on Machine Learning*, 2024.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are Minimax Optimal Nonparametric In-Context Learners. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024a.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024b.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-Purpose In-Context Learning by Meta-Learning Transformers. *arXiv preprint arXiv:2212.04458*, 2022.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, New York, NY, 2 edition, 1998. doi: 10.1007/b98854.

Brian Lester, Rami Al-Rfou, and Noah Constant. ”the power of scale for parameter-efficient prompt tuning”. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation, 2021.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as Algorithms: Generalization and Stability in In-context Learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23l.html.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30135–30188. PMLR, 21–27 Jul 2024.

Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One Step of Gradient Descent is Provably the Optimal In-Context Learner with One Layer of Linear Self-Attention. In *The Twelfth International Conference on Learning Representations*, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL http://probml.github.io/book1.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A Comprehensive Overview of Large Language Models, 2024. URL https://arxiv.org/abs/2307.06435.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=xQTSvP57C3.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained Transformer Efficiently Learns Low-Dimensional Target Functions In-Context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=uHcG5Y6fdB.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-Context Learning through the Bayesian Prism. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HX5ujdsSon.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073.

Víctor H. Peña and Evarist Giné. *Decoupling*. Probability and Its Applications. Springer New York, NY, 1999. doi: 10.1007/978-1-4612-0537-1.

David Pfau. A Generalized Bias-Variance Decomposition for Bregman Divergences. *arXiv preprint arXiv:2511.08789*, 2025. URL https://arxiv.org/abs/2511.08789.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models Are Unsupervised Multitask Learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online Learning: Random Averages, Combinatorial Parameters, and Learnability. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/e00406144c1e7e35240afed70f34166a-Paper.pdf.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015. doi: 10.1007/s00440-013-0545-5.

Arik Reuter, Tim G. J. Rudner, Vincent Fortuin, and David Rügamer. Can Transformers Learn Full Bayesian Inference in Context? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9Ip6fihKbc.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, 2nd edition, 2007. doi: 10.1007/0-387-71599-1.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Liang Shi, Zhengju Tang, Nan Zhang, Xiaotong Zhang, and Zhi Yang. A Survey on Employing Large Language Models for Text-to-SQL Tasks. *arXiv preprint*, abs/2407.15186, 2024. URL https://arxiv.org/abs/2407.15186. Submitted July 2024.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, mar 2025. doi: 10.1038/s41591-024-03423-7. URL https://www.nature.com/articles/s41591-024-03423-7.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A Family of Highly Capable Multimodal Models, 2025. URL https://arxiv.org/abs/2312.11805.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, NY, 2009. doi: 10.1007/b13794.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, second edition, 2023. ISBN 978-3-031-29038-1. doi: 10.1007/978-3-031-29040-4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BGvkwZEGt7.

Zhijie Wang, Bo Jiang, and Shuai Li. In-context Learning on Function Classes Unveiled for Transformers. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=rJkGOARXns.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How Many Pretraining Tasks Are Needed for In-Context Learning of Linear Regression? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vSh5ePa0ph.

Shiguang Wu, Yaqing Wang, and Quanming Yao. Why In-Context Learning Models are Good Few-Shot Learners? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iLUcsecZJp.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002. URL https://www.sciencedirect.com/science/article/pii/S0893608017301545.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and How does In-Context Learning Learn? Bayesian Model Averaging, Parameterization, and Generalization. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1684–1692. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/zhang25d.html.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, 2025. URL https://arxiv.org/abs/2303.18223.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

APPENDIX

A NOTATION AND DEFINITIONS

This section provides a comprehensive list of notations and definitions used throughout the paper for ease of reference.

General Mathematical Notation

- $\mathbb{R}^d$: The d-dimensional Euclidean space.
- $\|\cdot\|_2$: The Euclidean (ℓ_2) norm for vectors and the spectral (operator) norm for matrices.
- $\|\cdot\|_1$: The ℓ_1 norm of a vector.
- $\|\cdot\|_0$: The ℓ_0 pseudo-norm of a vector, counting the number of non-zero elements.
- $\mathbf{1}$: A vector of all ones, with its dimension inferred from the context.
- Δ^{m-1}: The standard probability simplex in $\mathbb{R}^m$, defined as $\Delta^{m-1} = \{\boldsymbol{s} \in [0,1]^m : \sum_{j=1}^m s_j = 1\}$.
- $\mathcal{U}, \mathcal{C}$: The example domain (the space of $(\boldsymbol{x}_i, y_i)$) and the query domain (the space of $\boldsymbol{x}_{k+1}$), respectively.
- F_i: The function space for tasks of type i.
- Θ: The parameter space for the neural network model M_θ.
- $\mathrm{diam}(A) := \sup_{\boldsymbol{x},\boldsymbol{y}\in A} \|\boldsymbol{x}-\boldsymbol{y}\|_2$: The diameter of a set A.
- $B(\boldsymbol{a}, R)$: The closed Euclidean ball of radius $R \geq 0$ centered at $\boldsymbol{a}$, $B(\boldsymbol{a}, R) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}-\boldsymbol{a}\|_2 \leq R\}$, where the ambient dimension d is understood from context.
- $\mathrm{Lip}(f)$: The Lipschitz constant of a function f.
- $f \asymp g$: Indicates that f and g are of the same order, i.e., there exist constants $c_1, c_2 > 0$ such that $c_1 g \leq f \leq c_2 g$.
- $f \lesssim g$: Indicates that f is less than or equal to g up to a constant factor, i.e., $f \leq Cg$ for some universal constant $C > 0$.
- $\tilde{O}(\cdot)$: Asymptotic notation that hides polylogarithmic factors.
- $\mathrm{polylog}(\cdot) := (\log(\cdot))^{O(1)}$, i.e., $\log^c(\cdot)$ for some constant $c > 0$.
- $\sigma(\cdot)$: The Rectified Linear Unit (ReLU) activation function, $\sigma(u) = \max\{u, 0\}$, applied element-wise.
- $\mathrm{clip}_{[a,b]}(x) := \max(a, \min(b, x))$: The clipping function.

Probability and Statistics

- $\mathcal{P}_X, \mathcal{P}_\varepsilon, \ldots$: Probability distributions of random variables $X, \varepsilon, \ldots$.
- $\mathbb{E}_{X\sim\mathcal{P}_X}[\cdot]$ or simply $\mathbb{E}[\cdot]$: The expectation with respect to the distribution of the random variable(s) specified in the subscript. If no subscript is present, the expectation is taken over all relevant random variables.
- $\mathrm{Var}(\cdot)$: The variance of a random variable.
- $\mathrm{Emp}_k(\boldsymbol{u}_{1:k}) := \frac{1}{k}\sum_{t=1}^k \delta_{\boldsymbol{u}_t}$: The empirical measure of the context.
- $\Sigma_X := \mathbb{E}\big[(\boldsymbol{x}-\mathbb{E}\boldsymbol{x})(\boldsymbol{x}-\mathbb{E}\boldsymbol{x})^\top\big]$: The covariance matrix of $\boldsymbol{x}$.
- $\Pr(\cdot)$: The probability of an event.
- $X \sim \mathcal{P}_X$: The random variable X is drawn from the distribution $\mathcal{P}_X$.
- $\overset{\text{i.i.d.}}{\sim}$: A symbol for "is independently and identically distributed as".
- $X \perp Y$: The random variables X and Y are statistically independent.
- $\mathcal{P}_{X,Y|f}$: The joint distribution of (X, Y) conditional on a function f.

- $\mathcal{P}^{\otimes k}$: The k-fold product measure, corresponding to k i.i.d. draws from the distribution $\mathcal{P}$.
- $I \sim \mathrm{Categorical}(\boldsymbol{\alpha})$: I is a discrete random variable on $\{1, \ldots, T\}$ with $\Pr(I = i) = \alpha_i$. $\sum_{i=1}^{T} \alpha_i = 1$.
- $\mathcal{P}(f \mid D^k)$: The marginal posterior distribution of the task function f given the context data D^k.
- $\pi_i(D^k) := \Pr(I = i \mid D^k)$: The marginal posterior probability of task type (task family) i given the context D^k.
- $\mathcal{G}_k$: The σ-algebra generated by the random variables D^k, representing the information available at step k.
- $\mathcal{G}'_k$: The σ-algebra generated by the random variables $(D^k, \boldsymbol{x}_{k+1})$.
- Sub-Gaussian: A centered random variable X is sub-Gaussian with proxy variance σ^2 if $\mathbb{E}e^{\lambda X} \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$.
- Sub-exponential: A centered random variable X is (ν, b)-sub-exponential if $\mathbb{E}e^{\lambda X} \leq \exp(\nu^2\lambda^2/2)$ for all $|\lambda| \leq 1/b$.
- $\mathrm{KL}(P\|Q)$: Kullback–Leibler divergence between distributions P and Q, used to quantify separation between task types.
- $m_i(D^k) := \int \prod_{t=1}^{k} p(y_t \mid \boldsymbol{x}_t, f)\, \mathcal{P}_{F_i}(\mathrm{d}f)$: The marginal likelihood of the context D^k under task type i.
- $\mu_{i,t}(\boldsymbol{x}),\ s^2_{i,t}(\boldsymbol{x})$: Predictive mean and variance for task type i after observing $t-1$ examples.
- $\mathcal{N}(\mu, \sigma^2)$: Gaussian (normal) distribution with mean μ and variance σ^2.
- Truncated Gaussian: A Gaussian distribution restricted to a bounded support set and renormalized to integrate to 1.
- $\frac{\mathrm{d}P}{\mathrm{d}Q}$: Radon–Nikodym derivative of P with respect to Q (when P is absolutely continuous with respect to Q).

Meta-learning Setup

- T: The total number of distinct task types (task families).
- p: The maximum number of in-context examples (i.e., the context length).
- $i^\star$: The index of the true task type at inference time.
- N: The number of prompts in the pretraining dataset.
- d_{feat}: The dimensionality of the input features $\boldsymbol{x}$.
- $d_{\mathrm{eff}} := d_{\mathrm{feat}} + 1$: The effective dimensionality of an example pair $(\boldsymbol{x}_i, y_i)$.
- m: The dimensionality of the feature vector produced by the encoder, $\phi_\theta(\boldsymbol{x}_i, y_i)$.
- $P = (\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_p, y_p, \boldsymbol{x}_{p+1})$: A full prompt of length p.
- $P^k = (\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \boldsymbol{x}_{k+1})$: A partial prompt of length k.
- $D^k = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{k}$: The context data, consisting of k example pairs.
- $M, M_\theta, M_{\hat{\theta}}$: A generic predictor, the uniform-attention Transformer parameterized by θ, and the uniform-attention Transformer obtained by empirical risk minimization (ERM), respectively.
- ϕ_θ: The feature encoder network that maps an example $(\boldsymbol{x}, y)$ to a feature vector in Δ^{m-1}.
- ρ_θ: The decoder network that predicts the output from the aggregated features and the query input.
- $\ell(u, v) = (u - v)^2$: The squared error loss function used throughout the paper.
- $S(\mathcal{T}_\theta) = \prod_{\ell=1}^{L} \|W^{(\ell)}\|_2$: The spectral product of the weight matrices of a neural network $\mathcal{T}_\theta$.
- $\mathrm{Renorm}_\tau(\boldsymbol{s})$: A specific renormalization layer that maps a vector $\boldsymbol{s} \in \mathbb{R}^m$ to the probability simplex Δ^{m-1}.
- B_f, B_X, B_M: Uniform bounds on $|f(\boldsymbol{x})|$, $\|\boldsymbol{x}\|_2$, and $|M(P^k)|$, respectively (as assumed).
- $S(\phi_\theta), S(\rho_\theta)$: Layerwise spectral-product bounds (or induced Lipschitz budgets) for the feature network ϕ_θ and decoder ρ_θ used in generalization and stability analyses.

Theoretical Quantities

- $R(M)$: The in-context learning (ICL) risk of a predictor M. $R(M) := \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}[(M(P^k) - f(\boldsymbol{x}_{k+1}))^2]$
- $M_{\text{Bayes}}(P^k) := \mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid D^k, \boldsymbol{x}_{k+1}]$: The Bayes predictor, which corresponds to the posterior mean of the query output given the context and is the optimal predictor for the squared error loss.
- $R_{\text{BG}}(M)$: The Bayes Gap, measuring the squared difference between the predictor M and the Bayes predictor, averaged over prompts. This term is reducible by training the model.
- $R_{\text{BG},k}(M) := \mathbb{E}[\{M(P^k) - M_{\text{Bayes}}(P^k)\}^2]$, $R_{\text{PV},k} := \mathbb{E}[\text{Var}(f(\boldsymbol{x}_{k+1}) \mid D^k)]$: Per-$k$ versions used in the risk decomposition.
- $R_{\text{BG}}^{(\mathsf{P})}(M) := \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{P^k \sim \mathsf{P}}[\{M(P^k) - M_{\text{Bayes}}(P^k)\}^2]$: Bayes Gap evaluated under a prompt distribution P (used in OOD analysis).
- R_{PV}: The Posterior Variance, which is the irreducible error corresponding to the variance of the posterior predictive distribution. This term is independent of the model M.
- $R_k^{\star}(F_{i^\star})$: The minimax risk for predicting a function from the true task class $F_{i^\star}$ given k examples.
- $R_k^{\star}(F_{i^\star}; \mathsf{R})$: The minimax risk for predicting a function from the true task class $F_{i^\star}$ under prompt distribution R (default R is the pretraining domain).
- L, α: Constants that define the Hölder condition on the Bayes predictor (see Lemma 5 and Theorem 2).
- $Z_{j,t} := \log \frac{p_j(y_t|\boldsymbol{x}_t, D^{t-1})}{p_{i^\star}(y_t|\boldsymbol{x}_t, D^{t-1})}$: The predictive log-likelihood ratio increment.
- $D_j,\ \nu_j,\ b_j,\ C$: Identification-rate constants for the wrong task $j \neq i^\star$; D_j is the negative drift, (ν_j, b_j) are sub-exponential parameters, and C controls exponential concentration of the posterior mass on incorrect types.
- S_k: The symmetric group on $\{1, \ldots, k\}$; $\mathcal{S}[M]$ denotes the symmetrized predictor obtained by averaging M over all permutations in S_k.
- Predictable tree: A depth-p tree $z = \{z_t(\xi_{1:t-1})\}_{t \le p}$ whose node z_t depends only on past signs $\xi_{1:t-1} \in \{\pm 1\}^{t-1}$.
- ℓ_2 sequential metric: For a depth-p tree z and predictable sequences v, v', and for a path $\xi \in \{\pm 1\}^p$, define

$$d_{2,\xi}(v, v'; z) := \left[\frac{1}{p}\sum_{t=1}^{p}\{v_t(z_t(\xi_{1:t-1})) - v_t'(z_t(\xi_{1:t-1}))\}^2\right]^{1/2}.$$

- $N_2^{\text{seq}}(\alpha, \mathcal{F}; z)$: The sequential covering number (Rakhlin et al., 2010; 2015) is the minimal size of a predictable α-cover on a predictable tree z with respect to $d_{2,\xi}(\cdot, \cdot; z)$ such that, for all $\xi \in \{\pm 1\}^p$ and all $f \in \mathcal{F}$, there exists v in the cover with $d_{2,\xi}(f \circ z, v; z) \le \alpha$.
- $N_2^{\text{seq}}(\alpha, \mathcal{F}, p)$: The depth-$p$ ℓ_2 sequential covering number is the worst–tree version $N_2^{\text{seq}}(\alpha, \mathcal{F}, p) := \sup_z N_2^{\text{seq}}(\alpha, \mathcal{F}; z)$, where the supremum ranges over all predictable trees z of depth p.
- Sequential Rademacher complexity: $\mathfrak{R}_p^{\text{seq}}(\mathcal{F}) := \sup_z \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}} \frac{1}{p}\sum_{t=1}^{p} \xi_t f\big(z_t(\xi_{1:t-1})\big)\right]$, where $\xi_t \overset{\text{i.i.d.}}{\sim} \text{Unif}\{\pm 1\}$.
- $W_1(\mu, \nu; d)$: The 1-Wasserstein distance between probability measures μ, ν with ground metric d. The specialized distances $W_\alpha^{(u)}$ and $\mathsf{W}_\alpha^{(k)}$ below are instances of $W_1(\cdot, \cdot; \cdot)$ with particular choices of d.
- $W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t})$: Discrete 1-Wasserstein on Δ^{m-1} with grid $\{\boldsymbol{r}_j\} \subset \mathcal{U}$ and cost $c^{(u)}(j, \ell) = \|\boldsymbol{r}_j - \boldsymbol{r}_\ell\|_2^\alpha$ $(0 < \alpha \le 1)$; $W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t}) = \min_{\pi \ge 0} \sum_{j,\ell} c^{(u)}(j, \ell)\pi_{j\ell}$ subject to the usual marginal constraints. On the simplex, $W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t}) \le \frac{\text{diam}(\mathcal{U})^\alpha}{2}\|\boldsymbol{s} - \boldsymbol{t}\|_1$.

- $\mathcal{P}_X,\ \mathcal{Q}_X$: Source (pretraining) and target (test) input distributions used in OOD analysis.
- $\mathcal{L}_P(P^k),\ \mathcal{L}_Q(P^k)$: Distributions of length-k prompts under the source and target domains, respectively.
- $\overline{d}_{k,\alpha}\big((\boldsymbol{u}_{1:k},\boldsymbol{c}),(\boldsymbol{u}'_{1:k},\boldsymbol{c}')\big) := \frac{1}{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{u}'_i\|_2^\alpha + \|\boldsymbol{c}-\boldsymbol{c}'\|_2^\alpha$: Prompt-level ground metric $(0<\alpha\le 1)$.
- $\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k),\mathcal{L}_Q(P^k)\big) := W_1\big(\mathcal{L}_P(P^k),\mathcal{L}_Q(P^k);\overline{d}_{k,\alpha}\big)$: Prompt-level Wasserstein distance used in OOD bounds.
- $\mathsf{P},\ \mathsf{Q}$: Generic prompt distributions used when evaluating risks.

B Experiments

In this section, we empirically investigate (i) how the Bayes Gap of a Transformer depends on the number of pretraining prompts N and their length p, and (ii) how the actual ICL prediction by the Transformer depends on the in-context length k.

Prompt Generation We set the feature dimension $d_{\text{feat}} = 5$. Tasks are generated from a mixture of two regression families: Family 1 is linear regression $y = \boldsymbol{w}^\top \boldsymbol{x} + \varepsilon$, and Family 2 is regression via an element-wise nonlinear map $\phi(\boldsymbol{x}) = \tanh(\boldsymbol{x})$, i.e., $y = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + \varepsilon$. The mixture weights are $(\alpha_1,\alpha_2) = (0.5, 0.5)$. The weight vector $\boldsymbol{w}$ for each family is drawn from $N(\boldsymbol{\mu}_i, \tau^2 I)$ with $\tau = 1$, $\boldsymbol{\mu}_1 = (3,3,3,3,3)$ and $\boldsymbol{\mu}_2 = (-3,-3,-3,-3,-3)$. The observation noise ε is zero-mean Gaussian with $\sigma_\varepsilon = 0.1$. This process generates prompts $P^k = (\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \boldsymbol{x}_{k+1}) \in (\mathbb{R}^{d_{\text{feat}}} \times \mathbb{R})^k \times \mathbb{R}^{d_{\text{feat}}}$ and the target $y_{k+1} \in \mathbb{R}$.

Architecture As in Oko et al. (2024) and Garg et al. (2022)[2], we adopt a GPT-2 model (Radford et al., 2019; Wolf et al., 2020), with 12 layers, 8 heads, a hidden size of 256, and positional embeddings removed. The input is a token sequence of length $2k+1$: $P^k = (\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \boldsymbol{x}_{k+1})$.

Evaluation of Bayes Gap For any prompt $(D^k, \boldsymbol{x}_{k+1})$, since each family admits a conjugate Bayesian update, we can compute the mixture posterior $\pi_i(D^k)$ and the Bayes predictor M_{Bayes} in closed form (Gelman et al., 2013). At inference time, using token sequences of length $2p+1$, we compute the difference between the model's prediction and the Bayes prediction $M_{\text{Bayes}}(D^k, \boldsymbol{x}_{k+1})$. We average this over 1000 trials and report the mean as the Bayes Gap.

- **N-Sweep (fixed p):** For $p \in \{5, 10, 15\}$, we train the model on data with $N \in \{128, 256, \ldots, 8192\}$ and report the Bayes Gap.
- **p-Sweep (fixed N):** For $N \in \{500, 1000, 2000\}$, we vary $p \in \{1, 2, 4, 6, 8, 10, 12, 14\}$ and report the Bayes Gap.

During pretraining, we perform the ERM as in (1) using Adam (Kingma & Ba, 2015; Paszke et al., 2019) with a learning rate of 0.00005, a batch size of 64, and 10000 steps.

Figure 2 illustrates how the Bayes Gap depends jointly on the number of pretraining prompts N and the context length p. In the N-sweep (left panel), larger p values start with smaller gaps and maintain this advantage as N grows. In the p-sweep (right panel), increasing p steadily reduces the gap for any fixed N, with higher N curves lying lower overall.

Evaluation of In-Context Error We pretrain the model on a dataset with $N = 12800000$ and $p = 15$ using Adam, which involves 200000 steps of online learning, sampling 64 data points per step, similar to (Garg et al., 2022). On 1000 test prompts, the Bayes Gap is below 10^{-2}, so the pretrained Transformer reasonably approximates M_{Bayes} (cf. Theorem 2). We then investigate: (i) the prediction accuracy of the Transformer via the mean squared error (MSE) with respect to y, and (ii) Bayes-like behavior of the Transformer via linear probing: generating inputs, recovering the regression coefficients implied by the Transformer's outputs (inverting the induced linear map), and comparing against the optimal Bayesian coefficients. The probed coefficients are averaged over 2000 independent input–output sets.

[2] https://github.com/dtsip/in-context-learning, MIT License.

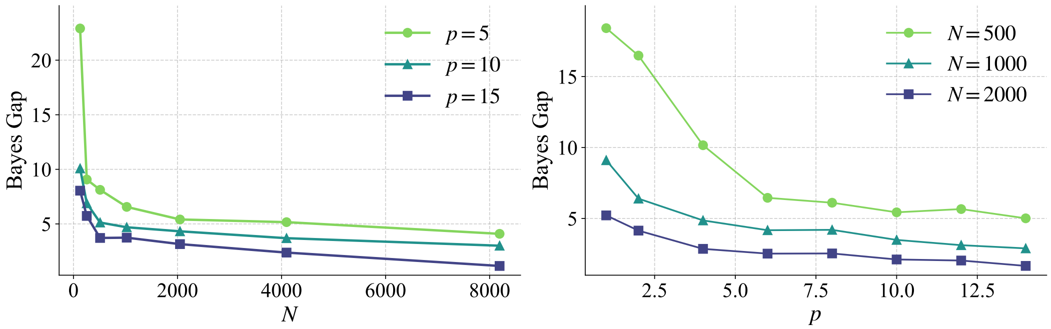

Figure 2: **Behavior of the Bayes Gap (left: N-sweep, right: p-sweep).** The left panel fixes $p \in \{5, 10, 15\}$ and varies the number of pretraining prompts N; the right panel fixes $N \in \{500, 1000, 2000\}$ and varies the context length p in pretraining. In both cases, the Bayes Gap decreases generally as N or p increases, demonstrating that longer contexts and more pretraining improve approximation to the Bayes predictor.

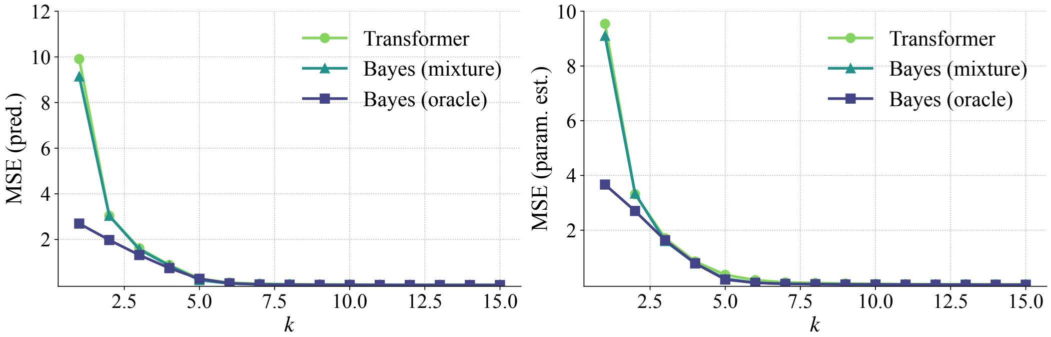

Figure 3: **In-context error under task mixtures (left: predictive MSE; right: parameter-estimation MSE).** As the context length k increases, both predictive error for the next label y_{k+1} from P^k (left) and parameter-estimation error (right) decrease monotonically. The Transformer closely tracks the mixture Bayes predictor and, with sufficient context, approaches the oracle Bayes curve that knows the true task family. This demonstrates the rapid concentration of the task-index posterior under growing context and the corresponding shrinkage of the irreducible term.

Figure 3 illustrates the inference-time behavior of the Transformer. Note that Bayes (mixture) corresponds to the Bayes-optimal in-context predictor M_{Bayes} that conditions on the mixture prior over task families, whereas Bayes (oracle) denotes the Bayes predictor that knows the true task family at inference time. First, with sufficiently large pretraining, the Transformer tracks the Bayes (mixture) curve almost exactly and its error decays toward zero, consistent with our theory (Theorem 2). In terms of parameter estimation, the Transformer likewise reproduces the Bayes-mixture behavior, supporting the view that it performs Bayesian inference in context. Moreover, the gap between the Transformer and Bayes (oracle) diminishes rapidly as the number of in-context examples k increases and essentially vanishes around $k \approx$ 3–4. This indicates that the task type identification error quickly diminishes and the remaining bottleneck is the intrinsic (optimal) error of the true task family (Theorem 3).

C Permutation Invariance and Justification for Uniform-Attention Transformers

This section formalizes the permutation invariance of the Bayes predictor under the prompt-generating process (Definition 1). In summary, by Proposition 1, the Bayes predictor depends on D^k only via its empirical measure. Hence, any architecture that can approximate functionals of empirical distributions, e.g., uniform-attention Transformers, matches the symmetry of the optimal predictor. Moreover, in view of Theorem 5, replacing any non-invariant model by its permutation average never increases risk.

Recall that the loss is the squared error, and all random objects live on standard Borel spaces, so regular conditional distributions exist.

Under Definition 1, once the task (I, f) is fixed, the context pairs $(\boldsymbol{x}_t, y_t)$ are i.i.d. draws. The following lemma says that the context can be treated as a multiset rather than an ordered list.

Lemma 1 (Conditional exchangeability)**.** *Fix (I, f) with $I \in \{1, \dots, T\}$ and $f \in F_I$. Under Definition 1, the context pairs $(\boldsymbol{x}_t, y_t)_{t=1}^k$ are i.i.d. from $\mathcal{P}_{X,Y|f}$; hence for any permutation π of $\{1, \dots, k\}$,*

$$\mathcal{L}((\boldsymbol{x}_1, y_1, \dots, \boldsymbol{x}_k, y_k) \,|\, I, f) = \mathcal{L}\big((\boldsymbol{x}_{\pi(1)}, y_{\pi(1)}, \dots, \boldsymbol{x}_{\pi(k)}, y_{\pi(k)}) \,\big|\, I, f\big)\,,$$

where $\mathcal{L}(Z \mid W)$ denotes the conditional law (distribution) of Z given W.

If the order of the context is uninformative, averaging any predictor over all permutations should not increase risk. This symmetrization principle justifies restricting attention to permutation-invariant models.

Theorem 5 (Risk–reducing symmetrization)**.** *For any measurable predictor M and any $k \in \{1, \dots, p\}$, define the permutation–averaged predictor*

$$\mathcal{S}[M](P^k) := \mathbb{E}_\Pi\big[M\big((\boldsymbol{x}_{\Pi(1)}, y_{\Pi(1)}, \dots, \boldsymbol{x}_{\Pi(k)}, y_{\Pi(k)}),\, \boldsymbol{x}_{k+1}\big) \mid D^k, \boldsymbol{x}_{k+1}\big]\,,$$

where Π is uniform on the symmetric group S_k and independent of everything else. Then the ICL risk satisfies

$$R(\mathcal{S}[M]) \;\le\; R(M).$$

Hence, by convexity of the squared loss and conditional exchangeability (Lemma 1), permutation-averaging is a risk-reducing ensembling step. This is an instance of Rao–Blackwellization by group averaging under convex loss (Lehmann & Casella, 1998). Therefore, uniform-attention (mean-pooling) architectures are not only natural but also without loss of optimality in this setting.

Proof of Theorem 5. Write $R(M) = \frac{1}{p}\sum_{k=1}^{p} R_k(M)$ with $R_k(M) := \mathbb{E}\big[(f(\boldsymbol{x}_{k+1}) - M(P^k))^2\big]$. Fix k and condition on $(D^k, \boldsymbol{x}_{k+1}, I, f)$. By Jensen's inequality applied to the convex map $v \mapsto (f(\boldsymbol{x}_{k+1}) - v)^2$,

$$\mathbb{E}_\Pi\big[(f(\boldsymbol{x}_{k+1}) - M_\Pi)^2 \mid D^k, \boldsymbol{x}_{k+1}, I, f\big] \;\ge\; \big(f(\boldsymbol{x}_{k+1}) - \mathcal{S}[M]\big)^2,$$

where $M_\Pi := M\big((\boldsymbol{x}_{\Pi(1)}, y_{\Pi(1)}), \dots, (\boldsymbol{x}_{\Pi(k)}, y_{\Pi(k)}), \boldsymbol{x}_{k+1}\big)$. By Lemma 1, $\mathbb{E}_\Pi[(f(\boldsymbol{x}_{k+1}) - M_\Pi)^2 \mid I, f] = \mathbb{E}[(f(\boldsymbol{x}_{k+1}) - M(P^k))^2 \mid I, f]$. Taking expectations proves $R_k(\mathcal{S}[M]) \le R_k(M)$ and summing over k yields the claim. □

The previous theorem immediately yields that an optimal predictor can be chosen permutation-invariant:

Corollary 1 (Existence of permutation-invariant minimizers)**.** *There exists a risk minimizer that is permutation-invariant in the k context items. In particular, when analyzing architectures it is without loss of generality to restrict to permutation-invariant (set-valued) models, e.g., uniform-attention/mean-pooling Transformers.*

Analytically, we may restrict our hypothesis class to set-function architectures (e.g., uniform-attention Transformers) without sacrificing optimality.

With a mixture over task families, the optimal predictor must both identify the task type and perform within-family inference. Bayes' rule exposes this computational structure explicitly.

Theorem 6 (Hierarchical posterior factorization)**.** *Assume that for each i all predictive distributions $P_i(y \mid \boldsymbol{x}, f)$ are dominated by a common reference measure so that Radon–Nikodym derivatives $p(y \mid \boldsymbol{x}, f)$ exist. Then for any context D^k,*

$$\mathcal{P}_{F_i}(\mathrm{d}f \mid D^k, I = i) \;\propto\; \Big\{\prod_{t=1}^{k} p(y_t \mid \boldsymbol{x}_t, f)\Big\}\, \mathcal{P}_{F_i}(\mathrm{d}f), \qquad \pi_i(D^k) = \frac{\alpha_i\, m_i(D^k)}{\sum_{j=1}^{T} \alpha_j\, m_j(D^k)},$$

where $m_i(D^k) := \int \prod_{t=1}^{k} p(y_t \mid \boldsymbol{x}_t, f)\, \mathcal{P}_{F_i}(\mathrm{d}f)$ *and* $\pi_i(D^k) := \Pr(I = i \mid D^k)$*. Consequently, the Bayes predictor decomposes as*

$$M_{\text{Bayes}}(P^k) = \sum_{i=1}^{T} \pi_i(D^k)\, \mathbb{E}_{f \sim \mathcal{P}_{F_i}(\cdot \mid D^k, I=i)}[f(\boldsymbol{x}_{k+1})] \,.$$

The Bayes predictor is a mixture of within-family posterior means with weights $\pi_i(D^k)$ determined by marginal likelihoods $m_i(D^k)$. Because these weights depend on the product of likelihood factors, they are invariant to permutations of the context, foreshadowing the permutation invariance results below and validating architectures that first summarize the context before decoding.

Proof of Theorem 6. Bayes' rule and conditional i.i.d. of $(\boldsymbol{x}_t, y_t)$ given (I, f) yield the displayed formulas. The final expression follows from the tower property applied to $\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid D^k, \boldsymbol{x}_{k+1}]$. □

Corollary 2 (Permutation invariance of the Bayes predictor)**.** *For any permutation* π *of* $\{1, \ldots, k\}$*,*

$$M_{\text{Bayes}}((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_k, y_k), \boldsymbol{x}_{k+1}) = M_{\text{Bayes}}\big((\boldsymbol{x}_{\pi(1)}, y_{\pi(1)}), \ldots, (\boldsymbol{x}_{\pi(k)}, y_{\pi(k)}), \boldsymbol{x}_{k+1}\big) \,.$$

Proof of Corollary 2. In Theorem 6, both the within-family posterior $\mathcal{P}_{F_i}(\cdot \mid D^k, I = i)$ and the weight $\pi_i(D^k)$ depend on D^k only through the product $\prod_{t=1}^{k} p(y_t \mid \boldsymbol{x}_t, f)$, which is invariant under reindexing $t \mapsto \pi(t)$. Substituting this into the mixture formula yields the claim. □

Proposition 1 (Empirical–measure representation)**.** *Let* $\mathrm{Emp}_k : \mathcal{U}^k \to \mathcal{P}(\mathcal{U})$ *be the empirical measure map* $\mathrm{Emp}_k(\boldsymbol{u}_{1:k}) = \frac{1}{k} \sum_{t=1}^{k} \delta_{\boldsymbol{u}_t}$*, where* $\mathcal{P}(\mathcal{U})$ *is endowed with the Borel* σ*-algebra of the weak topology. Then there exists a measurable map* $\Psi : \mathcal{P}(\mathcal{U}) \times \mathcal{C} \to \mathbb{R}$ *such that*

$$M_{\text{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) = \Psi(\mathrm{Emp}_k(\boldsymbol{u}_{1:k}),\, \boldsymbol{c}) \qquad (\forall\, \boldsymbol{u}_{1:k} \in \mathcal{U}^k,\ \boldsymbol{c} \in \mathcal{C}).$$

Once the context is summarized as an empirical distribution, a mean-pooled soft histogram is a faithful finite-dimensional proxy. This is precisely the representation approximated by our feature map ϕ_θ and decoder ρ_θ (cf. Lemma 5, Theorem 2). It also explains why the analysis avoids a dependence on the sequence length p beyond averaging.

Proof of Proposition 1. By Corollary 2, $M_{\text{Bayes}}(\cdot, \boldsymbol{c})$ is invariant under permutations on $\mathcal{U}^k$. The quotient of a standard Borel space by a finite group action is standard Borel; thus any measurable, permutation-invariant map factors measurably through the canonical invariant Emp_k. Define $\Psi(\mu, \boldsymbol{c})$ to be the common value of $M_{\text{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c})$ on the fiber $\{\boldsymbol{u}_{1:k} : \mathrm{Emp}_k(\boldsymbol{u}_{1:k}) = \mu\}$; well-definedness follows from invariance, measurability from the quotient factorization. □

Remark 2 (When permutation invariance may fail)**.** The invariance arguments rely on the conditional i.i.d. structure of Definition 1. If inputs are chosen adaptively (active learning, bandit-style data acquisition), or if the within-prompt distribution drifts over time, $(\boldsymbol{x}_t, y_t)$ are not conditionally i.i.d. given (I, f) and order may carry information. In such cases, non-uniform attention or explicitly sequential models can be beneficial, and our uniform-attention analysis should be viewed as the principled baseline for the i.i.d. prompt regime.

For further investigation of exchangeability of the Bayes predictor in the standard Bayesian statistics context, refer to Bernardo & Smith (1994); Gelman et al. (2013); Ghosal & van der Vaart (2017).

D Further Details on Out-of-Distribution Generalization

Throughout this section, $\mathcal{P}_X$ denotes the source input distribution used during pretraining and $\mathcal{Q}_X$ an input distribution at inference time. The task distribution and the noise distribution are unchanged. We work with ground metrics and assume compact supports (finite diameters) so that constants remain finite; in particular, $\mathrm{diam}(\mathcal{U}) < \infty$ and $\mathrm{diam}(\mathcal{C}) < \infty$. (Unit-diameter rescaling, e.g., $\mathrm{diam}(\mathcal{U}) \le 1$ and $\mathrm{diam}(\mathcal{C}) \le 1$, is a convenient normalization and only rescales constants.)

Remark 3 (High-probability boundedness)**.** As in Step 0 of Theorem 2, define $\mathcal{E}_\delta := \{\max_{t\le p+1}|\varepsilon_t| \le t_\delta\}$ with $t_\delta = \sigma_\varepsilon\sqrt{2\log(4p/\delta)}$. On $\mathcal{E}_\delta$, the domains $\mathcal{U}, \mathcal{C}$ have finite diameters since $|y| \le B_f + t_\delta$. Theorems in this section can thus be established on $\mathcal{E}_\delta$, while the contribution of $\mathcal{E}_\delta^{\mathsf{c}}$ is controlled by sub-Gaussian tails, yielding an additional $O(\delta\log(1/\delta))$ term.

For a metric space $(\mathcal{Z}, d)$, we write $W_1(\mu,\nu; d) := \inf_{\pi\in\Pi(\mu,\nu)} \int d(z,z')\pi(\mathrm{d}z,\mathrm{d}z')$ for the 1-Wasserstein distance with ground metric d. For $0 < \alpha \le 1$ and $k \in \mathbb{N}_+$, define the prompt-level ground metric

$$\overline{d}_{k,\alpha}\big((\boldsymbol{u}_{1:k},\boldsymbol{c}),(\boldsymbol{u}'_{1:k},\boldsymbol{c}')\big) := \frac{1}{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{u}'_i\|_2^\alpha + \|\boldsymbol{c}-\boldsymbol{c}'\|_2^\alpha,$$

and abbreviate $\mathsf{W}_\alpha^{(k)}(\cdot,\cdot) := W_1(\cdot,\cdot;\overline{d}_{k,\alpha})$. Likewise, for a random pair $\boldsymbol{U} = (\boldsymbol{X}, Y)$, we write $\mathsf{W}_\alpha(\cdot,\cdot) := W_1(\cdot,\cdot;\|\cdot\|_2^\alpha)$. Note that for any metric d and $0 < \alpha \le 1$, d^α is a metric by concavity of $t \mapsto t^\alpha$.

Define

$$R_{\mathrm{BG}}^{(\mathsf{P})}(M_\theta) := \frac{1}{p}\sum_{k=1}^{p}\mathbb{E}_{P^k\sim\mathsf{P}}\left[\big(M_\theta(P^k) - M_{\mathrm{Bayes}}(P^k)\big)^2\right],$$

so that $R_{\mathrm{BG}}^{(\mathcal{P}_X)}$ (resp. $R_{\mathrm{BG}}^{(\mathcal{Q}_X)}$) means the expectation under $\mathcal{L}_P(P^k)$ (resp. $\mathcal{L}_Q(P^k)$). Since $|f| \le B_f$, we have $|M_{\mathrm{Bayes}}| \le B_f$ and hence $|M_\theta - M_{\mathrm{Bayes}}| \le B_M + B_f$.

Theorem 7 (Wasserstein stability: OOD upper bound for the Bayes Gap)**.** *Under Definition 1, Definition 2 and Assumptions 1–2, assume the Bayes predictor M_{Bayes} satisfies the same α-Hölder condition as in Theorem 2 with exponent $\alpha \in (0,1]$ and constant L. Then, for any θ,*

$$\big|R_{\mathrm{BG}}^{(\mathcal{Q}_X)}(M_\theta) - R_{\mathrm{BG}}^{(\mathcal{P}_X)}(M_\theta)\big| \;\le\; \frac{2(B_M+B_f)}{p}\sum_{k=1}^{p}\big(L+\Lambda_\alpha\big)\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k),\mathcal{L}_Q(P^k)\big),$$

where

$$\Lambda_\alpha := \big(L_s\mathrm{Lip}(\phi_\theta) + L_c\big)\big(\mathrm{diam}(\mathcal{U}) + \mathrm{diam}(\mathcal{C})\big)^{1-\alpha}.$$

In particular, when $\alpha = 1$, $\Lambda_1 = L_s\,\mathrm{Lip}(\phi_\theta) + L_c$.

Proof of Theorem 7. Fix $k \in \{1,\dots,p\}$ and abbreviate $\boldsymbol{z} = (\boldsymbol{u}_{1:k},\boldsymbol{c}) \in \mathcal{U}^k\times\mathcal{C}$, $s(\boldsymbol{z}) := \frac{1}{k}\sum_{i=1}^{k}\phi_\theta(\boldsymbol{u}_i) \in \Delta^{m-1}$, and $M_\theta(\boldsymbol{z}) := \mathrm{clip}_{[-B_M,B_M]}\big(\rho_\theta(s(\boldsymbol{z}),\boldsymbol{c})\big)$. Write the Bayes predictor as $M_{\mathrm{Bayes}}(\boldsymbol{z}) := M_{\mathrm{Bayes}}(\boldsymbol{u}_{1:k},\boldsymbol{c})$ and introduce

$$g_k(\boldsymbol{z}) := \big(M_\theta(\boldsymbol{z}) - M_{\mathrm{Bayes}}(\boldsymbol{z})\big)^2, \qquad h_k(\boldsymbol{z}) := M_\theta(\boldsymbol{z}) - M_{\mathrm{Bayes}}(\boldsymbol{z}).$$

Step 1 (Lipschitz modulus of M_θ under $\overline{d}_{k,\alpha}$). By the network size assumption and the 1-Lipschitzness of clipping,

$$\big|M_\theta(\boldsymbol{z}) - M_\theta(\boldsymbol{z}')\big| \le L_s\big\|s(\boldsymbol{z}) - s(\boldsymbol{z}')\big\|_2 + L_c\|\boldsymbol{c}-\boldsymbol{c}'\|_2.$$

Let $L_\phi := \mathrm{Lip}(\phi_\theta)$ (for our encoder with Renorm_τ, $L_\phi \le \frac{2\sqrt{m}}{\tau}\,S(g_\theta)$). Since ϕ_θ is L_ϕ–Lipschitz,

$$\|s(\boldsymbol{z}) - s(\boldsymbol{z}')\|_2 \le \frac{1}{k}\sum_{i=1}^{k}\|\phi_\theta(\boldsymbol{u}_i) - \phi_\theta(\boldsymbol{u}'_i)\|_2 \le \frac{L_\phi}{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{u}'_i\|_2.$$

Let $D_U := \mathrm{diam}(\mathcal{U})$ and $D_C := \mathrm{diam}(\mathcal{C})$, and put $D := D_U + D_C$. For $0 < \alpha \le 1$ and any $t \in [0, D]$ one has $t \le D^{1-\alpha}t^\alpha$; hence

$$\frac{1}{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{u}'_i\|_2 \le D_U^{1-\alpha}\frac{1}{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{u}'_i\|_2^\alpha, \qquad \|\boldsymbol{c}-\boldsymbol{c}'\|_2 \le D_C^{1-\alpha}\|\boldsymbol{c}-\boldsymbol{c}'\|_2^\alpha.$$

Using the prompt-level metric $\overline{d}_{k,\alpha}(\boldsymbol{z},\boldsymbol{z}') = \frac{1}{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{u}'_i\|_2^\alpha + \|\boldsymbol{c}-\boldsymbol{c}'\|_2^\alpha$, we obtain

$$\big|M_\theta(\boldsymbol{z}) - M_\theta(\boldsymbol{z}')\big| \le \big(L_s\mathrm{Lip}(\phi_\theta) + L_c\big)D^{1-\alpha}\overline{d}_{k,\alpha}(\boldsymbol{z},\boldsymbol{z}') = \Lambda_\alpha\overline{d}_{k,\alpha}(\boldsymbol{z},\boldsymbol{z}').$$

Step 2 (Lipschitz modulus of h_k and g_k). By the assumption on the Bayes predictor, M_{Bayes} is α–Hölder with constant L under the $\overline{d}_{k,\alpha}$, hence

$$|h_k(\boldsymbol{z}) - h_k(\boldsymbol{z}')| = \big|M_\theta(\boldsymbol{z}) - M_\theta(\boldsymbol{z}') - \big(M_{\text{Bayes}}(\boldsymbol{z}) - M_{\text{Bayes}}(\boldsymbol{z}')\big)\big| \le \big(\Lambda_\alpha + L\big)\overline{d}_{k,\alpha}(\boldsymbol{z}, \boldsymbol{z}').$$

Because $|M_\theta| \le B_M$ and $|M_{\text{Bayes}}| \le B_f$, the range of h_k is contained in $[-(B_M + B_f), B_M + B_f]$. Therefore, using $|a^2 - b^2| = |(a-b)(a+b)| \le 2(B_M + B_f)|a - b|$,

$$\big|g_k(\boldsymbol{z}) - g_k(\boldsymbol{z}')\big| \le 2(B_M + B_f)\big|h_k(\boldsymbol{z}) - h_k(\boldsymbol{z}')\big| \le 2(B_M + B_f)(L + \Lambda_\alpha)\overline{d}_{k,\alpha}(\boldsymbol{z}, \boldsymbol{z}').$$

Thus g_k is $\overline{d}_{k,\alpha}$–Lipschitz with modulus $2(B_M + B_f)(L + \Lambda_\alpha)$.

Step 3 (Kantorovich–Rubinstein duality and averaging over k). Let $\mathcal{L}_P(P^k)$ and $\mathcal{L}_Q(P^k)$ denote the distributions of the length-k prompts under the source and target domains, respectively. Kantorovich–Rubinstein duality for $W_1(\cdot, \cdot; \overline{d}_{k,\alpha})$ implies, for any Lipschitz g_k,

$$\big|\mathbb{E}_Q g_k(P^k) - \mathbb{E}_P g_k(P^k)\big| \le \text{Lip}_{\overline{d}_{k,\alpha}}(g_k)\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k), \mathcal{L}_Q(P^k)\big),$$

where $\mathsf{W}_\alpha^{(k)} := W_1(\cdot, \cdot; \overline{d}_{k,\alpha})$. By the definition of the Bayes Gap under a prompt distribution P,

$$R_{\text{BG}}^{(\mathsf{P})}(M_\theta) = \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{\mathsf{P}}\left[g_k(P^k)\right].$$

Combining the last two displays and the Lipschitz bound from Step 2 yields

$$\left|R_{\text{BG}}^{(\mathcal{Q}_X)}(M_\theta) - R_{\text{BG}}^{(\mathcal{P}_X)}(M_\theta)\right| \le \frac{2(B_M + B_f)}{p}\sum_{k=1}^{p}\big(L + \Lambda_\alpha\big)\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k), \mathcal{L}_Q(P^k)\big),$$

which is exactly the claimed inequality. □

The prompt $P^k = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k, \boldsymbol{C})$ contains dependent coordinates in general, because the context responses $\boldsymbol{U}_i = (\boldsymbol{X}_i, Y_i)$ share the latent task function f within a prompt. Therefore, a direct product of coordinate-wise optimal couplings is not a valid coupling of the prompt distributions. The following conditional coupling fixes this.

Remark 4 (Prompt-level Wasserstein via conditional coupling)**.** Let S be a latent seed that is shared across domains and determines the task index and task function. For instance, one may take $S = (I, f)$. Conditional on S, the prompt coordinates factorize as

$$\mathcal{L}_P(P^k \mid S) = \big(\mathcal{L}_P(U \mid S)\big)^{\otimes k} \times \mathcal{P}_X, \qquad \mathcal{L}_Q(P^k \mid S) = \big(\mathcal{L}_Q(U \mid S)\big)^{\otimes k} \times \mathcal{Q}_X,$$

where $\boldsymbol{U} = (\boldsymbol{X}, Y)$ and $\mathcal{P}_X, \mathcal{Q}_X$ are the (source/target) input distributions. In particular, conditional on S the k context pairs are i.i.d. under each domain. (If one prefers to carry a coupling of the additive noise across domains, introduce an exogenous noise seed that determines the noise distribution but not its realized sample path; this preserves conditional i.i.d.)

Lemma 2 (Conditional product-type upper bound for prompt-level Wasserstein)**.** *Under the setting of Remark 4, for every $k \ge 1$ and $0 < \alpha \le 1$,*

$$\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k), \mathcal{L}_Q(P^k)\big) \le \mathbb{E}_S\left[\mathsf{W}_\alpha\big(\mathcal{L}_P(\boldsymbol{U} \mid S), \mathcal{L}_Q(\boldsymbol{U} \mid S)\big)\right] + \mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X),$$

where the prompt-level ground metric is

$$\overline{d}_{k,\alpha}\big((\boldsymbol{u}_{1:k}, \boldsymbol{c}), (\boldsymbol{u}'_{1:k}, \boldsymbol{c}')\big) := \frac{1}{k}\sum_{i=1}^{k} \|\boldsymbol{u}_i - \boldsymbol{u}'_i\|_2^\alpha + \|\boldsymbol{c} - \boldsymbol{c}'\|_2^\alpha,$$

and for single pairs $\boldsymbol{U} = (\boldsymbol{X}, Y)$ we write $\mathsf{W}_\alpha(\cdot, \cdot) := W_1(\cdot, \cdot; \|\cdot\|_2^\alpha)$.

Proof of Lemma 2. **Step 1 (conditional product coupling).** Fix $S = s$. By Remark 4, under each domain the k context coordinates are i.i.d. with common conditional distribution $\mathcal{L}_P(\boldsymbol{U} \mid s)$ (resp. $\mathcal{L}_Q(\boldsymbol{U} \mid s)$), while the query coordinate has distribution $\mathcal{P}_X$ (resp. $\mathcal{Q}_X$) independent of the context. Let π_U^s be an optimal coupling for $\mathsf{W}_\alpha\big(\mathcal{L}_P(\boldsymbol{U} \mid s), \mathcal{L}_Q(\boldsymbol{U} \mid s)\big)$ with ground metric

$d_\alpha(\boldsymbol{u}, \boldsymbol{u}') := \|\boldsymbol{u} - \boldsymbol{u}'\|_2^\alpha$, and let π_C be an optimal coupling for $\mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X)$ with ground metric $d_\alpha(\boldsymbol{c}, \boldsymbol{c}') := \|\boldsymbol{c} - \boldsymbol{c}'\|_2^\alpha$. Construct a coupling Π_s of $\mathcal{L}_P(P^k \mid s)$ and $\mathcal{L}_Q(P^k \mid s)$ by drawing $(\boldsymbol{U}_i, \boldsymbol{U}_i') \overset{\text{i.i.d.}}{\sim} \pi_U^s$ for $i = 1, \dots, k$ and $(\boldsymbol{C}, \boldsymbol{C}') \sim \pi_C$, all independent across coordinates. Then, by the definition of the prompt-level ground metric,

$$\begin{aligned}\mathbb{E}_{\Pi_s}\Big[\overline{d}_{k,\alpha}\big((\boldsymbol{U}_{1:k}, \boldsymbol{C}), (\boldsymbol{U}_{1:k}', \boldsymbol{C}')\big)\Big] &= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_{\pi_U^s}\big[d_\alpha(\boldsymbol{U}_i, \boldsymbol{U}_i')\big] + \mathbb{E}_{\pi_C}\big[d_\alpha(\boldsymbol{C}, \boldsymbol{C}')\big] \\ &= \mathsf{W}_\alpha\big(\mathcal{L}_P(\boldsymbol{U} \mid s), \mathcal{L}_Q(\boldsymbol{U} \mid s)\big) + \mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X).\end{aligned}$$

Therefore,

$$\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k \mid s), \mathcal{L}_Q(P^k \mid s)\big) \le \mathsf{W}_\alpha\big(\mathcal{L}_P(\boldsymbol{U} \mid s), \mathcal{L}_Q(\boldsymbol{U} \mid s)\big) + \mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X).$$

Step 2 (disintegration and convexity). Write the unconditional prompt distributions as mixtures over S: $\mathcal{L}_P(P^k) = \int \mathcal{L}_P(P^k \mid s)\nu(\mathrm{d}s)$ and $\mathcal{L}_Q(P^k) = \int \mathcal{L}_Q(P^k \mid s)\nu(\mathrm{d}s)$, where ν is the (shared) distribution of S under both domains (task distribution and noise distribution are kept the same across domains). By convexity of $W_1(\cdot, \cdot; \overline{d}_{k,\alpha})$ in each argument,

$$\begin{aligned}\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k), \mathcal{L}_Q(P^k)\big) &\le \int \mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k \mid s), \mathcal{L}_Q(P^k \mid s)\big)\nu(\mathrm{d}s) \\ &\le \mathbb{E}_S\left[\mathsf{W}_\alpha\big(\mathcal{L}_P(\boldsymbol{U} \mid S), \mathcal{L}_Q(\boldsymbol{U} \mid S)\big)\right] + \mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X),\end{aligned}$$

which is the desired bound. □

Corollary 3 (Input-only reduction under Lipschitz tasks)**.** *Assume $Y = f(X) + \varepsilon$ with a shared noise coupling across domains (possibly conditional on S) and a task family that is uniformly L_f-Lipschitz in x: $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \le L_f\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ for all tasks. Then for every $k \ge 1$ and $0 < \alpha \le 1$,*

$$\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k), \mathcal{L}_Q(P^k)\big) \;\le\; (2 + L_f^\alpha)\mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X).$$

Proof of Corollary 3. Under the shared-noise coupling, by subadditivity of $t \mapsto t^\alpha$ for $\alpha \in (0, 1]$, $\|\boldsymbol{U} - \boldsymbol{U}'\|_2^\alpha \le \|\boldsymbol{X} - \boldsymbol{X}'\|_2^\alpha + |f(\boldsymbol{X}) - f(\boldsymbol{X}')|^\alpha \le (1 + L_f^\alpha)\|\boldsymbol{X} - \boldsymbol{X}'\|_2^\alpha$. Hence $\mathsf{W}_\alpha(\mathcal{L}_P(\boldsymbol{U} \mid S), \mathcal{L}_Q(\boldsymbol{U} \mid S)) \le (1 + L_f^\alpha)\mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X)$ for every S. Plug this into Lemma 2 and add the $\mathsf{W}_\alpha(\mathcal{P}_X, \mathcal{Q}_X)$ term for the query coordinate $\boldsymbol{C}$. □

Combining Theorem 1, Theorem 2, Theorem 7 with either Lemma 2 or Corollary 3, and absorbing polylogarithms into $\tilde{O}(\cdot)$, yields the same end-to-end OOD risk bound as in the main text, with the prompt-level Wasserstein term. The additional terms quantify distribution shift incurred during pretraining (via $\mathcal{L}_P(P^k)$ vs. $\mathcal{L}_Q(P^k)$); once θ is fixed, the inference-time predictor risk R_{PV} is evaluated under the target domain alone and does not carry extra estimation error from pretraining. Putting everything together, for the target domain $\mathcal{Q}_X$ we obtain

$$\begin{aligned}\mathbb{E}R^{(\mathcal{Q}_X)}(M_{\hat{\theta}}) \le \quad & \underbrace{\frac{1}{p}\sum_{k=1}^{p} R_k^\star(F_{i^\star}; \mathcal{Q}_X)}_{\text{oracle risk under the true task type in the target domain}} \qquad + \tilde{O}\left(m^{-\frac{2\alpha}{d_{\text{eff}}}} + \frac{m}{pN} + \frac{1}{N}\right) \\ & + \underbrace{\frac{2(B_M + B_f)}{p}\sum_{k=1}^{p}\big(L + \Lambda_\alpha\big)\,\mathsf{W}_\alpha^{(k)}\big(\mathcal{L}_P(P^k), \mathcal{L}_Q(P^k)\big)}_{\text{OOD penalty on the Bayes Gap}} + \underbrace{\frac{5B_f^2}{p}\left(\frac{\frac{1-\alpha_{i^\star}}{\alpha_{i^\star}}}{e^{D_{\min}/2} - 1} + \frac{T - 1}{e^{C} - 1}\right)}_{\text{mixture identification remainder}},\end{aligned}$$

where $R_k^\star(F_{i^\star}; \mathcal{Q}_X)$ denotes the minimax risk for predicting a function from the true task class $F_{i^\star}$ under prompt distribution $\mathcal{Q}_X$.

E On the Hölder Condition of the Bayes Predictor

Although y is not bounded in the prompt-generating process, the theorem imposes the Hölder condition on bounded examples and queries. This is because if the noise follows a sub-Gaussian distribution, boundedness holds with high probability. Hence, the unbounded cases do not significantly affect the final result. Moreover, since the statement of Theorem 2 is a bound on the expectation, it suffices that the Hölder condition holds with high probability.

In addition to Assumptions 1-2, assume the noise is Gaussian for simplicity. Under these conditions, the Bayes predictor M_{Bayes} is Hölder ($\alpha = 1$), with the family-specific Hölder constants listed below:

- Linear regression. $f_{(w,b)}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ with $\|\boldsymbol{w}\|_2 \le B_w$, $|b| \le B_b$ and feature map $\psi(\boldsymbol{x}) = [\boldsymbol{x}^\top, 1]^\top$. Then $\boldsymbol{x} \mapsto f_{(w,b)}(\boldsymbol{x})$ is B_w-Lipschitz.
- Finite-order series regression. $f_a(\boldsymbol{x}) = \sum_{j=1}^{R} a_j g_j(\boldsymbol{x})$ with $\|\boldsymbol{a}\|_1 \le A$, basis functions satisfying $\|g_j\|_\infty \le 1$ and $\|\nabla g_j(\boldsymbol{x})\|_2 \le L_g$ uniformly; take $\psi(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \ldots, g_R(\boldsymbol{x})]^\top$. Then $\boldsymbol{x} \mapsto f_a(\boldsymbol{x})$ is AL_g-Lipschitz.
- Finite convex dictionary. $f_a = \sum_{j=1}^{J} a_j f^{(j)}$ with $\boldsymbol{a} \in \Delta^{J-1}$, each atom obeying $|f^{(j)}(\boldsymbol{x})| \le B_f$ and $\|\nabla f^{(j)}(\boldsymbol{x})\|_2 \le L_f$ uniformly; take $\psi(\boldsymbol{x}) = [f^{(1)}(\boldsymbol{x}), \ldots, f^{(J)}(\boldsymbol{x})]^\top$. Then $\boldsymbol{x} \mapsto f_a(\boldsymbol{x})$ is L_f-Lipschitz. (An example of a distribution on Δ^{J-1} is the logistic-normal distribution (Aitchison & Shen, 1980).)

We consider these three regression models. For these models, we additionally assume the following conditions:

- For task family i, there exist a dimension $d_i \in \mathbb{N}$ and a parameter space $\Theta_i \subset \mathbb{R}^{d_i}$ such that $f_{\boldsymbol{\theta}} : \mathcal{C} \to \mathbb{R}$ for every $\boldsymbol{\theta} \in \Theta_i$. Moreover, the model is uniformly bounded and Hölder in the query: $\sup_{\boldsymbol{\theta} \in \Theta_i} \sup_{\boldsymbol{x} \in \mathcal{C}} |f_{\boldsymbol{\theta}}(\boldsymbol{x})| \le B_f$ and $\sup_{\boldsymbol{\theta} \in \Theta_i} \sup_{\boldsymbol{x} \ne \boldsymbol{x}'} \frac{|f_{\boldsymbol{\theta}}(\boldsymbol{x}) - f_{\boldsymbol{\theta}}(\boldsymbol{x}')|}{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^\alpha} \le L_{f,i}$.
- The distribution on $\boldsymbol{\theta}$ given $I = i$ has a density $\pi_i(\boldsymbol{\theta}) \propto \exp\{-V(\boldsymbol{\theta})\}$ on Θ_i, where V is twice continuously differentiable and $\nabla^2 V(\boldsymbol{\theta}) \succeq \lambda_0 I_{d_i}$ for all $\boldsymbol{\theta} \in \Theta_i$, for some $\lambda_0 > 0$. In particular, a Gaussian distribution $\mathcal{N}(0, \Lambda_0^{-1})$ satisfies this with $\lambda_0 = \lambda_{\min}(\Lambda_0)$.
- Let $u = (\boldsymbol{x}, y) \in \mathcal{U}$ and define the per-sample loss $\tilde{\ell}(\boldsymbol{\theta}; u) := \frac{1}{2\sigma_\varepsilon^2}\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y\big)^2$. There exists a constant $L_{\boldsymbol{\theta},i} < \infty$ such that, for all $\boldsymbol{\theta} \in \Theta_i$ and all $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in \mathcal{U}$, $\big\|\nabla_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{u}) - \nabla_{\boldsymbol{\theta}} \tilde{\ell}(\boldsymbol{\theta}; \tilde{\boldsymbol{u}})\big\|_2 \le L_{\boldsymbol{\theta},i} \, \|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_2^\alpha$.
- There exists b such that $\frac{1}{k} \sum_{t=1}^{k} \psi(\boldsymbol{x}_t) \psi^\top(\boldsymbol{x}_t) \succeq bI$.

In the mixture setting, the Bayes predictor decomposes as $M_{\mathrm{Bayes}}(P^k) = \sum_{i=1}^{T} \pi_i(D^k)\, \mu_i(D^k, \boldsymbol{c})$ with $\mu_i(D^k, \boldsymbol{c}) := \mathbb{E}[f(\boldsymbol{c}) \mid D^k, I = i]$ and $\pi_i(D^k) \propto \alpha_i \, m_i(D^k)$, $m_i(D^k) := \int \exp\big\{ -(2\sigma_\varepsilon^2)^{-1} \sum_{r=1}^{k} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_r) - y_r)^2 \big\} \, \mathrm{d}\mathcal{P}_i(\boldsymbol{\theta})$. The above single-family arguments and the assumptions imply that each μ_i is α-Hölder with a constant $L_{\mu,i}$ independent of k. However, the mixture weights π_i depend on the entire context through the marginal evidences $m_i(D^k)$. Then, there exists C_i such that $\log m_i(D^k)$ is kC_i-Hölder in D^k when measured by the average per-sample metric: $|\log m_i(D^k) - \log m_i(\tilde{D}^k)| \le k\, C_i \, \frac{1}{k} \sum_{r=1}^{k} \big(\|\boldsymbol{u}_r - \tilde{\boldsymbol{u}}_r\|_2\big)$. Hence, in the worst case, the softmax gating $D^k \mapsto \pi(D^k)$ is $O(k)$-Hölder under the same metric, and $|M_{\mathrm{Bayes}}(P^k) - M_{\mathrm{Bayes}}(\tilde{P}^k)| \le \big(\max_i L_{\mu,i} + B_f C k\big) \, \frac{1}{k} \sum_{r=1}^{k} \|(\boldsymbol{u}_r, \boldsymbol{c}) - (\tilde{\boldsymbol{u}}_r, \tilde{\boldsymbol{c}})\|_2$ for some C.

If, in addition, the standard log-likelihood-ratio conditions in Theorem 3 hold, the task posterior $\mathcal{P}_{I|D^k}$ concentrates exponentially fast on the true index $i^\star$. From Step 3 in proof of Theorem 3, with probability at least $1 - e^{-C_1 k}$, $\sum_{i \ne i^\star} \pi_i(D^k) \le C_2 T k e^{-C_3 k}$ holds for some C_1, C_2, and C_3, implying $|M_{\mathrm{Bayes}}(P^k) - M_{\mathrm{Bayes}}(\tilde{P}^k)| \le \big(L_{\mu,i^\star} + 2B_f C_2 T k e^{-C_3 k}\big) \frac{1}{k} \sum_{r=1}^{k} \|(\boldsymbol{u}_r, \boldsymbol{c}) - (\tilde{\boldsymbol{u}}_r, \tilde{\boldsymbol{c}})\|_2$. In particular, the effective Hölder constant is independent of k up to an exponentially small remainder. Also, a uniform margin assumption $\min_{i \ne j} \frac{1}{k} |\log m_i(D^k) - \log m_j(D^k)| \ge \gamma > 0$ implies the same conclusion with $e^{-\gamma k}$ in place of $e^{-C_3 k}$.

F Details of Theorem 3

We concretely investigate Theorem 3 for a pair of task families: *linear regression* versus a *series (basis) regression* that excludes constant and linear terms.

Standing assumptions.

- Inputs are bounded and i.i.d.: $\boldsymbol{X} \sim \mathcal{P}_X$ with $\|\boldsymbol{X}\|_2 \leq B_X$ a.s. and $\mathbb{E}[\boldsymbol{X}] = 0$. Let $\Sigma_X := \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$, which we assume is positive definite on $\mathbb{R}^{d_{\text{feat}}}$ with $\lambda_{\min}(\Sigma_X) > 0$.
- Noise is Gaussian (a special case of sub-Gaussian): $\varepsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ independent of $(f, \boldsymbol{X})$.
- Boundedness of tasks. For the linear class
$$F_{\text{lin}} = \big\{f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{x} + b : \ \|\boldsymbol{w}\|_2 \leq B_w, \ |b| \leq B_b\big\},$$
we have $|f_{w,b}(\boldsymbol{x})| \leq B_w B_X + B_b =: B_f$ on the support of $\mathcal{P}_X$. For the series class
$$F_{\text{ser}} = \Big\{f_a(\boldsymbol{x}) = \sum_{r=r_0}^{R_{\max}} a_r g_r(\boldsymbol{x}) : \ \|\boldsymbol{a}\|_2 \leq B_a\Big\},$$
assume $r_0 \geq 2$ (so constant and linear terms are excluded), the basis $\{g_r\}_{r=r_0}^{R_{\max}}$ is orthonormal in $L^2(\mathcal{P}_X)$, orthogonal to linear functions, and bounded pointwise, i.e. $\sup_x |g_r(\boldsymbol{x})| \leq G_{\max}$. Then $|f_a(\boldsymbol{x})| \leq \|\boldsymbol{a}\|_2 \cdot \|\big(g_r(\boldsymbol{x})\big)_{r=r_0}^{R_{\max}}\|_2 \leq B_a\sqrt{R_{\max} - r_0 + 1}G_{\max} =: B_f$.
- Within each family we use a truncated Gaussian parameter distribution supported on the above bounded parameter sets (to respect $|f| \leq B_f$) and otherwise conjugate:
$$\boldsymbol{\theta}_{\text{lin}} := (\boldsymbol{w}, b) \sim \mathcal{N}(0, \tau_{\text{lin}}^2\mathbf{I}) \text{ truncated to } \{\|\boldsymbol{w}\| \leq B_w, \ |b| \leq B_b\},$$
$$\boldsymbol{\theta}_{\text{ser}} := \boldsymbol{a} \sim \mathcal{N}(0, \tau_{\text{ser}}^2\mathbf{I}) \text{ truncated to } \{\|\boldsymbol{a}\|_2 \leq B_a\}.$$
The truncation preserves boundedness; the standard Gaussian formulas below give upper bounds (hence valid constants) for the truncated case because the posterior covariances are $\preceq$ their untruncated analogues on the bounded domain.

A generic Gaussian–predictive bound for $Z_{j,t}$ Fix a time t and condition on $\mathcal{G}_{t-1}$ and $\boldsymbol{X}_t = \boldsymbol{x}$. Under any task type (task family) i, the (posterior) predictive distribution is Gaussian
$$p_i(y \mid \boldsymbol{x}, \mathcal{G}_{t-1}) = \mathcal{N}\big(\mu_{i,t}(\boldsymbol{x}), \ s_{i,t}^2(\boldsymbol{x})\big),$$
with mean $\mu_{i,t}(\boldsymbol{x})$ (the posterior mean of $f(\boldsymbol{x})$) and predictive variance
$$s_{i,t}^2(\boldsymbol{x}) = \sigma_\varepsilon^2 + \mathrm{Var}\big(f(\boldsymbol{x}) \mid \mathcal{G}_{t-1}, I = i\big).$$
For our conjugate priors,
$$\mu_{\text{lin},t}(\boldsymbol{x}) = \phi(\boldsymbol{x})^\top\boldsymbol{m}_{t-1}, \quad s_{\text{lin},t}^2(\boldsymbol{x}) = \sigma_\varepsilon^2 + \phi(\boldsymbol{x})^\top\Sigma_{t-1}\,\phi(\boldsymbol{x}), \quad \phi(\boldsymbol{x}) := \begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix},$$
and similarly
$$\mu_{\text{ser},t}(\boldsymbol{x}) = \psi(\boldsymbol{x})^\top\tilde{\boldsymbol{m}}_{t-1}, \quad s_{\text{ser},t}^2(\boldsymbol{x}) = \sigma_\varepsilon^2 + \psi(\boldsymbol{x})^\top\tilde{\Sigma}_{t-1}\,\psi(\boldsymbol{x}), \quad \psi(\boldsymbol{x}) := \big(g_r(\boldsymbol{x})\big)_{r=r_0}^{R_{\max}}.$$
Because $\|\phi(\boldsymbol{x})\|_2 \leq B_\phi := \sqrt{B_X^2 + 1}$ and $\|\psi(\boldsymbol{x})\|_2 \leq B_\psi := \sqrt{R_{\max} - r_0 + 1}\,G_{\max}$ and the posterior covariances are bounded by the prior covariances, we have a uniform variance upper bound
$$s_{i,t}^2(\boldsymbol{x}) \leq \sigma_\varepsilon^2 + \bar{V}, \qquad \bar{V} := \max\{\tau_{\text{lin}}^2 B_\phi^2, \ \tau_{\text{ser}}^2 B_\psi^2\}. \tag{2}$$

For two types i (true) and j (wrong), define the log-predictive increment[3]
$$Z_{j,t} := \log\frac{p_j(Y_t \mid \boldsymbol{X}_t, \mathcal{G}_{t-1})}{p_i(Y_t \mid \boldsymbol{X}_t, \mathcal{G}_{t-1})}.$$

[3] If the likelihood does not have a density function with respect to Lebesgue measure, assume that all predictive distributions are dominated by a common reference measure so that the Radon-Nikodym derivative exists. Then $Z_{j,t}$ can be rigorously defined.

A direct Gaussian calculation (writing $Y_t = \mu_{i,t}(\boldsymbol{X}_t) + s_{i,t}(\boldsymbol{X}_t)\,\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$) yields

$$\begin{aligned}\mathbb{E}[Z_{j,t} \mid \mathcal{G}_{t-1}, \boldsymbol{X}_t] &= -\mathrm{KL}\Big(\mathcal{N}(\mu_{i,t}, s_{i,t}^2) \,\Big\|\, \mathcal{N}(\mu_{j,t}, s_{j,t}^2)\Big) \\ &= -\frac{1}{2}\left\{\log\frac{s_{j,t}^2}{s_{i,t}^2} + \frac{s_{i,t}^2}{s_{j,t}^2} - 1 + \frac{(\mu_{i,t} - \mu_{j,t})^2}{s_{j,t}^2}\right\}.\end{aligned}$$

Consequently, for every $(t, \boldsymbol{x})$,

$$\mathbb{E}[Z_{j,t} \mid \mathcal{G}_{t-1}, \boldsymbol{X}_t = \boldsymbol{x}] \le -\frac{(\mu_{i,t}(\boldsymbol{x}) - \mu_{j,t}(\boldsymbol{x}))^2}{2\,s_{j,t}^2(\boldsymbol{x})} \le -\frac{(\mu_{i,t}(\boldsymbol{x}) - \mu_{j,t}(\boldsymbol{x}))^2}{2(\sigma_\varepsilon^2 + \bar{V})}. \tag{3}$$

Moreover, the centered increment $Z_{j,t} + D_{j,t}$ with $D_{j,t} := -\mathbb{E}[Z_{j,t} \mid \mathcal{G}_{t-1}, \boldsymbol{X}_t]$ is a quadratic polynomial in a standard normal,

$$Z_{j,t} + D_{j,t} = a_t\,\varepsilon + b_t\,(\varepsilon^2 - 1), \quad a_t := -\frac{(\mu_{i,t} - \mu_{j,t})\,s_{i,t}}{s_{j,t}^2}, \quad b_t := -\frac{1}{2}\left(\frac{s_{i,t}^2}{s_{j,t}^2} - 1\right),$$

hence sub-exponential. Calculating the mgf

$$\mathbb{E}e^{\lambda(a_t\varepsilon + b_t(\varepsilon^2-1))} = e^{-\lambda b_t}(1 - 2\lambda b_t)^{-1/2}\exp\left(\frac{\lambda^2 a_t^2}{2(1 - 2\lambda b_t)}\right),$$

and the elementary bound $-\ln(1-u) - u \le u^2$ valid for $|u| \le 1/2$ (note that $|b_t| \le \bar{V}/(2\sigma_\varepsilon^2)$, so $u = 2\lambda b_t \in [-1/2, 1/2]$ whenever $|\lambda| \le 1/b_j$), we obtain the uniform sub-exponential parameters (ν_j, b_j) in Theorem 3 with

$$\nu_j^2 \le \frac{8B_f^2\big(\sigma_\varepsilon^2 + \bar{V}\big)}{\sigma_\varepsilon^4} + \frac{\bar{V}^2}{\sigma_\varepsilon^4}, \qquad b_j := \frac{2\bar{V}}{\sigma_\varepsilon^2}.$$

Pair A: true linear vs. wrong series (degree $\ge$ 2) Assume the data are generated by some $f^\star(\boldsymbol{x}) = \boldsymbol{w}_\star^\top \boldsymbol{x} + b_\star \in F_{\mathrm{lin}}$ and the wrong family is F_{ser} with orthonormal $\{g_r\}_{r=r_0}^{R_{\max}}$, $r_0 \ge 2$, orthogonal to 1 and to all linear functionals of $\boldsymbol{X}$. Let Π_{ser} denote the $L^2(\mathcal{P}_X)$–orthogonal projection onto $\mathrm{span}\{g_r\}$.

By orthogonality, $\Pi_{\mathrm{ser}} f^\star \equiv 0$, hence the L^2–gap between the true function and the wrong family is

$$\Delta_{\mathrm{lin}\to\mathrm{ser}}^2 := \big\|f^\star - \Pi_{\mathrm{ser}} f^\star\big\|_{L^2(\mathcal{P}_X)}^2 = \mathbb{E}\big[(\boldsymbol{w}_\star^\top \boldsymbol{X} + b_\star)^2\big] = \boldsymbol{w}_\star^\top \Sigma_X \boldsymbol{w}_\star + b_\star^2.$$

For conjugate normal models with bounded regressors ϕ, ψ and positive definite design covariances, standard ridge-risk bounds in series regression (§3.4 in van der Vaart & Wellner, 2023) give

$$\|\mu_{\mathrm{lin},t} - f^\star\|_{L^2(\mathcal{P}_X)}^2 = O\left(\frac{d_{\mathrm{feat}} + 1}{t}\right), \qquad \|\mu_{\mathrm{ser},t} - \Pi_{\mathrm{ser}} f^\star\|_{L^2(\mathcal{P}_X)}^2 = O\left(\frac{R_{\max} - r_0 + 1}{t}\right).$$

Thus, taking $t_0 = \tilde{O}\left(\frac{d_{\mathrm{feat}} + R_{\max} - r_0 + 1}{\Delta_{\mathrm{lin}\to\mathrm{ser}}^2}\right)$, for all $t \ge t_0$, we have

$$\mathbb{E}_X\left[(\mu_{\mathrm{lin},t}(\boldsymbol{X}) - \mu_{\mathrm{ser},t}(\boldsymbol{X}))^2\right] \ge \frac{1}{2}\,\Delta_{\mathrm{lin}\to\mathrm{ser}}^2.$$

Combining with (3) and $s_{\mathrm{ser},t}^2 \le \sigma_\varepsilon^2 + \bar{V}$ gives the uniform negative drift (for all $t \ge t_0$)

$$\mathbb{E}[Z_{j,t} \mid \mathcal{G}_{t-1}] \le -D_j, \qquad D_j := \frac{\Delta_{\mathrm{lin}\to\mathrm{ser}}^2}{4\big(\sigma_\varepsilon^2 + \bar{V}\big)}.$$

From Theorem 3, the posterior mass on the wrong family is

$$\frac{1 - \alpha_{i^\star}}{\alpha_{i^\star}}\exp\left(-\frac{D_j}{2}\,k\right) + \exp(-C_j k), \quad C_j := \frac{D_j^2}{8\big(\nu_j^2 + b_j D_j/2\big)}.$$

Therefore, to make the mixture identification remainder $\le \eta$, it suffices (up to absolute constants and polylog factors) to take

$$k = \tilde{O}\left(\frac{\sigma_\varepsilon^2 + \bar{V}}{\Delta_{\mathrm{lin}\to\mathrm{ser}}^2}\log\frac{1}{\eta} \;\vee\; \left[\frac{(\sigma_\varepsilon^2 + \bar{V})^2}{\Delta_{\mathrm{lin}\to\mathrm{ser}}^4\,\sigma_\varepsilon^4}\left(B_f^2(\sigma_\varepsilon^2 + \bar{V}) + \bar{V}^2\right) + \frac{(\sigma_\varepsilon^2 + \bar{V})\,\bar{V}}{\Delta_{\mathrm{lin}\to\mathrm{ser}}^2\,\sigma_\varepsilon^2}\right]\log\frac{1}{\eta}\right). \tag{4}$$

The first term is the dominant, interpretable signal-to-noise scaling:

$$k \asymp \frac{\sigma_\varepsilon^2 + \bar{V}}{\boldsymbol{w}_\star^\top \Sigma_X \boldsymbol{w}_\star + b_\star^2}\log\frac{1}{\eta}.$$

Pair B: true series (degree $\geq$ 2) vs. wrong linear Now the data come from $f^\star(\boldsymbol{x}) = \sum_{r=r_0}^{R_{\max}} a_r^\star g_r(\boldsymbol{x})$ with $\|\boldsymbol{a}^\star\|_2 \leq B_a$ and the wrong family is linear. Orthogonality gives $\Pi_{\text{lin}} f^\star \equiv 0$ (since $r_0 \geq 2$ and $\mathbb{E}[\boldsymbol{X}] = 0$), hence

$$\Delta^2_{\text{ser}\to\text{lin}} := \left\|f^\star - \Pi_{\text{lin}} f^\star\right\|^2_{L^2(\mathcal{P}_X)} = \|f^\star\|^2_{L^2(\mathcal{P}_X)} = \sum_{r=r_0}^{R_{\max}} (a_r^\star)^2.$$

Exactly the same argument as above yields, for $t \geq t_0 = \tilde{O}\Big(\frac{d_{\text{feat}}+R_{\max}-r_0+1}{\Delta^2_{\text{ser}\to\text{lin}}}\Big)$,

$$D_j = \frac{\Delta^2_{\text{ser}\to\text{lin}}}{4\big(\sigma_\varepsilon^2 + \bar{V}\big)}, \qquad \nu_j^2 \leq \frac{8B_f^2(\sigma_\varepsilon^2+\bar{V}) + \bar{V}^2}{\sigma_\varepsilon^4}, \qquad b_j = \frac{2\bar{V}}{\sigma_\varepsilon^2},$$

and the same k–order as in (4) with $\Delta_{\text{lin}\to\text{ser}}$ replaced by $\Delta_{\text{ser}\to\text{lin}}$.

Remarks and extensions All bounds above use only: (i) $|f| \leq B_f$ on the support of $\mathcal{P}_X$; (ii) the uniform predictive variance upper bound (2); and (iii) $L^2(\mathcal{P}_X)$–orthogonality for the two families considered. Using truncated conjugate priors guarantees (i) and keeps (ii) finite with the explicit $\bar{V}$ given. The sub-exponential constants remain valid for truncated conjugate posteriors because truncation can only decrease posterior covariances, hence decrease $|a_t|$ and $|b_t|$.

If $\{g_r\}$ is merely linearly independent (non-orthonormal) let Π_{ser} be the $L^2(\mathcal{P}_X)$–projection onto $\operatorname{span}\{g_r\}$. Then the formulas hold with

$$\Delta^2_{\text{lin}\to\text{ser}} = \left\|f^\star - \Pi_{\text{ser}} f^\star\right\|^2_{L^2(\mathcal{P}_X)}, \qquad \Delta^2_{\text{ser}\to\text{lin}} = \left\|f^\star - \Pi_{\text{lin}} f^\star\right\|^2_{L^2(\mathcal{P}_X)},$$

and the same (ν_j, b_j) (with the same $\bar{V}$) because the mgf bound depended only on boundedness and Eq. (2).

G Technical Lemmas

Lemma 3 (Posterior variance is bounded by the true task's minimax risk)**.** *Suppose the prompt-generating process is as described in Definition 1 and that Assumptions 1-2 hold. Fix a task-type index $i^\star \in \{1,\ldots,T\}$ and recall that $F_{i^\star} = \operatorname{supp}(\mathcal{P}_{F_{i^\star}})$ is the corresponding function class (support of the true task type prior). For any $k \geq 1$,*

$$\mathbb{E}_{f\sim\mathcal{P}_{F_{i^\star}}} \mathbb{E}_{D^k\sim\mathcal{P}^{\otimes k}_{X,Y|f}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_X}\left[\operatorname{Var}_{f\sim\mathcal{P}_{F_{i^\star}|D^k}}(f(\boldsymbol{x}))\right] \leq \inf_M \sup_{f\in F_{i^\star}} \mathbb{E}_{P^k}\left[\big(f(\boldsymbol{x}_{k+1}) - M(P^k)\big)^2 \Big| f\right],$$

where the left-hand side is the conditional Posterior Variance average under the true task type and M belongs to the bounded and measurable function space.

This suggests that if the true task type is given, the Posterior Variance is smaller than the minimax L_2 prediction error.

Lemma 4 (Sequential covering bound)**.** *Fix $k \in \mathbb{N}_+$. Let $\mathcal{U} \subset \mathbb{R}^{d_{\text{eff}}}$ and $\mathcal{C} \subset \mathbb{R}^{d_{\text{feat}}}$ be bounded with $\sup_{\boldsymbol{u}\in\mathcal{U}} \|\boldsymbol{u}\|_2 \leq R_U$ and $\operatorname{diam}(\mathcal{C}) < \infty$. For $\theta \in \Theta$, consider the uniform-attention architecture*

$$M_\theta(P^k) = \rho_\theta\left(\frac{1}{k}\sum_{i=1}^k \phi_\theta(\boldsymbol{u}_i), \boldsymbol{c}\right), \qquad P^k = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k,\boldsymbol{c}) \in \mathcal{U}^k \times \mathcal{C},$$

where the query c is shared across the k context items within each P^k (i.e., $\boldsymbol{c}$ does not depend on i inside the mean $\frac{1}{k}\sum_{i=1}^k \phi_\theta(\boldsymbol{u}_i)$). Assume:

(i) $\phi_\theta : \mathcal{U} \to \Delta^{m-1}$ is L_ϕ–Lipschitz, where $L_\phi := \operatorname{Lip}(\phi_\theta) \leq \frac{2\sqrt{m}}{\tau} S(g_\theta)$ for our encoder with $\operatorname{Renorm}_\tau$, and the ReLU component satisfies $S(g_\theta) \leq C_\phi m^{1/d_{\text{eff}}}$. Moreover, $(\phi_\theta)_j \in [0,1]$ and $\sum_{j=1}^m (\phi_\theta)_j \equiv 1$, and ϕ_θ admits a realization with $\tilde{O}(m)$-weights and $O(\log m)$-layers. Put $B_\phi := \sup_j \|(\phi_\theta)_j\|_\infty \leq 1$.

(ii) $\rho_\theta : \Delta^{m-1} \times \mathcal{C} \to \mathbb{R}$ *is a ReLU network with spectral product* $S(\rho_\theta) \le C_\rho m^{1/2}$*, is jointly Lipschitz,*

$$|\rho_\theta(\boldsymbol{s}, \boldsymbol{c}) - \rho_\theta(\boldsymbol{s}', \boldsymbol{c}')| \le L_s \|\boldsymbol{s} - \boldsymbol{s}'\|_2 + L_c \|\boldsymbol{c} - \boldsymbol{c}'\|_2, \quad L_s, L_c \le C_\rho m^{1/2},$$

and its (clipped) output is bounded, $|\rho_\theta| \le B_M$.

Let $\mathcal{H} := \{P^k \mapsto M_\theta(P^k) - M_{\text{Bayes}}(P^k) : \theta \in \Theta\}$ *be the centered class for any fixed target* M_{Bayes}*. Denote by* $N_2^{\text{seq}}(\delta, \cdot; z)$ *the sequential covering number under the* ℓ_2 *sequential metric on a depth-*k *predictable tree* z*. Then, for all* $\delta \in (0, 2B_M]$,

$$\sup_z \log N_2^{\text{seq}}(\delta, \mathcal{H}; z) \lesssim m \log\left(\frac{\sqrt{m}}{\delta}\right) + k \log\left(\frac{1}{\delta}\right).$$

Lemma 5 (Approximation error of the Bayes predictor by a uniform-attention Transformer)**.** *Let* $\mathcal{U} \subset \mathbb{R}^{d_{\text{eff}}}$ *and* $\mathcal{C} \subset \mathbb{R}^{d_{\text{feat}}}$ *be non-empty compact sets with* $\operatorname{diam}(\mathcal{U}) \le 1$*. For every* $k \in \mathbb{N}_+$*, consider a permutation-invariant map* $M_{\text{Bayes}} : \mathcal{Z}^k \to \mathbb{R}$ *on* $\mathcal{Z} := \mathcal{U} \times \mathcal{C}$ *satisfying the Hölder condition*

$$|M_{\text{Bayes}}(\boldsymbol{z}_{1:k}) - M_{\text{Bayes}}(\boldsymbol{z}'_{1:k})| \le \; L \frac{1}{k} \sum_{i=1}^{k} \|\boldsymbol{z}_i - \boldsymbol{z}'_i\|_2^\alpha, \quad \alpha \in (0, 1], \; \boldsymbol{z}_i = (\boldsymbol{u}_i, \boldsymbol{c}), \; \boldsymbol{z}'_i = (\boldsymbol{u}'_i, \boldsymbol{c}').$$

Then, for any $\eta \in (0, e^{-1})$*, there exists an integer* $m \asymp \eta^{-d_{\text{eff}}/\alpha}$ *and a* C^∞ *partition of unity* $\phi = (\phi_1, \ldots, \phi_m) : \mathcal{U} \to [0, 1]^m$ *with* $\sum_{j=1}^m \phi_j \equiv 1$ *such that, writing* $s(\boldsymbol{u}_{1:k}) := \frac{1}{k} \sum_{i=1}^k \phi(\boldsymbol{u}_i) \in \Delta^{m-1}$*, one can construct a (clipped) ReLU decoder* $\rho_\theta : \Delta^{m-1} \times \mathcal{C} \to \mathbb{R}$ *so that*

$$\sup_{c \in \mathcal{C}} \sup_{\boldsymbol{u}_{1:k} \in \mathcal{U}^k} |M_{\text{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) - \rho_\theta(s(\boldsymbol{u}_{1:k}), \boldsymbol{c})| \le C(d_{\text{eff}}) L \eta.$$

Furthermore, ρ_θ *is uniformly Lipschitz and bounded with respect to* $(\boldsymbol{s}, \boldsymbol{c})$*, and the layer-wise spectral product can be controlled as follows:*

$$\left|\rho_\theta(\boldsymbol{s}, \boldsymbol{c}) - \rho_\theta(\boldsymbol{s}', \boldsymbol{c}')\right| \le L_s \|\boldsymbol{s} - \boldsymbol{s}'\|_2 + L_c \|\boldsymbol{c} - \boldsymbol{c}'\|_2, \qquad |\rho_\theta| \le B_M,$$

$$L_s \le CL\sqrt{m}, \qquad L_c \le CLm^{(1-\alpha)/d_{\text{eff}}} \le CL\sqrt{m}, \qquad S(\rho_\theta) \le CL\sqrt{m}.$$

In addition, ϕ *can be uniformly approximated by a ReLU network with* $O(\log m)$*-layers and* $O(m \log m)$*-weights, and its implementation satisfies* $\sum_j (\phi_\theta)_j \equiv 1$*,* $(\phi_\theta)_j \in [0, 1]$*, with the spectral product satisfying* $S(\phi_\theta) \le C_\phi m^{1/d_{\text{eff}}}$.

This lemma guarantees that the uniform-attention Transformer we are analyzing has the capacity to adequately represent smooth Bayesian predictors. This yields a fixed-length, permutation-invariant representation independent of context length p with provable approximation rates that feed directly into the sequential generalization analysis.

Lemma 6 (Oracle inequality for R_{BG})**.** *Let* $\mathcal{D}_{\text{train}} = \left\{\{(P_j^k, y_{j,k+1})\}_{k=1}^p\right\}_{j=1}^N$ *be draws from the prompt-generating process in Definition 1. Let* $M_{\hat{\theta}}$ *be the ERM* (1) *of the Transformer (Definition 2). Suppose Assumptions 1–2 hold. If* $\inf_{\theta \in \Theta} R_{\text{BG}}(M_\theta) = O(\frac{1}{N}(\frac{m}{p} + 1))$,

$$\mathbb{E} R_{\text{BG}}(M_{\hat{\theta}}) \lesssim \inf_{\theta \in \Theta} R_{\text{BG}}(M_\theta) + \frac{m}{pN} \operatorname{polylog}(pN) + \frac{1}{N} \operatorname{polylog}(pN),$$

where $\operatorname{polylog}(pN)$ *denotes a factor that is a polynomial in* $\log pN$*, the expectation is taken with respect to* $\mathcal{D}_{\text{train}}$ *and* $\mathcal{M} := \{M_\theta : \theta \in \Theta\}$.

The generalization error is $\tilde{O}(\frac{m}{pN}) + N^{-1}$. Here, m represents the complexity of $\mathcal{M}$, and increasing m improves the approximation ability (Lemma 5), but also increases the variance, which appears here.

H Proofs of the Main Results

Proof of Theorem 1. Let $R_k(M) := \mathbb{E}_{i=I\sim\mathcal{P}_I, f\sim\mathcal{P}_{F_i}, D^k\sim\mathcal{P}_{X,Y|f}^{\otimes k}, \boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[(f(\boldsymbol{x}_{k+1}) - M(P^k))^2\right]$. Then, $R(M) = \frac{1}{p}\sum_{k=1}^{p} R_k(M)$.

For any k-context D^k and query $\boldsymbol{x}_{k+1}$, define $M_{\text{Bayes}}(P^k) = \mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$. By simple algebra:

$$\begin{aligned}
R_k(M) &= \mathbb{E}_{I,f,P^k}[(f(\boldsymbol{x}_{k+1}) - M(P^k))^2] \\
&= \mathbb{E}_{I,f,P^k}[(f(\boldsymbol{x}_{k+1}) - M_{\text{Bayes}}(P^k))^2] + \mathbb{E}_{I,f,P^k}[(M_{\text{Bayes}}(P^k) - M(P^k))^2] \\
&\quad + 2\mathbb{E}_{I,f,P^k}[(f(\boldsymbol{x}_{k+1}) - M_{\text{Bayes}}(P^k))(M_{\text{Bayes}}(P^k) - M(P^k))]. \qquad (5)
\end{aligned}$$

Let $\mathcal{G}'_k$ be the σ-algebra generated by $(D^k, \boldsymbol{x}_{k+1})$. Since f is almost surely finite and $(M_{\text{Bayes}}(P^k) - M(P^k))$ is $\mathcal{G}'_k$-measurable, by the tower property of conditional expectation:

$$\begin{aligned}
&\mathbb{E}_{I,f,P^k}[(f(\boldsymbol{x}_{k+1}) - M_{\text{Bayes}}(P^k))(M_{\text{Bayes}}(P^k) - M(P^k))] \\
&= \mathbb{E}_{I,f,P^k}\left[(M_{\text{Bayes}}(P^k) - M(P^k))\mathbb{E}_{I,f}[f(\boldsymbol{x}_{k+1}) - M_{\text{Bayes}}(P^k) \mid \mathcal{G}'_k]\right].
\end{aligned}$$

$M_{\text{Bayes}}(P^k) = \mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})]$ implies the inner expectation equals zero.

From (5) with vanishing cross-term, $R_k(M)$ is decomposed as $R_k(M) = R_{\text{PV},k} + R_{\text{BG},k}(M)$, where

$$\begin{aligned}
R_{\text{PV},k} &:= \mathbb{E}_{I,f,P^k}[(f(\boldsymbol{x}_{k+1}) - M_{\text{Bayes}}(P^k))^2] \\
&= \mathbb{E}_{P^k}[\mathbb{E}_{f\sim\mathcal{P}(f|D^k)}(f(\boldsymbol{x}_{k+1}) - M_{\text{Bayes}}(P^k))^2] \\
&= \mathbb{E}_{D^k,\boldsymbol{x}_{k+1}}[\text{Var}_{f\sim P(f|D^k)}(f(\boldsymbol{x}_{k+1}))],
\end{aligned}$$

and

$$\begin{aligned}
R_{\text{BG},k}(M) &:= \mathbb{E}_{P^k}[\{M_{\text{Bayes}}(P^k) - M(P^k)\}^2] \\
&= \mathbb{E}_{P^k}[\{\mathbb{E}_{f\sim\mathcal{P}(f|D^k)}[f(\boldsymbol{x}_{k+1})] - M(P^k)\}^2].
\end{aligned}$$

Hence,

$$R(M) = \frac{1}{p}\sum_{k=1}^{p} R_{\text{PV},k} + \frac{1}{p}\sum_{k=1}^{p} R_{\text{BG},k}(M) = R_{\text{PV}} + R_{\text{BG}}(M).$$

□

Proof of Theorem 2. **Step 0 (clipping via a high-probability event).** Let $t_\varepsilon := \sigma_\varepsilon\sqrt{2\log(4p/\delta)}$ for $\delta \in (0, e^{-1})$, and define

$$\mathcal{E} := \left\{\max_{1\le i\le p+1} |\varepsilon_i| \le t_\varepsilon\right\}.$$

By sub-Gaussian tails and a union bound, $\Pr(\mathcal{E}^c) \le \delta$. On $\mathcal{E}$, writing $\boldsymbol{z}_i := (\boldsymbol{x}_i, y_i, \boldsymbol{x}_{k+1})$, we have $\boldsymbol{z}_i \in B(0, R_{rad})$ with radius $R_{rad} := C(B_X + B_f + t_\varepsilon)$, hence $\boldsymbol{z}_i \in \mathcal{Z}_R := B(0, R_{rad}) \subset \mathbb{R}^{2d_{\text{feat}}+1}$ (compact). Rescale $\tilde{\boldsymbol{z}} := \boldsymbol{z}/(2R_{rad})$ so that $\text{diam}(\tilde{\mathcal{Z}}_R) \le 1$.

Step 1 (approximation & aggregation noise on $\mathcal{E}$). Apply Lemma 5 with the shared variable $\boldsymbol{c} := \boldsymbol{x}_{k+1}$ and $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i, \boldsymbol{x}_{k+1})$. With grid scale $\eta \asymp m^{-1/d_{\text{eff}}}$, on $\mathcal{E}$, squared error is

$$C_1(2R_{rad})^{2\alpha}\eta^{2\alpha}.$$

Since $R_{rad} \lesssim \sqrt{\log(p/\delta)}$, the factor $(2R_{rad})^{2\alpha}$ is polylogarithmic and is absorbed into $\tilde{O}()$. Choosing $\eta \asymp m^{-1/d_{\text{eff}}}$ gives $m^{-2\alpha/d_{\text{eff}}}$ up to polylogarithmic factors.

Step 2 (estimation error and combination). From Lemma 6, the estimation term $\tilde{O}\left(\frac{m}{pN} + \frac{1}{N}\right)$. Combining with Step 1 gives

$$m^{-\frac{2\alpha}{d_{\text{eff}}}} + \frac{m}{pN} + \frac{1}{N}.$$

Optimizing over m yields the displayed rate (polylog factors absorbed into $\tilde{O}$).

Step 3 (contribution of $\mathcal{E}^c$). As in Step 7 of Lemma 6, using $(B_f{+}B_M)^2 + \sigma_\varepsilon^2$ as an envelope and sub-Gaussian tails, the contribution on $\mathcal{E}^c$ is $O\big(\delta + \delta\log(p/\delta)\big)$. With $\delta := (pN)^{-2}$, this is negligible compared to the main terms. □

Proof of Theorem 3. Recall that $D^k = (\boldsymbol{x}_1, y_1, \dots, \boldsymbol{x}_k, y_k)$. By the chain rule and the definition of $Z_{j,t}$,

$$\frac{p_j(D^k)}{p_{i^\star}(D^k)} = \prod_{t=1}^{k} \frac{p_j(\boldsymbol{x}_t, y_t \mid D^{t-1})}{p_{i^\star}(\boldsymbol{x}_t, y_t \mid D^{t-1})} = \prod_{t=1}^{k} \frac{p_j(y_t \mid \boldsymbol{x}_t, D^{t-1})}{p_{i^\star}(y_t \mid \boldsymbol{x}_t, D^{t-1})} = \exp\Big(\sum_{t=1}^{k} Z_{j,t}\Big). \tag{6}$$

Write $\pi_i(D^k) := \Pr(I = i \mid D^k)$ and $\mu_i(\boldsymbol{x}) := \mathbb{E}\left[f(\boldsymbol{x}) \mid I = i, D^k\right]$. By the law of total variance conditioning on I,

$$\mathrm{Var}\left(f(\boldsymbol{x}) \mid D^k\right) = \underbrace{\mathbb{E}_{I|D^k}\left[\mathrm{Var}\left(f(\boldsymbol{x}) \mid I, D^k\right)\right]}_{\text{(A)}} + \underbrace{\mathrm{Var}_{I\sim\mathcal{P}_{I|D^k}}\left(\mu_I(\boldsymbol{x})\right)}_{\text{(B)}}. \tag{7}$$

We compare the right-hand side with $\mathrm{Var}\left(f(\boldsymbol{x}) \mid I = i^\star, D^k\right)$.

Step 1 (term (A)**).** Using $|f(\boldsymbol{x})| \le B_f$,

$$\begin{aligned}\left|\mathbb{E}_{I|D^k}\left[\mathrm{Var}(f(\boldsymbol{x}) \mid I, D^k)\right] - \mathrm{Var}\left(f(\boldsymbol{x}) \mid I = i^\star, D^k\right)\right| &\le \sum_{j\ne i^\star} \pi_j(D^k)\left|\mathrm{Var}_j - \mathrm{Var}_{i^\star}\right| \\ &\le B_f^2 \sum_{j\ne i^\star} \pi_j(D^k),\end{aligned}$$

where $\mathrm{Var}_j := \mathrm{Var}(f(\boldsymbol{x}) \mid I = j, D^k) \le B_f^2$.

Step 2 (term (B)**).** For any $\mathcal{G}'_k$-measurable scalar a, $\mathrm{Var}_{I\sim\mathcal{P}_{I|D^k}}(\mu_I) \le \mathbb{E}_{I|D^k}(\mu_I - a)^2$. Choosing $a = \mu_{i^\star}(\boldsymbol{x})$ and using $|\mu_i(\boldsymbol{x})| \le B_f$,

$$\mathrm{Var}_{I\sim\mathcal{P}_{I|D^k}}\left(\mu_I(\boldsymbol{x})\right) \le \sum_j \pi_j(D^k)\left(\mu_j(\boldsymbol{x}) - \mu_{i^\star}(\boldsymbol{x})\right)^2 \le 4B_f^2 \sum_{j\ne i^\star} \pi_j(D^k).$$

Combining the two steps with (7),

$$\mathrm{Var}\left(f(\boldsymbol{x}) \mid D^k\right) \le \mathrm{Var}\left(f(\boldsymbol{x}) \mid I = i^\star, D^k\right) + 5B_f^2 \sum_{j\ne i^\star} \pi_j(D^k).$$

Taking $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_X}$ and then $\mathbb{E}_{D^k|I=i^\star}$ yields

$$\begin{aligned}&\mathbb{E}_{D^k,\boldsymbol{x}|I=i^\star}\left[\mathrm{Var}_{f|D^k}\{f(\boldsymbol{x})\}\right] \\ &\le \mathbb{E}_{D^k,\boldsymbol{x}|I=i^\star}\Big[\mathrm{Var}\left(f(\boldsymbol{x}) \mid I = i^\star, D^k\right)\Big] + 5B_f^2\,\mathbb{E}_{D^k|I=i^\star}\left[1 - \pi_{i^\star}(D^k)\right].\end{aligned} \tag{8}$$

Step 3 (posterior concentration of the task index). Let $S_{j,k} := \sum_{t=1}^{k} Z_{j,t}$ and $\lambda_{j,k} := e^{S_{j,k}}$. By the assumption, $\{Z_{j,t} + D_j\}$ are conditionally sub-exponential supermartingale differences. Applying a Bernstein-type supermartingale inequality (Theorem 2.6 in Fan et al., 2015), for each $j \ne i^\star$,

$$\Pr\left(S_{j,k} + kD_j \ge \tfrac{1}{2}kD_j \mid I = i^\star\right) \le e^{-C_j k}, \qquad C_j := \frac{D_j^2}{8(\nu_j^2 + b_j D_j/2)}.$$

Hence, by a union bound, there is an event $\mathcal{E}_k := \left\{\lambda_{j,k} \le e^{-D_j k/2} \forall j \ne i^\star\right\}$ with $\Pr(\mathcal{E}_k) \ge 1 - (T-1)e^{-Ck}$, where $C := \min_{j\ne i^\star} C_j$. On $\mathcal{E}_k$, using (6),

$$S_k := \sum_{j\ne i^\star} \frac{\alpha_j}{\alpha_{i^\star}} \lambda_{j,k} \le \frac{1-\alpha_{i^\star}}{\alpha_{i^\star}} e^{-D_{\min}k/2}, \qquad \pi_{i^\star}(D^k) = \frac{1}{1+S_k} \ge 1 - S_k.$$

Hence $1 - \pi_{i^\star}(D^k) \le S_k$ on $\mathcal{E}_k$, while trivially $1 - \pi_{i^\star}(D^k) \le 1$ on $\mathcal{E}_k^{\mathsf{c}}$. Therefore

$$\mathbb{E}_{D^k|i^\star}\left[1 - \pi_{i^\star}(D^k)\right] \le \frac{1-\alpha_{i^\star}}{\alpha_{i^\star}} e^{-D_{\min} k/2} + (T-1)e^{-Ck}.$$

Step 4 (conclusion). Plug the last inequality into (8) to obtain the displayed bound for $\mathbb{E}_{D^k,\boldsymbol{x}|I=i^\star}\left[\mathrm{Var}_{f|D^k}\{f(\boldsymbol{x})\}\right]$. Finally, apply Lemma 3 to bound $\mathbb{E}_{D^k,\boldsymbol{x}|I=i^\star}\Big[\mathrm{Var}\left(f(\boldsymbol{x}) \mid I = i^\star, D^k\right)\Big]$ by $\inf_M \sup_{f\in F_{i^\star}} \mathbb{E}_{P^k}\left[\left(f(\boldsymbol{x}_{k+1}) - M(P^k)\right)^2 \mid f\right]$. □

I Proofs of the Technical Lemmas

Proof of Lemma 3. Define the MSE at step k under f, $r_k(M,f) := \mathbb{E}_{P^k}\left[\left(f(\boldsymbol{x}_{k+1}) - M(P^k)\right)^2 \mid f\right]$, and the minimax risk at step k for the true task type $R_k^\star(F_{i^\star}) := \inf_M \sup_{f\in F_{i^\star}} r_k(M,f)$. For any fixed M and any measure Π supported on $F_{i^\star}$,

$$\sup_{f\in F_{i^\star}} r_k(M,f) \ge \int r_k(M,f)\mathrm{d}\Pi(f).$$

Taking $\Pi = \mathcal{P}_{F_{i^\star}}$ and then infimum over M,

$$R_k^\star(F_{i^\star}) \ge \inf_M \int r_k(M,f)\mathrm{d}\mathcal{P}_{F_{i^\star}}(f).$$

By Tonelli's theorem and the tower property,

$$\int r_k(M,f)\mathrm{d}\mathcal{P}_{F_{i^\star}}(f) = \mathbb{E}_{D^k,\boldsymbol{x}_{k+1}|I=i^\star}\Big[\mathbb{E}_{f\sim\mathcal{P}_{f|I=i^\star,D^k}}\left[\left(f(\boldsymbol{x}_{k+1}) - M(P^k)\right)^2\right]\Big].$$

Since $M(P^k)$ is $\mathcal{G}_k'$-measurable, the inner expectation is minimized pointwise (for each realized $D^k, \boldsymbol{x}_{k+1}$) by the posterior mean $\mathbb{E}\left[f(\boldsymbol{x}_{k+1}) \mid I = i^\star, D^k, \boldsymbol{x}_{k+1}\right]$, and its minimum value is $\mathrm{Var}\left(f(\boldsymbol{x}_{k+1}) \mid I = i^\star, D^k\right)$. Therefore

$$\inf_M \int r_k(M,f)\mathrm{d}\mathcal{P}_{F_{i^\star}}(f) = \mathbb{E}_{f\sim\mathcal{P}_{F_{i^\star}}} \mathbb{E}_{D^k\sim\mathcal{P}_{X,Y|f}^{\otimes k}} \mathbb{E}_{\boldsymbol{x}_{k+1}\sim\mathcal{P}_X}\left[\mathrm{Var}_{f\sim\mathcal{P}_{F_{i^\star}|D^k}}\left(f(\boldsymbol{x}_{k+1})\right)\right]$$

which proves the claim. □

Proof of Lemma 4. Step 1: Contraction via triangle inequality. Let $\mathcal{S} := \{P^k \mapsto \frac{1}{k}\sum_{i=1}^k \phi_\theta(\boldsymbol{u}_i)\}$ and $\mathcal{R} := \{(\boldsymbol{s},\boldsymbol{c}) \mapsto \rho_\theta(\boldsymbol{s},\boldsymbol{c}) : (\boldsymbol{s},\boldsymbol{c}) \in \Delta^{m-1} \times \mathcal{C}\}$. For any two predictors $\rho_\theta \circ S_\theta$ and $\rho_{\theta'} \circ S_{\theta'}$ evaluated along a predictable tree z, the (L_s, L_c)–Lipschitz property of ρ_θ in $\boldsymbol{s}$ and the triangle inequality give

$$\begin{aligned}&\left|\rho_\theta(S_\theta(P^t),\boldsymbol{c}_t) - \rho_{\theta'}(S_{\theta'}(P^t),\boldsymbol{c}_t)\right|\\ &\le L_s\left\|S_\theta(P^t) - S_{\theta'}(P^t)\right\|_2 + \left|\rho_\theta(S_{\theta'}(P^t),\boldsymbol{c}_t) - \rho_{\theta'}(S_{\theta'}(P^t),\boldsymbol{c}_t)\right|.\end{aligned}$$

Consequently, a $(\delta/(2L_s))$–cover of the pooled-feature class $\mathcal{S}$ together with a $(\delta/2)$–cover of the decoder outputs $\mathcal{R}$ produces a δ–cover of the composite class $\{\rho_\theta \circ S_\theta\}$ under the ℓ_2 sequential metric. Equivalently,

$$\sup_z \log N_2^{\mathrm{seq}}\left(\delta, \{\rho_\theta \circ S_\theta\}; z\right) \le \sup_z \log N_2^{\mathrm{seq}}\left(\tfrac{\delta}{2L_s}, \mathcal{S}; z\right) + \sup_z \log N_2^{\mathrm{seq}}\left(\tfrac{\delta}{2}, \mathcal{R}; z\right).$$

This single reduction step subsumes the earlier contraction and triangle-inequality arguments and will be followed by separate bounds for $\mathcal{S}$ (Step 2) and $\mathcal{R}$ (Step 3).

Step 2: Cover of the pooled features $\mathcal{S}$. Let $\Phi = \{\phi_\theta : \theta \in \Theta\}$. For any $\theta, \theta' \in \Theta$ and any prompt P^k,

$$\|S_\theta(P^k) - S_{\theta'}(P^k)\|_2 = \left\|\frac{1}{k}\sum_{i=1}^k (\phi_\theta(\boldsymbol{u}_i) - \phi_{\theta'}(\boldsymbol{u}_i))\right\|_2 \le \sup_{\boldsymbol{u}\in\mathcal{U}} \|\phi_\theta(\boldsymbol{u}) - \phi_{\theta'}(\boldsymbol{u})\|_2.$$

Fix $\eta \in (0, 1)$ and set $r := \eta/(4L_\phi)$, where $L_\phi := \mathrm{Lip}(\phi_\theta)$. Take an r-net $\mathcal{N} \subset \mathcal{U}$ of input space of ϕ_θ with

$$|\mathcal{N}| \;\le\; C(d_{\mathrm{eff}}) \left(\tfrac{\mathrm{diam}(\mathcal{U})}{r}\right)^{d_{\mathrm{eff}}} = C(d_{\mathrm{eff}}) \left(\tfrac{4L_\phi \,\mathrm{diam}(\mathcal{U})}{\eta}\right)^{d_{\mathrm{eff}}}.$$

By triangle inequality and Lipschitzness, for every $\boldsymbol{u} \in \mathcal{U}$, there exists $\boldsymbol{u}' \in \mathcal{N}$ such that

$$\|\phi_\theta(\boldsymbol{u}) - \phi_{\theta'}(\boldsymbol{u})\|_2 \le \|\phi_\theta(\boldsymbol{u}') - \phi_{\theta'}(\boldsymbol{u}')\|_2 + 2L_\phi\, r \le \|\phi_\theta(\boldsymbol{u}') - \phi_{\theta'}(\boldsymbol{u}')\|_2 + \eta/2.$$

Hence a cover of $\{\phi_\theta(\cdot)\}$ on $\mathcal{N}$ at scale $\eta/2$ yields a uniform cover on $\mathcal{U}$ at scale η.

Note that $\log N_{\infty,2}(\eta, \Phi; \mathcal{N}) \;\le\; \sum_{j=1}^m \log N_\infty\left(\frac{\eta}{\sqrt{m}}, \Phi_j; \mathcal{N}\right) \;\le\; \sum_{j=1}^m \mathrm{Pdim}(\Phi_j) \log \frac{C|\mathcal{N}|\sqrt{m}}{\eta}$. From Anthony & Bartlett (1999); Bartlett et al. (2019), using $\mathrm{Pdim}(\Phi) = \tilde{O}(m)$ for the coordinate-wise $[0, 1]$-bounded ReLU features, the finite-set (size $|\mathcal{N}|$) covering bound gives

$$\begin{aligned}\log N_{\infty,2}\left(\tfrac{\eta}{2}, \Phi; \mathcal{N}\right) &\lesssim \mathrm{Pdim}(\Phi) \left[\log\left(\tfrac{C\sqrt{m}}{\eta}\right) + d_{\mathrm{eff}} \log\left(\tfrac{C' L_\phi \mathrm{diam}(\mathcal{U})}{\eta}\right)\right] \\ &\lesssim m \left[\log\left(\tfrac{C\sqrt{m}}{\eta}\right) + d_{\mathrm{eff}} \log\left(\tfrac{C' L_\phi \mathrm{diam}(\mathcal{U})}{\eta}\right)\right].\end{aligned}$$

Substituting $\eta = \delta/(2L_s)$ from Step 1 yields the sequential bound

$$\sup_z \log N_2^{\mathrm{seq}}\left(\tfrac{\delta}{2L_s}, \mathcal{S}; z\right) \lesssim m \left[\log\left(\tfrac{\tilde{C} L_s \sqrt{m}}{\delta}\right) + d_{\mathrm{eff}} \log\left(\tfrac{\tilde{C}' L_\phi \mathrm{diam}(\mathcal{U}) L_s}{\delta}\right)\right],$$

uniformly in z.

Step 3: Uniform cover of the decoder $\mathcal{R}$. Fix a predictable input tree $z = \{(s_t(\xi_{1:t-1}), c_t(\xi_{1:t-1}))\}_{t\le k}$ with nodes in $\Delta^{m-1} \times \mathcal{C}$. Fix $\delta \in (0, 2B_M]$ and build a uniform grid on the output range:

$$\mathcal{G} := \{-B_M, -B_M + \delta, -B_M + 2\delta, \ldots, -B_M + J\delta\}, \qquad J := \left\lceil \tfrac{2B_M}{\delta} \right\rceil,$$

so that for any $y \in [-B_M, B_M]$ there exists $q(y) \in \mathcal{G}$ with $|y - q(y)| \le \delta/2$. Now consider the family $\mathcal{V}$ of depth-wise constant predictable trees $v = \{v_t\}_{t\le k}$ defined by choosing, independently for each depth t, a grid value $g_t \in \mathcal{G}$ and setting $v_t(\cdot) \equiv g_t$ (constant on all nodes at depth t). Then $|\mathcal{V}| = |\mathcal{G}|^k = (J + 1)^k$.

Fix any decoder $\rho_\theta \in \mathcal{R}$ and any path $\xi \in \{\pm 1\}^k$. Along this path, we observe the length-k sequence of decoder outputs $y_t := \rho_\theta\left(s_t(\xi_{1:t-1}), c_t(\xi_{1:t-1})\right) \in [-B_M, B_M]$. Define the depth-wise grid sequence $g_t := q(y_t) \in \mathcal{G}$ and take the corresponding $v^\star \in \mathcal{V}$ with $v_t^\star(\cdot) \equiv g_t$. Then, along the path ξ,

$$\frac{1}{k} \sum_{t=1}^k \left(v_t^\star(\xi_{1:t-1}) - y_t\right)^2 \le \frac{1}{k} \sum_{t=1}^k \left(\frac{\delta}{2}\right)^2 = \left(\frac{\delta}{2}\right)^2,$$

that is, $d_{2,\xi}\left(\rho_\theta \circ z, v^\star; z\right) \le \delta/2$. Since this holds for every ρ_θ and every path ξ, the set $\mathcal{V}$ is a sequential $(\delta/2)$–cover of $\mathcal{R}$ on z. Therefore,

$$N_2^{\mathrm{seq}}\left(\tfrac{\delta}{2}, \mathcal{R}; z\right) \le |\mathcal{V}| = (J + 1)^k \le \left(\tfrac{2B_M}{\delta} + 2\right)^k.$$

Taking logarithms yields

$$\sup_z \log N_2^{\mathrm{seq}}\left(\tfrac{\delta}{2}, \mathcal{R}; z\right) \le k \log\left(\tfrac{2B_M}{\delta} + 2\right) \lesssim k \log\left(\tfrac{CB_M}{\delta}\right).$$

□

Proof of Lemma 5. We will write $C, C(d), \ldots$ for positive constants depending only on displayed arguments. Note that, w.r.t. ℓ_2, the renormalization layer with parameter τ has Lipschitz constant $L_{\text{renorm}} \le \frac{2\sqrt{m}}{\tau}$. Since ReLU is 1-Lipschitz and biases do not affect Lipschitz constants, the global Lipschitz modulus satisfies $\mathrm{Lip}(\mathcal{T}_\theta) \le S(\mathcal{T}_\theta)$ for ReLU network $\mathcal{T}_\theta$.

Step 1 (feature map: soft histogram). Fix

$$\delta := \left(\frac{\eta}{8\sqrt{d_{\mathrm{eff}}}}\right)^{1/\alpha} \in (0, 1), \qquad r := \delta/4.$$

Let $U \supset \mathcal{U}$ be an axis-aligned cube with $\mathrm{dist}(\mathcal{U}, \partial U) \geq r$, where $\mathrm{dist}(\mathcal{U}, \partial U) := \inf\{\|\boldsymbol{u} - \boldsymbol{u}'\| : \boldsymbol{u} \in \mathcal{U}, \boldsymbol{u}' \in \partial U\}$ denotes the Euclidean distance between $\mathcal{U}$ and the boundary of U. Partition U into a regular grid of closed cubes $\{Q_j\}_{j=1}^m$ of side length δ, so that $m \asymp \delta^{-d_{\mathrm{eff}}}$; denote by $\boldsymbol{q}_j$ the center of Q_j and set the representative point

$$\boldsymbol{r}_j \in \arg\min_{\boldsymbol{u} \in \mathcal{U}} \|\boldsymbol{u} - \boldsymbol{q}_j\|_2.$$

Let $\eta \in C_c^\infty(\mathbb{R}^{d_{\mathrm{eff}}})$ be a nonnegative and radially symmetric mollifier with $\int \eta = 1$ and $\mathrm{supp}\eta \subset B(0,1)$. Put $\eta_r(\boldsymbol{x}) := r^{-d_{\mathrm{eff}}}\eta(\boldsymbol{x}/r)$ and define

$$\phi_j(\boldsymbol{x}) := (\mathbf{1}_{Q_j} * \eta_r)(\boldsymbol{x}).$$

Then $\mathrm{supp}\phi_j \subset Q_j^+ := \{\boldsymbol{q} : \mathrm{dist}(\boldsymbol{q}, Q_j) \leq r\}$. Since the pairwise intersections of the grid cells have Lebesgue measure zero, we have $\sum_j \mathbf{1}_{Q_j} = \mathbf{1}_U$ almost everywhere, and because $B(\boldsymbol{x}, r) \subset U$ for all $\boldsymbol{x} \in \mathcal{U}$, convolution with the unit-mass mollifier ignores these measure-zero discrepancies, yielding $\sum_j \phi_j(\boldsymbol{x}) = (\sum_j \mathbf{1}_{Q_j}) * \eta_r(\boldsymbol{x}) = \mathbf{1}_U * \eta_r(\boldsymbol{x}) = 1$ pointwise on $\mathcal{U}$. Also, by Young's inequality, $\|\nabla\phi_j\|_\infty \leq \|\mathbf{1}_{Q_j}\|_\infty \|\nabla\eta_r\|_1 = \|\nabla\eta\|_1\, r^{-1}$. Since $r = \delta/4$, we get $\|\nabla\phi_j\|_\infty \leq (4\|\nabla\eta\|_1)\,\delta^{-1} =: C\delta^{-1}$, uniformly in j. For $\boldsymbol{u}_{1:k} \in \mathcal{U}^k$, define the soft histogram

$$s_j := \frac{1}{k}\sum_{i=1}^k \phi_j(\boldsymbol{u}_i), \qquad \boldsymbol{s} = (s_1, \ldots, s_m) \in \Delta^{m-1}.$$

Step 2 (decoder construction). For each fixed $\boldsymbol{c}$, define the ground cost on indices by

$$c^{(u)}(j, \ell) := \|\boldsymbol{r}_j - \boldsymbol{r}_\ell\|_2^\alpha, \qquad 0 < \alpha \leq 1,$$

and let $W_\alpha^{(u)}$ be the discrete 1-Wasserstein distance on the simplex $\Delta^{m-1} = \{\boldsymbol{s} \in [0,1]^m : \sum_j s_j = 1\}$ with cost $c^{(u)}$:

$$W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t}) := \min_{\pi \geq 0} \sum_{j,\ell} c^{(u)}(j,\ell)\pi_{j\ell} \quad \text{s.t.} \quad \sum_\ell \pi_{j\ell} = s_j, \sum_j \pi_{j\ell} = t_\ell.$$

where $\boldsymbol{s}, \boldsymbol{t} \in \Delta^{m-1}$. Note that $c^{(u)}$ is a metric since $0 < \alpha \leq 1$. Let $\Delta_k := \{\frac{\boldsymbol{n}}{k} : \boldsymbol{n} \in \{0, \ldots, k\}^m, \sum_j n_j = k\}$. For $\boldsymbol{v} = \boldsymbol{n}/k \in \Delta_k$, define

$$\rho_{\boldsymbol{c}}(\boldsymbol{v}) := M_{\mathrm{Bayes}}(\underbrace{(\boldsymbol{r}_1, \boldsymbol{c}), \ldots, (\boldsymbol{r}_1, \boldsymbol{c})}_{n_1}, \ldots, \underbrace{(\boldsymbol{r}_m, \boldsymbol{c}), \ldots, (\boldsymbol{r}_m, \boldsymbol{c})}_{n_m}).$$

This is well-defined by permutation invariance of M_{Bayes}.

Let $\boldsymbol{s} = \boldsymbol{n}/k$ and $\boldsymbol{t} = \boldsymbol{n}'/k$ be points of Δ_k. Construct an integer matrix $A = (A_{j\ell})$ with row sums $\boldsymbol{n}$ and column sums $\boldsymbol{n}'$ (e.g., by the Northwest corner rule (Peyré & Cuturi, 2019)), and set $\pi := A/k$. Then $\pi \in \Pi(s,t)$ is a feasible transport plan. Enumerating the k pairs so that $(\boldsymbol{r}_{j(i)}, \boldsymbol{r}_{\ell(i)})$ appears exactly $A_{j\ell}$ times, the Hölder condition yields

$$|\rho_{\boldsymbol{c}}(\boldsymbol{s}) - \rho_{\boldsymbol{c}}(\boldsymbol{t})| \leq \frac{L}{k}\sum_{i=1}^k \|\boldsymbol{r}_{j(i)} - \boldsymbol{r}_{\ell(i)}\|_2^\alpha = L\sum_{j,\ell} c^{(u)}(j,\ell)\,\frac{A_{j\ell}}{k} = L\sum_{j,\ell} c^{(u)}(j,\ell)\,\pi_{j\ell},$$

where $c^{(u)}(j,\ell) := \|\boldsymbol{r}_j - \boldsymbol{r}_\ell\|_2^\alpha$. Since this bound holds for $\pi^* \in \Pi(s,t)$,

$$|\rho_{\boldsymbol{c}}(s) - \rho_{\boldsymbol{c}}(t)| \leq L\, W_\alpha^{(u)}(s,t), \tag{9}$$

which proves the L-Lipschitz property on Δ_k.

Extend to all $\boldsymbol{s} \in \Delta^{m-1}$ by the McShane-type formula

$$\rho_{\boldsymbol{c}}^\star(\boldsymbol{s}) := \inf_{\boldsymbol{v} \in \Delta_k}\left\{\rho_{\boldsymbol{c}}(\boldsymbol{v}) + LW_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{v})\right\}, \tag{10}$$

which satisfies $\rho^\star_{\boldsymbol{c}}(\boldsymbol{v}) = \rho_{\boldsymbol{c}}(\boldsymbol{v})$ for $\boldsymbol{v} \in \Delta_k$ and, by the inequality (9), the Lipschitz property

$$|\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho^\star_{\boldsymbol{c}}(\boldsymbol{t})| \le L W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t}) \qquad (\forall \boldsymbol{s}, \boldsymbol{t}).$$

By this construction, $\rho^\star_{\boldsymbol{c}}(\boldsymbol{v}) = \rho_{\boldsymbol{c}}(\boldsymbol{v})$ holds. Indeed, for $\boldsymbol{v} \in \Delta_k$, taking $\boldsymbol{t} = \boldsymbol{v}$ in Eq. (10) gives $\rho^\star_{\boldsymbol{c}}(\boldsymbol{v}) \le \rho_{\boldsymbol{c}}(\boldsymbol{v})$. Conversely, the inequality (9) implies $\rho_{\boldsymbol{c}}(\boldsymbol{v}) \le \rho_{\boldsymbol{c}}(\boldsymbol{t}) + L W_\alpha^{(u)}(\boldsymbol{t}, \boldsymbol{v})$ for every $\boldsymbol{t} \in \Delta_k$, hence $\rho_{\boldsymbol{c}}(\boldsymbol{v}) \le \inf_t \{\rho_{\boldsymbol{c}}(\boldsymbol{t}) + L W_\alpha^{(u)}(\boldsymbol{v}, \boldsymbol{t})\} = \rho^\star_{\boldsymbol{c}}(\boldsymbol{v})$. Therefore $\rho^\star_{\boldsymbol{c}}(\boldsymbol{v}) = \rho_{\boldsymbol{c}}(\boldsymbol{v})$.

We next show its L-Lipschitzness. For any $\boldsymbol{s}, \boldsymbol{t}$ and any $\boldsymbol{v} \in \Delta_k$, the triangle inequality yields $W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{v}) \le W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t}) + W_\alpha^{(u)}(\boldsymbol{t}, \boldsymbol{v})$. Taking infima over $\boldsymbol{v}$, $\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) \le \rho^\star_{\boldsymbol{c}}(\boldsymbol{t}) + L W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t})$ and $\rho^\star_{\boldsymbol{c}}(\boldsymbol{t}) \le \rho^\star_{\boldsymbol{c}}(s) + L W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t})$, so $|\rho^\star_{\boldsymbol{c}}(s) - \rho^\star_{\boldsymbol{c}}(\boldsymbol{t})| \le L W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{t})$.

We also note its piecewise linearity. By the Kantorovich–Rubinstein dual (Peyré & Cuturi, 2019) on a finite space,

$$W_\alpha^{(u)}(\boldsymbol{s}, \boldsymbol{v}) = \sup_{\boldsymbol{\varphi} \in \mathbb{R}^m : |\varphi_j - \varphi_\ell| \le c^{(u)}(j,\ell)} \langle \varphi, \boldsymbol{s} - \boldsymbol{v} \rangle,$$

so $\boldsymbol{s} \mapsto \rho^\star_{\boldsymbol{c}}(\boldsymbol{s})$ is the lower envelope of finitely many support functions and thus piecewise linear on Δ^{m-1}.

Step 3 (error decomposition and bounds). Adopt a half-open tie-breaking so that each $\boldsymbol{u}_i$ belongs to a unique cell $Q_{j(i)}$. Let the hard histogram be $\boldsymbol{h} := \frac{1}{k}(n_1^{\mathrm{hard}}, \ldots, n_m^{\mathrm{hard}})$ with $n_j^{\mathrm{hard}} := \#\{i : \boldsymbol{u}_i \in Q_j\}$. Then, with $\boldsymbol{z}_i = (\boldsymbol{u}_i, \boldsymbol{c})$ and using the Hölder condition while keeping $\boldsymbol{c}$ fixed,

$$\begin{aligned} |M_{\mathrm{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) - \rho_\theta(\boldsymbol{s}, \boldsymbol{c})| \le & \underbrace{\left| M_{\mathrm{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) - M_{\mathrm{Bayes}}\left((\boldsymbol{r}_{j(1)}, \boldsymbol{c}), \ldots, (\boldsymbol{r}_{j(k)}, \boldsymbol{c})\right) \right|}_{\text{quantization in } u} \\ & + \underbrace{|\rho^\star_{\boldsymbol{c}}(\boldsymbol{h}) - \rho^\star_{\boldsymbol{c}}(\boldsymbol{s})|}_{\text{hard-to-soft transport}} + \underbrace{|\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho_\theta(\boldsymbol{s}, \boldsymbol{c})|}_{\text{network approximation}}. \end{aligned}$$

Quantization: $\|\boldsymbol{u}_i - \boldsymbol{r}_{j(i)}\|_2 \le \sqrt{d_{\mathrm{eff}}}\,\delta$, the Hölder condition gives

$$\left| M_{\mathrm{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) - M_{\mathrm{Bayes}}\left((\boldsymbol{r}_{j(1)}, \boldsymbol{c}), \ldots, (\boldsymbol{r}_{j(k)}, \boldsymbol{c})\right) \right| \le \frac{L}{k} \sum_{i=1}^k \|\boldsymbol{u}_i - \boldsymbol{r}_{j(i)}\|_2^\alpha \le C(d_{\mathrm{eff}}) L \delta^\alpha.$$

Moreover, $M_{\mathrm{Bayes}}\left((\boldsymbol{r}_{j(1)}, \boldsymbol{c}), \ldots, (\boldsymbol{r}_{j(k)}, \boldsymbol{c})\right) = \rho_{\boldsymbol{c}}(\boldsymbol{h}) = \rho^\star_{\boldsymbol{c}}(\boldsymbol{h})$.

Transport: Define a coupling π between $\boldsymbol{h}$ and $\boldsymbol{s}$ by moving, for each i, the mass $1/k$ placed at $\boldsymbol{r}_{j(i)}$ to the mixture $\sum_{j=1}^m \phi_j(\boldsymbol{u}_i) \delta_{\boldsymbol{r}_j}$:

$$\pi^{(i)}_{j(i) \to j} := \frac{1}{k} \phi_j(\boldsymbol{u}_i), \qquad \pi := \sum_{i=1}^k \sum_{j=1}^m \pi^{(i)}_{j(i) \to j}.$$

Because $\sum_j \phi_j \equiv 1$, π has marginals $\boldsymbol{h}$ and $\boldsymbol{s}$, hence is feasible for $W_\alpha^{(u)}$. If $\phi_j(u_i) > 0$ then $u_i \in Q_j^+$, and by the triangle inequality together with Step 1,

$$\|\boldsymbol{r}_{j(i)} - \boldsymbol{r}_j\|_2 \le \|\boldsymbol{r}_{j(i)} - \boldsymbol{u}_i\|_2 + \|\boldsymbol{u}_i - \boldsymbol{r}_j\|_2 \le C(d_{\mathrm{eff}}) \delta.$$

Therefore, with $W_\alpha^{(u)}$,

$$W_\alpha^{(u)}(\boldsymbol{h}, \boldsymbol{s}) \le \sum_{i=1}^k \sum_{j=1}^m \pi^{(i)}_{j(i) \to j} \|\boldsymbol{r}_{j(i)} - \boldsymbol{r}_j\|_2^\alpha \le C(d_{\mathrm{eff}}) \delta^\alpha,$$

and since $\rho^\star_{\boldsymbol{c}}$ is L-Lipschitz w.r.t. $W_\alpha^{(u)}$,

$$|\rho^\star_{\boldsymbol{c}}(\boldsymbol{h}) - \rho^\star_{\boldsymbol{c}}(\boldsymbol{s})| \le L W_\alpha^{(u)}(\boldsymbol{h}, \boldsymbol{s}) \le C(d_{\mathrm{eff}}) L \delta^\alpha.$$

Combining the three bounds and using $\mathrm{diam}(\mathcal{U}) \le 1$ (so that $c^{(u)}(j,\ell) \le 1$ and $W_\alpha^{(u)} \le \mathrm{TV} = \frac{1}{2}\|\cdot\|_1$), we obtain

$$|M_{\mathrm{Bayes}}(\boldsymbol{u}_{1:k}, \boldsymbol{c}) - \rho_\theta(\boldsymbol{s}, \boldsymbol{c})| \le C(d_{\mathrm{eff}}) L \delta^\alpha + |\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho_\theta(\boldsymbol{s}, \boldsymbol{c})|.$$

Finally choose ρ_θ so that $\sup_{(\boldsymbol{s},\boldsymbol{c})} |\rho^\star_{\boldsymbol{c}}(s) - \rho_\theta(\boldsymbol{s},\boldsymbol{c})| \le CL\delta^\alpha$ (Step 4(iii)). Then

$$\sup_{\boldsymbol{c}} \sup_{\boldsymbol{u}_{1:k}\in\mathcal{U}^k} |M_{\text{Bayes}}(\boldsymbol{u}_{1:k},\boldsymbol{c}) - \rho_\theta(\boldsymbol{s},\boldsymbol{c})| \le C(d_{\text{eff}})L\delta^\alpha.$$

Choosing $\delta \asymp \eta^{1/\alpha}$ and $m \asymp \delta^{-d_{\text{eff}}}$ yields the claimed bound $C(d_{\text{eff}})L\eta$.

Step 4 (Neural implementation). We first consider the joint regularity of $(\boldsymbol{s},\boldsymbol{c}) \mapsto \rho^\star_{\boldsymbol{c}}(\boldsymbol{s})$ on the compact domain $\Delta^{m-1} \times \mathcal{C}$.

(i) Joint Lipschitz in $(\boldsymbol{s},\boldsymbol{c})$. By Step 2, for each fixed $\boldsymbol{c}$ and all $\boldsymbol{s},\boldsymbol{s}' \in \Delta^{m-1}$,

$$|\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho^\star_{\boldsymbol{c}}(\boldsymbol{s}')| \le LW^{(u)}_\alpha(\boldsymbol{s},\boldsymbol{s}').$$

On the simplex, we have $W^{(u)}_\alpha(\boldsymbol{s},\boldsymbol{s}') \le \frac{\operatorname{diam}(\mathcal{U})^\alpha}{2}\|\boldsymbol{s}-\boldsymbol{s}'\|_1 \le \frac{\operatorname{diam}(\mathcal{U})^\alpha}{2}\sqrt{m}\|\boldsymbol{s}-\boldsymbol{s}'\|_2$: it follows from the trivial plan that transports the total variation mass across at most $\operatorname{diam}(\mathcal{U})^\alpha$. Since $\operatorname{diam}(\mathcal{U}) \le 1$,

$$|\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho^\star_{\boldsymbol{c}}(\boldsymbol{s}')| \le CL\sqrt{m}\|\boldsymbol{s}-\boldsymbol{s}'\|_2.$$

Next, fix $\boldsymbol{s}$ and vary $\boldsymbol{c},\boldsymbol{c}'$. From the Hölder assumption on M_{Bayes} applied to $\boldsymbol{z}_i = (\boldsymbol{r}_{j(i)},\boldsymbol{c})$ and $\boldsymbol{z}'_i = (\boldsymbol{r}_{j(i)},\boldsymbol{c}')$ we obtain $|\rho_{\boldsymbol{c}}(\boldsymbol{v}) - \rho_{\boldsymbol{c}'}(\boldsymbol{v})| \le L\|\boldsymbol{c}-\boldsymbol{c}'\|_2^\alpha$ for all $\boldsymbol{v} \in \Delta_k$. By the McShane envelope (10), $(\boldsymbol{s},\boldsymbol{c}) \mapsto \rho^\star_{\boldsymbol{c}}(\boldsymbol{s})$ is α-Hölder in $\boldsymbol{c}$:$|\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho^\star_{\boldsymbol{c}'}(\boldsymbol{s})| \le L\|\boldsymbol{c}-\boldsymbol{c}'\|_2^\alpha$. To meet the size of networks in Definition 2, we first apply a McShane-type α-Hölder extension to the whole space $\mathbb{R}^{d_{\text{feat}}}$, and then convolve only in the $\boldsymbol{c}$-direction with a standard mollifier (Appendix C.5 in Evans, 2010) η_h. This yields, for any $(\boldsymbol{s},\boldsymbol{c})$, $|\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho^\sharp_{\boldsymbol{c}}(\boldsymbol{s})| \le \left|\int \left(\rho^\star_{\boldsymbol{c}}(\boldsymbol{s}) - \rho^\star_{\boldsymbol{c}-h\boldsymbol{z}}(\boldsymbol{s})\right)\eta(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\right| \le \int L\,\|h\boldsymbol{z}\|^\alpha\,\eta(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \le C_\eta\,L\,h^\alpha$, and $\operatorname{Lip}_c(\rho^\sharp) \lesssim h^{\alpha-1}$ uniformly in $(\boldsymbol{s},\boldsymbol{c})$. In what follows we approximate $\rho^\sharp$ by a ReLU network and keep the same notation ρ_θ.

(ii) ReLU approximation of the feature map. As in the current proof, each $\boldsymbol{u} \mapsto \phi_j(\boldsymbol{u})$ is C^∞ on $[0,1]^{d_{\text{eff}}}$ with $\|\nabla\phi_j\|_\infty \lesssim \delta^{-1}$, hence by ReLU approximation (Yarotsky, 2017) there exists a ReLU network of depth $O(\log(1/\eta_\phi))$ and size $O(m\log(1/\eta_\phi))$ that uniformly approximates $\phi = (\phi_1,\dots,\phi_m)$ on $\mathcal{U}$ with error $\eta_\phi \in (0,e^{-1})$. Additionally, we set spectral product

$$S(\phi_\theta) \asymp \delta^{-1} = m^{1/d_{\text{eff}}},$$

matching size of the Transformer in Definition 2. After applying the fixed renormalization layer $\text{Renorm}_\tau : \mathbb{R}^m \to \Delta^{m-1}$, the features are simplex-valued

(iii) ReLU approximation of the decoder. On the compact set $\Delta^{m-1} \times \mathcal{C}$, the map $(\boldsymbol{s},\boldsymbol{c}) \mapsto \rho^\sharp_{\boldsymbol{c}}(\boldsymbol{s})$ is jointly Lipschitz with moduli (L_s, L_c) from (i), and for each fixed $\boldsymbol{c}$ it is piecewise-linear in $\boldsymbol{s}$ (lower envelope of affine forms by the KR dual). Therefore, by standard approximation results for Lipschitz targets on a compact domain (Yarotsky, 2017), there exists a ReLU network $\rho_\theta : \Delta^{m-1} \times \mathcal{C} \to \mathbb{R}$ such that

$$\sup_{(\boldsymbol{s},\boldsymbol{c})} |\rho^\sharp_{\boldsymbol{c}}(\boldsymbol{s}) - \rho_\theta(\boldsymbol{s},\boldsymbol{c})| \le CL\delta^\alpha.$$

Moreover, by spectral normalization of the linear layers, we can enforce

$$\operatorname{Lip}_s(\rho_\theta) \le cL_s = cCL\sqrt{m}, \qquad \operatorname{Lip}_{\boldsymbol{c}}(\rho_\theta) \le cL_c = cL\delta^{\alpha-1},$$

so the decoder's spectral product can be taken as

$$S(\rho_\theta) = O\big(L\sqrt{m} + L\delta^{\alpha-1}\big)$$

under the ℓ_2-metric used. Note that $\delta^{\alpha-1} = O(m^{(1-\alpha)/d_{\text{eff}}}) = O(\sqrt{m})$ as $d_{\text{eff}} \ge 2$. Note that the number of parameters of the decoder does not affect the upper bound of the predictive risk in Theorem 2. Instead, we evaluate the complexity regarding the decoder by counting the number of δ-cubes to cover the space of length-k sequences (see proof of Lemma 4, Step 3).

Finally, combining (ii)–(iii) with Step 4 and taking $\eta_\phi = 1/m$, we obtain

$$\sup_{\boldsymbol{c}} \sup_{\boldsymbol{u}_{1:k}\in\mathcal{U}^k} \left|M_{\text{Bayes}}(\boldsymbol{u}_{1:k},\boldsymbol{c}) - \rho_\theta\Big(\tfrac{1}{k}\sum_{i=1}^k \phi(\boldsymbol{u}_i),\boldsymbol{c}\Big)\right| \le C(d_{\text{eff}})L\delta^\alpha.$$

Choosing $\delta \asymp \eta^{1/\alpha}$ and $m \asymp \delta^{-d_{\text{eff}}}$ yields the lemma. □

Proof of Lemma 6. Recall that we work on standard Borel spaces (the Borel spaces associated with Polish spaces) so that regular conditional distributions (Durrett, 2019) exist. Accordingly, $\Pr(f \in \cdot \mid D^k)$ and the quantities $\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid D^k]$, $\mathrm{Var}(f(\boldsymbol{x}_{k+1}) \mid D^k)$ are well-defined.

A technical point concerns the measurability of suprema over the parameter space Θ, which is required for expectations to be well-defined. Note that under our assumptions, the parameter space Θ is separable and, for any fixed sample, $\theta \mapsto (y - M_\theta(P))^2$ is continuous, so the relevant random suprema are measurable.

Step 1 (Reduction via a centered, Bayes-offset objective). For each block j, write

$$\Lambda_j(\theta) := \frac{1}{p}\sum_{k=1}^{p} \left(y_{j,k+1} - M_\theta(P_j^k)\right)^2 = A_j(\theta) + B_j(\theta) + C_j,$$

with

$$A_j(\theta) := \frac{1}{p}\sum_{k=1}^{p} \left(M_{\text{Bayes}}(P_j^k) - M_\theta(P_j^k)\right)^2,$$

$$B_j(\theta) := \frac{2}{p}\sum_{k=1}^{p} \left(y_{j,k+1} - M_{\text{Bayes}}(P_j^k)\right)\left(M_{\text{Bayes}}(P_j^k) - M_\theta(P_j^k)\right),$$

and $C_j := \frac{1}{p}\sum_{k=1}^{p} \left(y_{j,k+1} - M_{\text{Bayes}}(P_j^k)\right)^2$, which does not depend on θ. Define the centered (Bayes-offset) empirical objective

$$\widehat{\mathcal{R}}(\theta) := \frac{1}{N}\sum_{j=1}^{N} \widetilde{\Lambda}_j(\theta), \qquad \widetilde{\Lambda}_j(\theta) := \Lambda_j(\theta) - C_j = A_j(\theta) + B_j(\theta).$$

Then $\arg\min_\theta \frac{1}{N}\sum_j \Lambda_j(\theta) = \arg\min_\theta \widehat{\mathcal{R}}(\theta)$, i.e., the ERM $\hat{\theta}$ is unchanged by the offset. Define the population counterpart $\mathcal{R}(\theta) := \mathbb{E}[\widetilde{\Lambda}_j(\theta)]$; using $\mathbb{E}[y - M_{\text{Bayes}}(P) \mid P] = 0$,

$$\mathcal{R}(\theta) = \mathbb{E}\left[(M_{\text{Bayes}}(P) - M_\theta(P))^2\right] = R_{\text{BG}}(M_\theta).$$

Let $\theta^\star \in \arg\min_\theta \mathcal{R}(\theta)$. Then

$$\begin{aligned} R_{\text{BG}}(M_{\hat{\theta}}) - R_{\text{BG}}(M_{\theta^\star}) &= \mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\star) \\ &= \mathcal{R}(\hat{\theta}) - \hat{\mathcal{R}}(\hat{\theta}) + \hat{\mathcal{R}}(\hat{\theta}) - \hat{\mathcal{R}}(\theta^\star) + \hat{\mathcal{R}}(\theta^\star) - \mathcal{R}(\theta^\star) \\ &\le \mathcal{R}(\hat{\theta}) - \hat{\mathcal{R}}(\hat{\theta}) + \hat{\mathcal{R}}(\theta^\star) - \mathcal{R}(\theta^\star) \end{aligned} \tag{11}$$

and hence

$$\mathbb{E}[R_{\text{BG}}(M_{\hat{\theta}}) - R_{\text{BG}}(M_{\theta^\star})] \le \mathbb{E}[\mathcal{R}(\hat{\theta}) - \hat{\mathcal{R}}(\hat{\theta})] \le \mathbb{E}\left[\sup_\theta |\mathcal{R}(\theta) - \hat{\mathcal{R}}(\theta)|\right].$$

Step 2 (Localization at worst–path sequential radius). Let $h_\theta := M_\theta - M_{\text{Bayes}}$. Fix a $\mathcal{Z}$–valued predictable tree $Z = (Z_k)_{k=1}^p$ of depth p that is decoupled tangent to the prompt process, in the sense that for each depth k and each past $\xi_{1:k-1} \in \{\pm 1\}^{k-1}$, the conditional distribution of $Z_k(\xi_{1:k-1})$ equals the conditional distribution of P^k given D^{k-1} (Peña & Giné, 1999; Rakhlin et al., 2015); namely $Z_k(\xi_{1:k-1}) \mid D^{k-1} \overset{d}{=} P^k \mid D^{k-1}$. Conditioning on a realization $Z = z$, we refer to such z as a data-containing tangent tree. For any realization z of Z, define the worst–path sequential ℓ_2 radius on z by

$$\|h\|_{\text{seq},2;z} := \left\{ \sup_{\xi \in \{\pm 1\}^p} \frac{1}{p} \sum_{k=1}^{p} h\left(z_k(\xi_{1:k-1})\right)^2 \right\}^{1/2}.$$

For $r > 0$, we localize by the uniform worst–path radius

$$\mathcal{H}(r) := \left\{ h_\theta = M_\theta - M_{\text{Bayes}} : \sup_z \|h_\theta\|_{\text{seq},2;z} \le r \right\}.$$

Then, for any θ such that $h_\theta \in \mathcal{H}(r)$, since h_θ is a bounded measurable function,

$$R_{\mathrm{BG}}(M_\theta) = \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}_{P^k}\left[h_\theta\left(P^k\right)^2\right] = \mathbb{E}_{Z,\xi}\left[\frac{1}{p}\sum_{k=1}^{p} h_\theta\big(Z_k(\xi_{1:k-1})\big)^2\right] \le \sup_z \|h_\theta\|^2_{\mathrm{seq},2;z} \le r^2.$$

Hence, $h_\theta \in \mathcal{H}(r)$ implies $R_{\mathrm{BG}}(M_\theta) \le r^2$.

Step 3 (High–probability envelope for the squared loss). Let $\delta := (pN)^{-2}$ and define the event

$$\mathcal{E} := \left\{\max_{j\in[N],k\in[p]} |\varepsilon_{j,k+1}| \le t_\delta\right\}, \qquad t_\delta := \sigma_\varepsilon\sqrt{2\log\left(\frac{2pN}{\delta}\right)}.$$

By the sub-Gaussian tail bound and a union bound, $\Pr(\mathcal{E}^c) \le \delta$. On $\mathcal{E}$, for every (j,k) and every $\theta \in \Theta$ we have

$$\left|y_{j,k+1} - M_\theta(P_j^k)\right| \le |f(\boldsymbol{x}_{j,k+1})| + |\varepsilon_{j,k+1}| + |M_\theta(P_j^k)| \le B_f + t_\delta + B_M =: \widetilde{B},$$

hence, using $\delta = (pN)^{-2}$,

$$\widetilde{B} = B_f + B_M + \sigma_\varepsilon\sqrt{2\log\left(\frac{2pN}{\delta}\right)} \le B_f + B_M + \sigma_\varepsilon\sqrt{6\log(2pN)}.$$

We first carry out the analysis on $\mathcal{E}$ (where the above envelope holds) and add a negligible $O(\delta)$ contribution to expectations in Step 7.

Step 4 (Block symmetrization for the centered objective). We work directly with the centered blocks $\widetilde{\Lambda}_j(\theta) = A_j(\theta) + B_j(\theta)$ and their mean:

$$\sup_{\theta\in\Theta}\left|(\widehat{\mathcal{R}} - \mathcal{R})(\theta)\right| = \sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\widetilde{\Lambda}_j(\theta) - \mathbb{E}\widetilde{\Lambda}_j(\theta)\right|.$$

Since $\widetilde{\Lambda}_1, \ldots, \widetilde{\Lambda}_N$ are i.i.d., standard symmetrization with Rademacher variables $(\epsilon_j)_{j=1}^N$, Cauchy–Schwarz inequality and Jensen inequality give

$$\mathbb{E}\sup_\theta\left|(\widehat{\mathcal{R}} - \mathcal{R})(\theta)\right| \;\le\; \frac{2}{N}\mathbb{E}\sup_\theta\left|\sum_{j=1}^{N}\epsilon_j\,\widetilde{\Lambda}_j(\theta)\right| \;\le\; \frac{C}{\sqrt{N}}\left(\mathbb{E}\sup_\theta\widetilde{\Lambda}_1(\theta)^2\right)^{1/2}.$$

We decompose $\widetilde{\Lambda}_1(\theta) = A_1(\theta) + B_1(\theta)$. From the definition of $\mathcal{H}(r)$, $\mathbb{E}[\sup_{\theta\in\mathcal{H}(r)} A_1^2(\theta)]^{1/2} \le r^2$. We then analyze $B_1^2(\theta) = \{\frac{2}{p}\sum_{k=1}^{p}(y_{1,k+1} - M_{\mathrm{Bayes}}(P_1^k))(M_{\mathrm{Bayes}}(P_1^k) - M_\theta(P_1^k))\}^2$. Note that $|M_{\mathrm{Bayes}}(P_1^k) - M_\theta(P_1^k)| \le B_f + B_M$ and B_1 is constructed by a martingale difference sequence with filtration $\mathcal{G}'_k$. Since $\mathbb{E}[X^2] = 2\int_0^\infty t\Pr(|X| > t)\mathrm{d}t \le 2\int_{t_0}^\infty t\Pr(|X| > t)\mathrm{d}t + 2\int_0^{t_0} t\mathrm{d}t$, evaluation of the tail probability from Lemma 8 in Rakhlin et al. (2015) yields

$$\mathbb{E}\left[\sup_\theta|B_1(\theta)|^2\right]^{1/2} \lesssim \tilde{B}\log^3 p\mathfrak{R}_p^{\mathrm{seq}}\left(\mathcal{H}(r)\right),$$

where $\mathfrak{R}_p^{\mathrm{seq}}(\mathcal{F}) := \sup_z\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}}\frac{1}{p}\sum_{t=1}^{p}\xi_t f(z_t(\xi_{1:t-1}))\right]$ is the depth-p sequential Rademacher complexity. Therefore

$$\mathbb{E}\sup_{\theta\in\mathcal{H}(r)}\left|(\widehat{\mathcal{R}} - \mathcal{R})(\theta)\right| \lesssim \frac{1}{\sqrt{N}}\left\{\log^3 p\,\mathfrak{R}_p^{\mathrm{seq}}(\mathcal{H}(r)) + r^2\right\}.$$

Step 5 (Sequential Dudley bound). The sequential Dudley integral bound (Block et al., 2021, Corollary 10) gives, for an absolute constant $C > 0$,

$$\mathfrak{R}_p^{\mathrm{seq}}\left(\mathcal{H}(r)\right) \le C\inf_{\alpha>0}\left\{\alpha + \frac{1}{\sqrt{p}}\int_\alpha^{\mathrm{diam}(\mathcal{H}(r))}\sup_z\sqrt{\log N'\left(\delta, \mathcal{H}(r); z\right)}\mathrm{d}\delta\right\},$$

where N' denotes the fractional covering number (Block et al., 2021). Note that since every $h \in \mathcal{H}(r)$ satisfies $\|h\|_{\text{seq},2;z} \le r$, the diameter under the path ℓ_2 metric is at most $2r$, so the upper limit can be replaced by $2r$, from Lemma 7 in Block et al. (2021),

$$\mathfrak{R}_p^{\text{seq}}\left(\mathcal{H}(r)\right) \le C \inf_{\alpha>0} \left\{ \alpha + \frac{1}{\sqrt{p}} \int_{\alpha}^{2r} \sup_z \sqrt{\log N_2^{\text{seq}}\left(\delta, \mathcal{H}(r); z\right) \mathrm{d}\delta} \right\}. \tag{12}$$

From Lemma 4, for universal constants $C_0, C_1 > 0$ and all $\delta \in (0, 2r]$,

$$\sup_z \log N_2^{\text{seq}}\left(\delta, \mathcal{H}(r); z\right) \le C_0 m \log\left(\tfrac{\sqrt{m}}{\delta}\right) + C_1 p \log\left(\tfrac{1}{\delta}\right). \tag{13}$$

Plugging (13) into (12) and optimizing over α absorbs polylogarithmic factors to give the succinct bound

$$\mathfrak{R}_p^{\text{seq}}\left(\mathcal{H}(r)\right) \lesssim r \frac{\sqrt{m+p}}{\sqrt{p}} \sqrt{\log\left(\frac{m}{r}\right)}.$$

Step 6 (Self–bounding fixed point).

Let $\Delta_\theta(P^k) := M_\theta(P^k) - M_{\text{Bayes}}(P^k)$, $\ell_\theta(P^k, y_{k+1}) := \{y_{k+1} - M_\theta(P^k)\}^2$, $\ell^{\text{Bayes}}(P^k, y_{k+1}) := \{y_{k+1} - M_{\text{Bayes}}(P^k)\}^2$. Then $R_{\text{BG}}(M_\theta) = \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}[\Delta_\theta(P^k)^2]$ and

$$\ell_\theta - \ell^{\text{Bayes}} = \Delta_\theta^2 - 2\Delta_\theta\{y - M_{\text{Bayes}}(P^k)\}.$$

Hence $\mathbb{E}[\ell_\theta - \ell^{\text{Bayes}} \mid P^k] = \Delta_\theta^2$, and using $|\Delta_\theta| \le B_M + B_f$ and $\mathbb{E}\{y - M_{\text{Bayes}}(P^k)\}^2 \le C(B_f, B_M, \sigma_\varepsilon)$,

$$\mathbb{E}\big[(\ell_\theta - \ell^{\text{Bayes}})^2\big] \le C_0\, \mathbb{E}[\Delta_\theta^2] = C_0\, R_{\text{BG}}(M_\theta),$$

that is, a Bernstein condition with exponent 1 for the excess loss holds.

For $r > 0$, set

$$\Theta(r) := \left\{ \theta \in \Theta :\ R_{\text{BG}}(M_\theta) - \inf_{\vartheta \in \Theta} R_{\text{BG}}(M_\vartheta) \le r \right\}.$$

By the standard symmetrization, we have $\mathbb{E}\sup_{\theta\in\Theta(r)} \big|(\widehat{\mathcal{R}} - \mathcal{R})(\theta) - (\widehat{\mathcal{R}} - \mathcal{R})(\theta^\star)\big| \lesssim \frac{1}{\sqrt{N}}\{\mathfrak{R}_p^{\text{seq}}\big(\mathcal{H}(r)\big) + \sqrt{r + R_{\text{BG}}(M_{\theta^\star})}\}$. Then, from Lemma 4 and Corollary 10 in Block et al. (2021), there exists a constant $c > 0$ (range rescaling absorbed into c) such that $\mathbb{E}\sup_{\theta\in\Theta(r)} \big|(\widehat{\mathcal{R}} - \mathcal{R})(\theta) - (\widehat{\mathcal{R}} - \mathcal{R})(\theta^\star)\big| \lesssim \sqrt{\frac{r + R_{\text{BG}}(M_{\theta^\star})}{N}}\big(1 + \sqrt{\log \frac{c_1}{r + R_{\text{BG}}(M_{\theta^\star})}} + \sqrt{\frac{m}{p}}\ \sqrt{\log \frac{c_2\sqrt{m}}{r + R_{\text{BG}}(M_{\theta^\star})}}\big)$.

By the basic inequality in (11), if $\mathbb{E}\sup_{\theta\in\Theta(r)} \Big|(\widehat{\mathcal{R}} - \mathcal{R})(\theta) - (\widehat{\mathcal{R}} - \mathcal{R})(\theta^\star)\Big| \le \frac{r}{8} + cR_{\text{BG}}(M_{\theta^\star})$ with $c = o(1)$, then the ERM satisfies $R_{\text{BG}}(M_{\hat\theta}) - (1+c)R_{\text{BG}}(M_{\theta^\star}) \le r/2$. Let the critical radius $r_\star$ be the smallest $r > 0$ solving $\frac{r}{8} \asymp \sqrt{\frac{r}{N}}(\sqrt{\frac{m}{p}} + 1)$. Hence $r_\star \asymp \frac{1}{N}(\frac{m}{p} + 1)$. Then, the ERM obeys $\mathbb{E}R_{\text{BG}}(M_{\hat\theta}) \lesssim \inf_{\theta\in\Theta} R_{\text{BG}}(M_\theta) + \frac{1}{N}(\frac{m}{p} + 1)$.

Step 7 (Control on $\mathcal{E}^c$). Recall $\mathcal{E} := \{\max_{j,k} |\varepsilon_{j,k+1}| \le t_\delta\}$ with $t_\delta := \sigma_\varepsilon\sqrt{2\log(2pN/\delta)}$. By sub-Gaussian tails and a union bound,

$$\Pr(\mathcal{E}^c) = \Pr\Big(\exists (j,k) : |\varepsilon_{j,k+1}| > t_\delta\Big) \le 2pN \exp\Big(-\frac{t_\delta^2}{2\sigma_\varepsilon^2}\Big) \le \delta.$$

Let

$$\tilde{T} := \sup_{\theta\in\Theta} \Big|(\widehat{\mathcal{R}} - \mathcal{R})(\theta)\Big|$$

and note that, by the identity $(y - M_\theta)^2 - (y - M_{\text{Bayes}})^2 = (M_{\text{Bayes}} - M_\theta)\{2y - M_\theta - M_{\text{Bayes}}\}$, Assumptions 1 and 2 imply a quadratic envelope of the form

$$\tilde{T} = \sup_{\theta\in\Theta} \Big|(\widehat{\mathcal{R}} - \mathcal{R})(\theta)\Big| \ \le\ C\Big\{(B_f + B_M)^2 + \frac{1}{pN}\sum_{j,k} \varepsilon_{j,k+1}^2 + \mathbb{E}\varepsilon^2\Big\}.$$

for some constant $C > 0$ (where $C, C', \dots$ below are universal constants). Thus,

$$\tilde{T} \le C\left\{(B_f + B_M)^2 + \frac{1}{pN}\sum_{j,k}\varepsilon_{j,k+1}^2 + \sigma_\varepsilon^2\right\}. \tag{14}$$

To bound the expectation of $\tilde{T}$ on $\mathcal{E}^c$, we bound the tail of the second moment of each ε. For any $t > 0$, from the sub-Gaussian (ψ_2) tail probability,

$$\mathbb{E}\left[\varepsilon^2 \mathbf{1}_{\{|\varepsilon|>t\}}\right] = \int_t^\infty 2x \Pr(|\varepsilon| > x)\mathrm{d}x \le 2\int_t^\infty 2x \exp\Big(-\frac{x^2}{2\sigma_\varepsilon^2}\Big)\mathrm{d}x \le 4\sigma_\varepsilon^2 \exp\Big(-\frac{t^2}{2\sigma_\varepsilon^2}\Big).$$

Substituting $t = t_\delta$ yields $\mathbb{E}[\varepsilon^2 \mathbf{1}_{\{|\varepsilon|>t_\delta\}}] \le 2\sigma_\varepsilon^2\delta/(pN)$. Furthermore, by the decomposition

$$\varepsilon_{j,k+1}^2 \mathbf{1}_{\mathcal{E}^c} \le \varepsilon_{j,k+1}^2 \mathbf{1}_{\{|\varepsilon_{j,k+1}|>t_\delta\}} + t_\delta^2 \mathbf{1}_{\mathcal{E}^c},$$

it follows that

$$\frac{1}{pN}\sum_{j,k}\mathbb{E}\left[\varepsilon_{j,k+1}^2 \mathbf{1}_{\mathcal{E}^c}\right] \le \underbrace{\frac{1}{pN}\sum_{j,k}\mathbb{E}\left[\varepsilon_{j,k+1}^2 \mathbf{1}_{\{|\varepsilon_{j,k+1}|>t_\delta\}}\right]}_{\le 2\sigma_\varepsilon^2\delta/(pN)} + t_\delta^2 \Pr(\mathcal{E}^c). \tag{15}$$

Combining (14) and (15), we get

$$\mathbb{E}\left[\tilde{T}\mathbf{1}_{\mathcal{E}^c}\right] \le C\left\{(B_f + B_M)^2 + \sigma_\varepsilon^2\right\}\Pr(\mathcal{E}^c) + C'\left\{\sigma_\varepsilon^2\delta + t_\delta^2 \Pr(\mathcal{E}^c)\right\}.$$

Substituting $t_\delta^2 = 2\sigma_\varepsilon^2 \log\frac{2pN}{\delta}$ and $\Pr(\mathcal{E}^c) \le \delta$,

$$\mathbb{E}\left[\tilde{T}\mathbf{1}_{\mathcal{E}^c}\right] \le C(B_f + B_M)^2\delta + C'\sigma_\varepsilon^2\delta\log\frac{2pN}{\delta} + C''\sigma_\varepsilon^2\delta.$$

Finally, by using $\delta = (pN)^{-2}$, the right-hand side becomes $O\big(\sigma_\varepsilon^2(\log pN)/(pN)^2\big)$, which is negligible compared to the main term from Step 5. □

J LLM Usage Disclosure

We used LLMs for English proofreading and style suggestions on early drafts (grammar, phrasing, and minor clarity edits). All technical statements, theorems, and proofs were written and verified by the authors, who take full responsibility for the content.