Autoregressive Learning in Joint KL: Sharp Oracle Bounds and Lower Bounds*

Yunbei Xu[1] **Yuzhe Yuan**[1] **Ruohan Zhan**[2]
[1]National University of Singapore [2]London Business School
yunbei@nus.edu.sg yuzhe.yuan@u.nus.edu rzhan@london.edu

Abstract

We study the fundamental problem of learning long sequences in autoregressive modeling and next-token prediction under model misspecification, measured by the joint Kullback–Leibler (KL) divergence. Our goal is to characterize how the sequence horizon H affects both approximation and estimation errors in this joint-distribution, sequence-level regime. By establishing matching upper and lower bounds, we provide, to our knowledge, the first complete characterization of long-horizon error behavior under the natural joint KL objective, with improved rates and optimality justification relative to existing work. On the approximation side, we show that joint KL admits a horizon-free approximation factor, in sharp contrast to Hellinger-based analyses that exhibit an $\Omega(H)$ dependence for computationally efficient methods; this isolates the choice of divergence as the source of approximation amplification. On the estimation side, we prove fundamental information-theoretic lower bounds of order $\Omega(H)$ that hold both for decomposable policy classes, even under convexified prediction, and for fully shared policy classes under proper learning, matching the $\widetilde{O}(H)$ upper bounds achieved by computationally efficient algorithms. Our analysis clarifies the landscape of recent autoregressive learning results by aligning the log-loss training objective, the sequence-level evaluation metric, and the approximation metric through a sharp joint-KL oracle theory. We further show that these joint-KL guarantees imply policy learning regret bounds at rates matching prior imitation learning literature.

1 Introduction

Learning distributions over long sequences is a central problem in modern AI and in applications of large language models. Autoregressive (AR) modeling addresses this challenge by predicting the next token given the history and repeating this step to generate a full sequence. It does so by factorizing the joint distribution through conditional next-step laws, enabling likelihood-based training and principled probabilistic evaluation. This formulation has become the dominant template for sequence modeling, powering large-scale language models (Bengio et al., 2000) and naturally extending to multimodal generation (van den Oord et al., 2016), sequential density estimation (Larochelle and Murray, 2011), and model-based forecasting (Salinas et al., 2020).

We formalize AR modeling as learning a policy that induces a joint distribution over length-H sequences under the maximum-likelihood objective. Here we adopt the useful dynamical-systems lens (Rohatgi et al., 2025; Hazan et al., 2025), to regard the history as a state, the next token as an action or output, and the data-generating mechanism as a policy. Formally, let $\pi^\star = (\pi_1^\star, \ldots, \pi_H^\star)$ denote the data-generating policy for a length-H sequence $S = (u_1, \ldots, u_H)$, where $\pi_h^\star(\cdot \mid u_{<h})$ is the conditional law of the h-th token given its history $u_{<h}$. Let $P^{\pi^\star}$ be the induced joint distribution

*Authors are listed in alphabetical order. This project was conducted collaboratively. Yunbei Xu and Ruohan Zhan acknowledge Yuzhe Yuan's significant contributions to this work.

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

over S. For any candidate policy $\pi = (\pi_1, \ldots, \pi_H) \in \Pi$, denote by P^π the induced joint distribution. AR modeling exploits the sequential structure through the autoregressive factorization

$$P^\pi(u_{1:H}) \;=\; \prod_{h=1}^{H} \pi_h(u_h \mid u_{<h})\,, \qquad u_{<h} = (u_1, \ldots, u_{h-1}). \tag{1}$$

The population maximum-likelihood (equivalently, cross-entropy) objective (Goodfellow et al., 2016, §20), used explicitly in autoregressive language modeling since GPT-1 (Radford et al., 2018), aims to minimize the expected negative log-likelihood under $P^{\pi^\star}$ and is therefore equivalent to minimizing the *joint* Kullback–Leibler (KL) divergence:

$$\min_{\pi\in\Pi} \mathbb{E}_{S\sim P^{\pi^\star}}\left[-\sum_{h=1}^{H} \log \pi_h(u_h|u_{<h})\right] \iff \min_{\pi\in\Pi} \mathrm{KL}\Big(P^{\pi^\star} \,\|\, P^\pi\Big)\,. \tag{2}$$

Accordingly, the joint-KL objective (2) is the natural target in generative modeling and sequence prediction practice; we refer to learning under this objective as *autoregressive learning*.

Despite its practical importance, autoregressive learning is not a standard i.i.d. statistical learning problem. The joint-KL objective is a trajectory-level criterion: later prefixes are generated by earlier tokens, so the covariate distribution evolves with the horizon rather than being fixed exogenously. This endogenous sequential structure is the source of the price of long horizons studied in this paper.

In this work, we study the generalization properties of autoregressive learning under *misspecification*, where the true policy $\pi^\star$ may not belong to the specified policy class Π. A standard decomposition of the generalization risk of an estimator $\hat{\pi}$ takes the form

$$d(P^{\pi^\star}, P^{\hat{\pi}}) \le \underbrace{C_{\text{apx}} \cdot \min_{\pi\in\Pi} d(P^{\pi^\star}, P^\pi)}_{\text{approximation error}} + \underbrace{C_{\text{est}} \cdot \frac{\log(|\Pi_0|/\delta)}{n} + \tilde{\mathcal{O}}(\frac{H}{n})}_{\text{estimation error with fast rate}},$$

where n denotes the number of *i.i.d.* trajectory samples and Π_0 denotes the per-step base policy class (Section 2). Throughout, we evaluate models by their *joint* KL to the data-generating process as in (2), which means $d(P, Q) = \mathrm{KL}(P\|Q)$ and is exactly the population objective of log-loss training. Unlike standard in-sample conditional density estimation, (2) places autoregressive learning in a distinct joint-distribution regime: the learned conditional factors are judged by the full sequence law they induce, so estimation error and misspecification may accumulate through the autoregressive factorization. This raises a basic question: *how does the sequence length H enter the approximation and estimation terms above under joint sequence-level evaluation, and which of these dependencies are intrinsic rather than consequences of the chosen metric or analysis?*

This question is timely and connects to recent theory on autoregressive learning, next-token prediction, and in-context learning. Motivated by the maximum-likelihood objective used in GPTs (Radford et al., 2018), early studies (Xie et al., 2022; Wies et al., 2023; Wang et al., 2023; Jiang, 2023) developed Bayesian or latent-variable interpretations of in-context learning. These works explain how autoregressive models can perform implicit inference from context, but they do not provide a misspecified joint-KL oracle theory. Building on this line, Zhang et al. (2023) formulate next-token prediction as autoregressive density estimation under joint metrics and misspecification, with their main result stated in joint total variation distance. Although joint-KL-type quantities appear in their proofs and in-context-learning consequences, their results do not yield a sharp estimation–approximation decomposition for the joint KL metric. Relatedly, Yüksel and Flammarion (2025) obtain KL-type guarantees for log-loss training of autoregressive processes, but their approximation term is measured by a worst-case total-variation criterion rather than by the joint KL metric used for evaluation. None of these works develops lower bounds or an optimality theory that distinguishes how horizon dependence changes across different structural restrictions on the policy class.

A second line studies autoregressive and imitation learning under squared Hellinger distance through the minimax-optimality lens. Foster et al. (2024a) analyze log-loss training for general behavior cloning and obtain finite-sample guarantees that imply multi-step control under Hellinger-type evaluation, with autoregressive learning as a special case. Rohatgi et al. (2025) focus more directly on misspecified autoregressive learning and show that, under squared Hellinger evaluation, computationally tractable algorithms can suffer an $\Omega(H)$ amplification of approximation error. These lower

	Approximation error	Estimation error	
		Decomposable class	Fully-shared class
KL divergence (our results)	$\Theta(1)$	$O\left(\frac{H \log\left(H\lvert\Pi_0\rvert\right)}{n}\right)$	$\Theta\left(\frac{H \log \lvert\Pi_0\rvert}{n}\right)$
Squared Hellinger distance (Rohatgi et al., 2025)	$\Theta(H)$	$\Theta\left(\frac{H \log \lvert\Pi_0\rvert}{n}\right)$	$\Theta\left(\frac{\log \lvert\Pi_0\rvert}{n}\right)$

Table 1.1: Comparison results for proper learners in sequence horizon H and sample size n; Π_0 is the per-step base policy class.

bounds are highly informative, but these results also suggest that the horizon behavior of autoregressive learning is metric-dependent. Squared Hellinger does not coincide with the population objective induced by log-loss training, and its pseudo-chain-rule behavior leads to a different approximation story from the exact KL chain-rule decomposition studied here. Our work gives a sharp joint-KL analysis in a general autoregressive learning setting. Both the evaluation risk and the approximation error are measured in the same joint-KL metric, so the result directly characterizes the population objective corresponding to next-token log-loss training. We provide upper and lower bounds for both approximation and estimation, and treat both fully-shared and decomposable policy classes. Our work provides a unified answer by developing a sharp horizon scaling map for misspecified autoregressive learning under log-loss. At a high level, Table 1.1 summarizes the canonical behaviors of approximation and estimation under joint KL versus squared Hellinger, and clarifies how sharing across steps affects the statistical term. Beyond the proper-learning comparison, we also investigate Bayesian aggregation and convexified prediction, examining how different aggregation mechanisms interact with policy-class structure under joint KL. With this picture in place, we can state the main contributions.

1.1 Main Results and Takeaways

We summarize our main results and takeaways below; proofs and extensions are deferred to the appendix.

- **Clarifying the KL formulation in the joint-distribution-level regime.** We study the new joint-distribution-level regime of autoregressive learning and organize how prior work has approached this paradigm under TV-, Hellinger-, and KL-type analyses. By disentangling the training objective, sequence-level evaluation metric, and approximation metric, we further clarify the relationships among existing results and identify joint KL as the metric-aligned formulation for log-loss training.
- **Misspecified guarantees under joint KL (and what drives horizon effects).** We establish a sharp misspecified generalization guarantee for log-loss, showing that the approximation ratio satisfies $C_{\mathrm{apx}} = O(1)$ under the *joint* KL objective. In particular, in contrast to the $\Omega(H)$ barrier under squared Hellinger distance (Rohatgi et al., 2025), this shows that horizon growth in C_{apx} can arise from the choice of divergence rather than an inherent limitation of autoregressive factorization (see Theorem 3.1).
- **Algorithmic upper bounds under joint KL.** We analyze ERM and Bayesian aggregation learners and derive joint-KL generalization bounds with estimation ratio $C_{\mathrm{est}} = \tilde{O}(H)$. The bounds make explicit the tradeoff between approximation and estimation across two canonical policy-class regimes (decomposable versus fully shared). For general dependent classes, we further develop computationally tractable stepwise learning over a lifted policy space, removing cross-horizon compatibility constraints (see Section 3.2). The bounds further imply rollout-level policy learning guarantees to be highlighted below (see also Section 5).
- **Information-theoretic lower bounds (unavoidable linear horizon dependence).** We establish matching lower bounds showing that the estimation ratio C_{est} can scale as $\Omega(H)$ under joint KL. For decomposable classes, this dependence persists even with trajectory-level convexified prediction, that is, even when the learner may output any mixture in the convex hull of the induced trajectory laws (which covers Bayesian aggregation); while for fully shared classes, it remains unavoidable under proper learning, that is, for learners that

must output an element of the class (see Theorems 4.2 and 4.3). These results refine the interpretation of "error amplification" and show that the linear horizon dependence in our upper bounds is tight up to logarithmic factors, and clarify that the fundamental obstruction under joint KL is statistical rather than approximation-driven.

Two points are worth highlighting in more detail:

Advantages of KL over Hellinger in our setting. Our analysis and accompanying insights point to two reasons to favor KL over the squared Hellinger distance adopted in Rohatgi et al. (2025). First, since *estimation error can be significantly reduced with larger sample size* n, it is preferable for any horizon dependence H to appear in the estimation term rather than the approximation term; comparing the two rows of Table 1.1 shows that our KL-based formulation yields the more favorable rate in this regime. Second, joint KL aligns with the canonical objective in modern generative modeling and large language model training (Ermon, 2023; Radford et al., 2018).

Policy learning guarantee. Our joint-KL bounds translate directly into rollout-level policy learning guarantees: by Pinsker's inequality, they control the worst-case return gap over bounded trajectory rewards at the usual $\sqrt{\cdot}$ order. Hence joint KL matches the rollout control order obtained from squared Hellinger distance, while preserving our estimation–approximation separation.

2 Problem Setting

Let $\pi^\star = (\pi_1^\star, \ldots, \pi_H^\star)$ denote the data-generating model and $P^{\pi^\star}$ be the induced joint distribution over length-H token sequences. Given n i.i.d. samples $(S_1, \cdots, S_n)$, where each sequence $S_i \sim P^{\pi^\star}$ consists of tokens $\{u_1^i, \cdots, u_H^i\}$, our goal is to learn a model π from a pre-specified class Π that best approximates $\pi^\star$ in KL divergence:

$$\min_{\pi \in \Pi} \mathrm{KL}\Big(P^{\pi^\star} \,\|\, P^{\pi}\Big). \tag{2}$$

In particular, we allow for model misspecification $\pi^\star \notin \Pi$. Our objective is therefore to characterize both the *approximation error* induced by restricting to Π and the *estimation error* arising from finite-sample statistical learning. Following standard terminology, a learner is *proper* if its output $\hat{\pi}$ always belongs to Π, and *improper* otherwise; for example, a mixture of elements of Π (such as the Bayesian aggregation predictor of Section 3.1) is an improper output. Unless stated otherwise, the comparisons in Table 1.1 concern proper learners.

Denote by $s_h = u_{<h}$ the length-$(h-1)$ prefix, and write $\pi = (\pi_1, ..., \pi_H)$ for a sequence model. We adopt a widely used assumption of bounded log density (Xie et al., 2022; Rohatgi et al., 2025).

Assumption 1 (Boundedness of log density)**.**

$$\sup_{\pi \in \Pi} \sup_{h \in [H]} \sup_{s_h,\, u_h} \left| \log \frac{\pi_h^\star(u_h|s_h)}{\pi_h(u_h|s_h)} \right| \le G.$$

We distinguish between two structural regimes for the policy class Π.

Decomposable policy classes (no parameter sharing). In the *decomposable* regime, often called *no parameter sharing* in prior work (e.g., Foster et al., 2024a; Rohatgi et al., 2025) and motivated by horizon-specific parameterizations in practice (e.g., Sabbaghi et al., 2024), the model class is Cartesian across steps, so the conditional policy at each step can be selected independently from its corresponding class. Concretely, the class factorizes across the horizon as

$$\Pi \;=\; \Pi_1 \times \cdots \times \Pi_H,$$

where Π_h denotes the set of admissible conditional distributions at step h. Viewing each prefix through its position-tagged state $(h, u_{<h})$, all conditional policies can be regarded as acting on a common state space. For ease of presentation, we focus on the homogeneous setting,

$$\Pi_1 = \cdots = \Pi_H =: \Pi_0,$$

so that $\Pi = \Pi_0^H$. Note that heterogeneous per-step classes $(\Pi_h)_{h=1}^H$ can always be reduced to this homogeneous form by taking the lifted base class $\Pi_0 := \bigcup_{h=1}^H \Pi_h$. More generally, one may consider

policy classes that do not decompose across steps, i.e., classes satisfying only $\Pi \subseteq \Pi_0^H$. We defer the analysis of such dependent (non-decomposable) classes to Section 3.2.

Fully-shared policy classes (parameter sharing). A second regime of primary interest is the *fully shared* (or *consistent*) policy class, obtained by using the same conditional policy across all horizons. This restriction is analytically cleaner and is adopted in prior joint-metric analyses (Zhang et al., 2023; Yüksel and Flammarion, 2025). Formally, this class is given by

$$\Pi_{\mathrm{cons}} := \big\{\pi = (\pi_0, \dots, \pi_0) \in \Pi_0^H : \pi_0 \in \Pi_0\big\},$$

so that $\pi_h \equiv \pi_0$ for all $h \in [H]$ and $\Pi_{\mathrm{cons}} \subseteq \Pi_0^H$. This regime captures the canonical parameter-sharing structure in autoregressive model training (e.g., the original Transformer architecture), while still allowing the induced joint distribution P^π to exhibit long-range dependence through the evolving history $u_{<h}$.

3 Upper Bounds of Autoregressive Learning in Joint KL

3.1 Joint-KL Guarantees: Constant Approximation and an Estimation Barrier

Two canonical log-loss-aligned learners: ERM and Bayesian aggregation. We begin by studying two classical learners for next-token prediction under the log-loss, both directly aligned with the *joint* KL objective in (2). The first is empirical risk minimization (ERM), which outputs any minimizer of the empirical negative log-likelihood:

$$\hat{\pi} \in \arg\min_{\pi \in \Pi} -\frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \log \pi_h(u_h^i \mid u_{<h}^i). \tag{3}$$

This is the canonical training paradigm for modern language models: Transformer pretraining is precisely maximum-likelihood learning over large text corpora.

As a complementary baseline, we consider a Bayesian aggregation learner, also known as an exponential-weights or Gibbs learner, that aggregates hypotheses according to their empirical log-likelihood. Concretely, starting from a prior over Π_h, the per-step posterior after observing n trajectories yields a mixture predictor

$$\hat{\pi}_h(\cdot \mid s_h) \propto \sum_{\pi_h \in \Pi_h} q_h(\pi_h)\, \pi_h(\cdot \mid s_h), \tag{4}$$

where q_h is the posterior weight updated from the prior by the empirical likelihood $\prod_{i=1}^{n} \pi_h(u_h^i | u_{<h}^i)$. This is a classical improper exponential-weights aggregation procedure under log-loss: the output mixture $\widehat{\pi}_h$ belongs to the convex hull of the original class Π_h. All posterior weights are fitted offline from the n observed trajectories and remain fixed during the generation of a new sequence.

We next show that *both* learners admit oracle-type guarantees under joint KL in the misspecified regime, with horizon-free approximation and horizon-dependent estimation terms. Throughout this paper, we use the notation $f \lesssim g$ to mean that there exists an absolute constant $c > 0$ such that $f \le c \cdot g$.

Theorem 3.1 (ERM and Bayesian aggregation under joint KL)**.** *Let $S_1, \dots, S_n \overset{\mathrm{iid}}{\sim} P^{\pi^\star}$ with $\pi^\star \notin \Pi$, where Π is a finite decomposable model class. Under Assumption 1, ERM algorithm* (3) *and Bayesian aggregation algorithm* (4) *guarantee that for any $\delta \in (0,1)$ and any $\varepsilon > 0$, with probability at least $1 - \delta$ and for a universal constant $C > 0$,*

$$\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \le \underbrace{(1+\varepsilon) \min_{\pi \in \Pi} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})}_{\textit{approximation error}} + \underbrace{C\, \frac{(1+\varepsilon)^2}{\varepsilon} \frac{(HG+1)\log(H|\Pi_0|\delta^{-1})}{n}}_{\textit{estimation error}}. \tag{5}$$

For fully shared model class Π_{cons}, the result would be

$$\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \le \underbrace{(1+\varepsilon) \min_{\pi \in \Pi_{\mathrm{cons}}} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})}_{\textit{approximation error}} + \underbrace{C\, \frac{(1+\varepsilon)^2}{\varepsilon} \frac{(HG+1)\log(|\Pi_0|\delta^{-1})}{n}}_{\textit{estimation error}}. \tag{6}$$

The two regimes trade off approximation and estimation: since $\Pi_{\text{cons}} \subset \Pi$*, the decomposable class can only reduce the approximation term, while its estimation term differs from the shared case only by an additional* $\log H$ *factor.*

The first terms in (5) and (6) show that the approximation error does not amplify with the horizon (i.e., sequence length): the approximation ratio C_{apx} on $\min_{\pi\in\Pi} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})$ is $(1+\varepsilon)$, independent of H.

Comparison and technical distinction relative to Hellinger analyses (Rohatgi et al., 2025). The reason why joint KL can achieve horizon independent approximation ratio comes from the *exact chain rule* for KL, which decomposes the joint KL into a sum of conditional KLs along the sequence:

$$\mathrm{KL}\Big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\Big) = \sum_{h=1}^{H} \mathbb{E}_{s_h\sim P^{\pi^\star}(u_{<h})}\Big[\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h) \,\|\, \hat{\pi}_h(\cdot \mid s_h)\big)\Big].$$

By contrast, the squared Hellinger distance satisfies only a *pseudo chain rule*. A convenient form is

$$\frac{1}{7}\, D_H^2\big(P^{\pi^\star}, P^{\hat{\pi}}\big) \le \sum_{h=1}^{H} \mathbb{E}_{s_h\sim P^{\pi^\star}(u_{<h})}\Big[D_H^2\big(\pi_h^\star(\cdot\,|\,s_h),\ \hat{\pi}_h(\cdot\,|\,s_h)\big)\Big] \le 2H\, D_H^2\big(P^{\pi^\star}, P^{\hat{\pi}}\big),$$

see, e.g., Foster et al. (2024b). Consequently, stepwise Hellinger analyses incur a linear in H factor, explaining the $\Omega(H)$ dependence reported in prior work (Rohatgi et al., 2025), whereas the KL chain rule prevents such amplification. This clarifies that the $\Omega(H)$ growth is a metric-induced artifact.

The proof techniques are also distinct. Their analysis exploits a sequence-level log-loss–Hellinger relation, together with the boundedness property and the triangle inequality of the Hellinger distance, which simplifies the horizon-dependence analysis but is specific to Hellinger and leads to the decomposable class gap reported in Table 1.1; see the discussion around (9). Our joint-KL analysis must instead track prefixes and union-bound across horizons explicitly. Theorem 3.1 shows that, even under this sharper metric, decomposable classes incur only $H \log H$, rather than the naive H^2, horizon dependence.

Improvements over TV and KL analyses (Zhang et al., 2023; Yüksel and Flammarion, 2025). Zhang et al. (2023) state their main guarantee in joint total variation. Since Pinsker converts KL to TV only at square-root scale, a TV guarantee derived from log-loss complexity naturally yields an $n^{-1/2}$-type rate, whereas direct joint-KL analysis can retain the n^{-1}-type oracle rate under standard bounded log-loss conditions.

It is also useful to compare Theorem 3.1 with Yüksel and Flammarion (2025), whose KL objective is additionally normalized by H; below we use our sequence-level normalization for comparison. Their Theorems 3.4 and 3.5 correspond closely to Step 1 and Step 2 in our proof sketch, but their approximation term is measured in worst-case total variation, which is not a joint metric:

$$\mathcal{E}_{\text{app}} \lesssim G \cdot \inf_{\pi\in\Pi} \sum_{h=1}^{H} \sup_{s_h} \mathrm{TV}\left(\pi_h^\star(\cdot|s_h), \pi_h(\cdot|s_h)\right).$$

Even strengthening their analysis to sequence-level TV would not give a sharp oracle bound in the evaluation metric we use, namely joint KL. By basic inequalities, under Assumption 1 we have

$$\mathrm{KL}(P^{\pi^\star}\|P^{\pi}) \lesssim G \cdot \mathbb{E}_{P^{\pi^\star}}\left[\sum_{h=1}^{H} \mathrm{TV}(\pi_h^\star(\cdot|s_h), \pi_h(\cdot|s_h))\right] \lesssim G\sqrt{H\mathrm{KL}(P^{\pi^\star}\|P^{\pi})},$$

showing that even a sequence-level TV approximation would yield horizon dependence in the worst case.

Moreover, applying the bounds of Zhang et al. (2023); Yüksel and Flammarion (2025) to decomposable classes leads to a naive H^2-level estimation term. Our ERM and posterior bounds instead scale linearly in H, again up to logarithms. This advantage is relevant for horizon- or position-dependent parameterizations used in practice (Sabbaghi et al., 2024).

Core proof ideas. The argument first proves a prefix-conditioned oracle inequality for the one-step log-loss at each fixed step h and prefix s_h. For ERM this uses a Bernstein-type concentration argument, while for the Bayesian aggregation it follows from the mixability of log-loss. We then

average these local inequalities over the random prefixes generated by the data; a Freedman-type martingale bound controls the deviation between this empirical prefix average and its population counterpart. The exact KL chain rule then converts the stepwise control into the desired joint-KL bound. In particular, for decomposable classes, the stepwise argument requires uniform control only over the coordinate-wise index set $\{(h, \pi_h) : h \in [H], \pi_h \in \Pi_0\}$, yielding the $\log(H|\Pi_0|)$ factor instead of the complexity of the full product class, improving the naive H^2-horizon dependence to $H \log H$.

On the impossibility of $C_{\text{apx}} = 1$ high-probability bounds. It is natural to ask whether the Bayesian aggregation can satisfy a high-probability oracle inequality under joint KL with leading constant 1 on the misspecification term and only a complexity-dependent remainder. While mixability of log-loss can deliver such sharpness at the prefix-conditioned level, it does not extend through the averaging-over-prefixes step needed to control the joint objective. As a result, sharp high-probability bounds under joint KL cannot generally be achieved. The following theorem formalizes this impossibility in the misspecified regime, for any proper learner.

Theorem 3.2 (No $C_{\text{apx}} = 1$ high-probability oracle bound under joint KL)**.** *There exist absolute constants $c > 0$ and $\delta_0 \in (0, 1)$ such that the following holds. For any $H, n \in \mathbb{N}$, one can construct a misspecified data-generating policy $\pi^\star$ and a base class Π_0 satisfying Assumption 1 such that for*

$$\Pi \in \Big\{\Pi_0^H,\ \Pi_{\text{cons}} := \{(\pi_0, \dots, \pi_0) : \pi_0 \in \Pi_0\}\Big\},$$

every (possibly randomized) learner outputting $\hat{\pi} \in \Pi$ from n i.i.d. trajectories satisfies

$$\Pr\left[\mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\hat{\pi}}\Big) \ \geq\ \inf_{\pi \in \Pi} \mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\pi}\Big) \ +\ c\,\frac{H}{\sqrt{n}}\right] \ \geq\ \delta_0. \tag{7}$$

In particular, a sharp high-probability oracle inequality with leading constant 1 and only a $\widetilde{O}(H \log |\Pi_0|/n)$ remainder is impossible in the misspecified regime.

3.2 General Dependent Policy Classes

We now consider the more general *dependent* regime beyond the two canonical structures above, where the admissible stepwise conditionals are coupled across horizons:

$$\Pi \subseteq \Pi_1 \times \cdots \times \Pi_H.$$

Without loss of generality, we take each Π_h to be the minimal induced marginal class

$$\Pi_h := \{\pi_h : \pi \in \Pi\},$$

so Π_h is exactly the collection of step-h conditionals realizable by some $\pi \in \Pi$. Such dependencies naturally arise from parameter sharing, nonconvex architectural constraints, or combinatorial structures, and they prevent choosing $(\pi_1, \dots, \pi_H)$ independently. Here in our discussion, "computational efficiency" is used in the specific sense of avoiding the horizon-dependent planning inefficiency highlighted in Section 3 of Rohatgi et al. (2025).

Coupled learning over the original class: computational and statistical barriers. Learning directly over Π requires respecting all cross-horizon compatibility constraints simultaneously, so the H conditional components can no longer be optimized independently. In the worst case, empirical log-loss minimization therefore entails a global search over exponentially many joint configurations, rather than H separate stepwise problems. The same coupling can also amplify statistical difficulty: even ignoring computational cost and allowing an arbitrary selector over the original class, the joint-KL estimation error can scale as $\Omega(HG^2 \log |\Pi|/n)$ in the worst case (see Corollary 1 in Appendix B). Thus, cross-horizon dependence can create both an exponential search burden and an additional statistical penalty.

Lifting: recovering stepwise separability. We therefore *lift* the search space to the stepwise product class $\widetilde{\Pi} := \Pi_0^H$ where $\Pi_0 := \bigcup_{h=1}^H \Pi_h$ is the marginal union class. This lifting removes the cross-horizon compatibility constraints and enables *stepwise* ERM or Bayesian aggregation updates, which are computationally tractable and yield an $O(H)$-type estimation term of order $\tilde{O}(H \log |\Pi_0|/n)$. Since $\widetilde{\Pi}$ may strictly contain Π, the resulting learner need not return a policy in the original coupled class. The following lemma relates the sizes of Π and Π_0, allowing a direct comparison between the coupled and lifted regimes.

Lemma 1 (Comparison between coupled and lifted classes)**.** *For any $\Pi \subset \Pi_1 \times \cdots \times \Pi_H$ where $\Pi_h := \{\pi_h : \pi \in \Pi\}$, writing $\Pi_0 := \bigcup_{h=1}^{H} \Pi_h$, we have that:*

$$\log|\Pi_0| \le \log|\Pi| + \log H$$
$$\log|\Pi| \le H \log|\Pi_0|$$

Therefore, the estimation-error gain from using the lifted class is sandwiched as

$$H \gtrsim \frac{\log|\Pi|}{\log|\Pi_0|} \gtrsim \frac{\log|\Pi|}{\log|\Pi| + \log H},$$

so the gain is on the order of a factor ranging from $O(1)$ to $O(H)$ in the horizon (modulo logarithms), at the cost of learning in the lifted space. We summarize the computational and statistical gains by lifting as the following theorem.

Theorem 3.3 (Computational and statistical gains by lifting)**.** *With $\widetilde{\Pi} = \Pi_0^H$ as above, stepwise ERM or Bayesian aggregation updates are computationally tractable and achieve $\widetilde{O}(H \log|\Pi_0|/n)$ estimation. In contrast, proper learning over a general dependent class Π can incur $\Omega(HG^2 \log|\Pi|/n)$ joint-KL estimation error even in the realizable setting. By Lemma 1, optimizing over the lifted space corresponds to an up to H gap in horizon dependence.*

4 Lower Bounds of Autoregressive Learning in Joint KL

In this section we discuss error amplification from both the approximation and estimation perspectives. To facilitate a clean comparison across metrics, we still assume that the candidate policy space is identical across steps and focus on two representative cases: decomposable classes and fully shared classes. In particular, Theorems 4.2 and 4.3 provide the paper's main lower bounds. This is a central contribution, as it identifies the intrinsic horizon-dependent statistical cost of autoregressive learning and clarifies the fundamental limits of earlier upper-bound analyses.

Approximation-side error amplification. As discussed earlier, squared-Hellinger analyses imply that, for decomposable classes, computationally efficient procedures may suffer an approximation ratio scaling as $C_{\mathrm{apx}} = \Omega(H)$ in long-horizon AR modeling. Here we interpret this phenomenon through the lens of *error amplification*. The best-in-class misspecification level $\inf_{\pi\in\Pi} d(P^{\pi^\star}, P^{\pi})$ typically grows with H simply because matching a length-H joint distribution becomes progressively harder. An $\Omega(H)$ inflation in C_{apx} therefore *amplifies* this horizon-growing approximation gap into a much larger sequence-level error, echoing the classical compounding-error intuition from imitation learning. Importantly, in the same decomposable regime, our joint-KL guarantee removes this $\Omega(H)$ blow-up in the approximation ratio. Holding the policy-class structure fixed therefore isolates a *metric-induced* effect: the amplification is not an inherent feature of sequential prediction but a consequence of the evaluation criterion.

Estimation-side error amplification on KL divergence. We now focus on the statistical term and make the dependence on the policy class structure explicit. For squared Hellinger distance, a standard proper-learning guarantee controls the estimation error directly through the complexity of the joint policy class. We state the realizable result below, which cleanly isolates this statistical term.

Theorem 4.1 (Foster et al. (2024a), Proposition 2.1)**.** *Fix any expert $\pi^\star \in \Pi$. Let $\hat{\pi}$ be the output of the ERM algorithm (3). Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$D_H^2\Big(P^{\hat{\pi}},\, P^{\pi^\star}\Big) \;\le\; 2\,\frac{\log\big(|\Pi|\,\delta^{-1}\big)}{n}. \tag{8}$$

The bound depends on the policy class only through $\log|\Pi|$. Thus, for the decomposable class $\Pi = \Pi_0^H$, it gives an estimation rate of order $H\log|\Pi_0|/n$, whereas for the fully-shared class $|\Pi| = |\Pi_0|$, the rate becomes $\log|\Pi_0|/n$. Hence, under proper learning, the Hellinger estimation rate directly benefits from parameter sharing through the reduced complexity of the joint policy class.

Unlike squared Hellinger distance, joint KL would retain a linear dependence on the horizon even under full parameter sharing. We establish two complementary lower bounds for misspecified excess-KL risk. Theorem 4.2 shows that, for decomposable classes, the linear dependence on H persists even when prediction from $\mathrm{conv}(\Pi)$ is allowed, covering the Bayesian mixture predictors considered in

Section 3.1. Theorem 4.3 establishes the same H-dependence for proper learners under full parameter sharing, despite the substantial reduction in class complexity.

Theorem 4.2 (Misspecified excess-KL lower bound for a decomposable class and convexified prediction)**.** *There exist absolute constants $c, c_0 > 0$ such that for any $M \geq 4$, any $n \geq c_0 \log M$, any horizon $H \in \mathbb{N}$, and any $G \in (0, 1]$, one can construct a finite decomposable policy class $\Pi = \Pi_0^H$ where $|\Pi_0| = M$ together with its induced trajectory-law class $\{P^\pi\}_{\pi\in\Pi}$ on a finite sample space. There also exists a family of misspecified trajectory laws $\{P^{\pi^\star} : \pi^\star \in \Pi^\star\}$ with $\pi^\star \notin \mathrm{conv}(\Pi)$ such that*

$$\inf_{\hat{\pi}\in\mathrm{conv}(\Pi)} \sup_{\pi^\star\in\Pi^\star} \mathbb{E}_{P^{\pi^\star}} \left[\mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\hat{\pi}}\Big) - \inf_{\pi\in\Pi} \mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\pi}\Big) \right] \geq c\, \frac{HG^2 \log M}{n}.$$

Theorem 4.3 (Misspecified excess-KL lower bound for a fully shared class and proper learners)**.** *There exist absolute constants $c, c_0 > 0$ such that for any $M \geq 4$, any $n \geq c_0 \log M$, any horizon $H \in \mathbb{N}$, and any $G \in (0, 1]$, one can construct a finite fully shared policy class $\Pi_{\mathrm{cons}} = \{\pi_1, \ldots, \pi_M\}$, together with its induced trajectory-law class $\{P^\pi : \pi \in \Pi_{\mathrm{cons}}\}$ on a finite sample space. There also exists a family of misspecified trajectory laws $\{P^{\pi^\star} : \pi^\star \in \Pi^\star\}$, with $P^{\pi^\star} \notin \mathrm{conv}\left(\{P^\pi : \pi \in \Pi_{\mathrm{cons}}\}\right)$ for every $\pi^\star \in \Pi^\star$, such that*

$$\inf_{\hat{\pi}\in\Pi_{\mathrm{cons}}} \sup_{\pi^\star\in\Pi^\star} \mathbb{E}_{(P^{\pi^\star})^{\otimes n}} \left[\mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\hat{\pi}}\Big) - \inf_{\pi\in\Pi_{\mathrm{cons}}} \mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\pi}\Big) \right] \geq c \frac{HG^2 \log M}{n}.$$

The infimum is over all possibly randomized measurable Π_{cons}-valued estimators; in particular, the result concerns proper learning.

Together, Theorems 4.2 and 4.3 show that the H-linear excess-KL cost is robust to both convexified prediction in the decomposable regime and complete parameter sharing under proper learning. Thus, within the proper-learning comparison of Table 1.1, sharing removes the statistical H-dependence under squared Hellinger distance but not under joint KL.

Why does full sharing help under squared Hellinger but not under joint KL? A central reason is that the squared Hellinger analysis exploits a special log-loss/Hellinger relationship at the level of the induced joint distributions (Rohatgi et al., 2025). Roughly, exponential-tail control of the log-likelihood ratio can be combined with the triangle inequality: for any comparator $\bar{\pi} \in \Pi$,

$$D^2_{\mathrm{H}}(P^{\hat{\pi}}, P^{\pi^\star}) \leq 2D^2_{\mathrm{H}}(P^{\hat{\pi}}, P^{\bar{\pi}}) + 2D^2_{\mathrm{H}}(P^{\bar{\pi}}, P^{\pi^\star}) \leq \frac{4\log(|\Pi|/\delta)}{n} + \frac{2}{n}\big(\hat{L}(P^{\bar{\pi}}) - \hat{L}(P^{\pi^\star})\big). \quad (9)$$

Here $\hat{L}$ denotes the empirical trajectory-level log-loss. Since this comparison is made at the level of the induced joint distributions, the resulting complexity term depends on the shared policy class as a whole, rather than on a uniform control over horizon-indexed conditional components. This is exactly why their analysis exhibits the *striking statistical gap between decomposable and fully shared classes*; see Table 1.1 of our paper.

However, this horizon-free behavior under full sharing does not generally carry over to joint KL. Under joint KL, the risk itself decomposes along the autoregressive sequence, so horizon-wise statistical accumulation can persist even under a fully-shared policy class. Our lower bound makes this point precise: even with full sharing, the estimation error can scale linearly with H. Thus, the H-dependence in our joint-KL bounds reflects a genuine statistical difficulty rather than an artifact of decomposable classes or a loose horizon-wise union bound.

Summary of error amplification under Hellinger and KL analyses. To summarize, Hellinger and KL exhibit different error amplification behaviors: Hellinger amplifies primarily through the approximation side, while KL typically shifts the horizon dependence to the estimation side. Since estimation error can be diluted by increasing the sample size n, this makes KL a particularly meaningful and practice-aligned metric despite its H-linear statistical accumulation in estimation. The following remark refines this picture by showing that beyond proper learners, full parameter sharing combined with trajectory-level aggregation can eliminate the statistical H-dependence.

Remark on improper aggregation. Full sharing combined with improper aggregation can remove the H-dependence. Applying Bayesian/exponential-weights aggregation directly to the induced trajectory laws yields a horizon-free remainder of order $\log|\Pi_0|/n$. The general finite-class result is given in Proposition 2.

5 From Joint KL Guarantees to Policy Learning Regret Bounds

Autoregressive generation can also be viewed as a policy-learning problem: a policy maps each prefix s_h to a distribution over the next token, and hence induces a distribution P^π over entire rollouts. This is the same object that appears in imitation learning (see Foster et al. (2024a) for more details), where one compares the rollout distribution of a learned policy with that of an expert policy. To make this connection explicit, let

$$J(\pi; r) := \mathbb{E}_{\tau \sim P^\pi}[r(\tau)]$$

denote the expected return of policy π under a non-negative trajectory-level reward function r. We call r R-bounded if $r(\tau) \in [0, R]$ for all trajectories τ, and for simplicity take $R = 1$. For the standard imitation-learning formulation, the worst-case, sequence-level regret can be bounded by joint TV as follows:

$$\sup_{r:\ 1\text{-bounded}} \{J(\pi^\star; r) - J(\widehat{\pi}; r)\} = D_{\mathrm{TV}}(P^{\pi^\star}, P^{\widehat{\pi}}).$$

Thus controlling the trajectory-level total variation distance is exactly equivalent to controlling the worst-case return gap over bounded rewards.

Our joint-KL guarantee fits directly into this policy-learning objective. Indeed, Pinsker's inequality immediately yields a rollout-control guarantee for the corresponding imitation-learning problem:

$$\sup_{r:\ 1\text{-bounded}} \{J(\pi^\star; r) - J(\widehat{\pi}; r)\} \le \sqrt{\frac{1}{2}\mathrm{KL}(P^{\pi^\star} \| P^{\widehat{\pi}})}.$$

Thus our bound provides a direct rollout-control guarantee for the corresponding policy-learning problem. At this level, joint KL yields the same $\sqrt{\cdot}$-order guarantee as squared Hellinger distance, since trajectory-level Hellinger control is also converted to total variation through a square-root comparison (see Theorem 3.1 of Foster et al. (2024a)).

This perspective clarifies that the joint-KL objective is not merely a next-token prediction surrogate: it is strong enough to control the rollout distribution induced by the learned policy. It also preserves the main estimation–approximation message of our analysis: approximation does not amplify with the horizon, while the linear horizon dependence in estimation is intrinsic under misspecification.

6 Conclusion

We studied misspecified autoregressive learning through the lens of joint KL divergence. We show that joint KL admits a horizon-free approximation factor, suggesting that the $\Omega(H)$ approximation blow-up under squared Hellinger distance can be metric-induced. At the same time, our lower bounds reveal a linear-in-H statistical barrier under joint KL, which persists even with convexified prediction for decomposable classes and with full parameter sharing for proper learners. We further develop computationally tractable stepwise Bayesian aggregation learning over a lifted policy space, achieving linear horizon scaling, and examine the role of aggregation beyond proper learning. Finally, we extend the autoregressive density estimation results to rollout-level policy learning guarantees. A natural direction for future work is to apply this framework to LLM-specific settings, including pre-training, in-context learning, and alignment, and to extend our results to transformer architectures and more realistic policy classes.

Acknowledgments and Disclosure of Funding

Yunbei Xu and Yuzhe Yuan were supported by the National University of Singapore [Start-Up Grant A-0010008-00-00]. Ruohan Zhan wishes to acknowledge the financial support of the Research and Materials Development Fund (RAMD_Zhan_R_26/27_3303) at London Business School.

References

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandit algorithms with supervised learning guarantees. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26, Fort Lauderdale, FL, USA. PMLR.

Ermon, S. (2023). Course notes in cs 236 deep generative modeling. https://deepgenerativemodels.github.io/syllabus.html.

Foster, D. J., Block, A., and Misra, D. (2024a). Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37:120602–120666.

Foster, D. J., Han, Y., Qian, J., and Rakhlin, A. (2024b). Online estimation via offline estimation: An information-theoretic framework.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Hazan, E., Shwartz, S. S., and Srebro, N. (2025). Research program: Theory of learning in dynamical systems.

Jiang, H. (2023). A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*.

Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 29–37, Fort Lauderdale, FL, USA. PMLR.

Mehta, N. (2017). Fast rates with high probability in exp-concave statistical learning. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1085–1093. PMLR.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.

Rohatgi, D., Block, A., Huang, A., Krishnamurthy, A., and Foster, D. J. (2025). Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification.

Sabbaghi, M., Pappas, G., Hassani, H., and Goel, S. (2024). Explicitly encoding structural symmetry is key to length generalization in arithmetic tasks. *arXiv preprint arXiv:2406.01895*.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.

van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1747–1756, New York, New York, USA. PMLR.

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

Wies, N., Levine, Y., and Shashua, A. (2023). The learnability of in-context learning.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2022). An explanation of in-context learning as implicit bayesian inference.

Yüksel, O. K. and Flammarion, N. (2025). Generalization bounds for autoregressive processes and in-context learning. In *EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM)*.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. (2023). What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*.

Contents

1 Introduction **1**

1.1 Main Results and Takeaways 3

2 Problem Setting **4**

3 Upper Bounds of Autoregressive Learning in Joint KL **5**

3.1 Joint-KL Guarantees: Constant Approximation and an Estimation Barrier 5

3.2 General Dependent Policy Classes 7

4 Lower Bounds of Autoregressive Learning in Joint KL **8**

5 From Joint KL Guarantees to Policy Learning Regret Bounds **10**

6 Conclusion **10**

A Proofs of Theorem 3.1 (Main Upper Bound) **13**

A.1 Proofs for step 1: prefix-conditioned control 14

A.2 Proofs for step 2: averaging over random prefixes 16

A.3 Proofs of Theorem 3.1 18

B Proofs in Section 4 (Main Lower Bounds) **19**

C Missing proofs in Section 3 **28**

D More Discussion on Squared Hellinger Distance **30**

E Technical Tools **32**

F Impact Statement and Limitations **33**

F.1 Broader Impacts 33

F.2 Limitations 33

A Proofs of Theorem 3.1 (Main Upper Bound)

We prove a generalization of Theorem 3.1 that allows for an infinite policy class via a log-ratio covering argument. The finite-class case is recovered by setting $r = 0$.

Definition A.1 (Log-ratio cover)**.** *Fix a policy class Π_0. For $r \geq 0$, we say that $\Pi' \subset \Pi_0$ is an r-cover of Π_0 if for each $\pi \in \Pi_0$, there is some $\pi' \in \Pi'$ with $\left|\log(\pi(u_h|s_h)/\pi'(u_h|s_h))\right| \leq r$ for any u_h and s_h. We write $\mathcal{N}_{\log}(\Pi_0, r)$ to denote the cardinality of the smallest r-cover of Π_0.*

The definition here is standard and widely used (Foster et al., 2021; Rohatgi et al., 2025).

The proof consists of two ingredients. The first is a *prefix-conditioned control*. Fix a step h and consider the random prefixes $\{s_h^i\}_{i=1}^n$ induced by the trajectories. The goal is to bound the empirical conditional KL of the learner in terms of any conditional comparator $\bar{\pi}_h$ within the class (up to a

complexity term), namely to control

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i)\,\|\,\hat{\pi}_h(\cdot \mid s_h^i)\big) \;-\; (1+\varepsilon)\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i)\,\|\,\bar{\pi}_h(\cdot \mid s_h^i)\big).$$

For ERM, this bound is proved in Lemma 2. For the Bayesian aggregation learner, the analogous prefix-conditioned bound is proved in Lemma 3.

The second ingredient is an *averaging over random prefixes* bound, which connects the empirical average over *visited* prefixes to the population quantity in the chain rule. Concretely, the goal is to control the deviation

$$\sum_{h=1}^{H}\mathbb{E}_{s_h\sim P^{\pi^\star}}\Big[\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\pi_h(\cdot \mid s_h)\big)\Big] \;-\; (1+\varepsilon)\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i)\,\|\,\pi_h(\cdot \mid s_h^i)\big),$$

uniformly over π in the policy class. This step is shared by both ERM and the Bayesian aggregation and is proved in Lemma 4. Combining the two ingredients and, moreover, using a Freedman-type deviation bound (Lemma 9) to control the empirical-to-population gap

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i)\,\|\,\bar{\pi}_h(\cdot \mid s_h^i)\big) \;-\; (1+\varepsilon)\mathbb{E}_{S\sim P^{\pi^\star}}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\bar{\pi}_h(\cdot \mid s_h)\big),$$

yields Theorem 3.1.

A.1 Proofs for step 1: prefix-conditioned control

Lemma 2 (Prefix-conditioned control for ERM)**.** *Let $S_1,\dots,S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ with $\pi^\star \notin \Pi = \Pi_0^H$ or Π_{cons}, where Π_0 has a finite log-ratio cover. Under Assumption 1, ERM algorithm guarantees that for any $\delta \in (0,1)$ and any $\varepsilon > 0$, with probability at least $1-\delta$, for all $\bar{\pi}_h$:*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_{1:h-1}^i)\,\|\,\hat{\pi}_h(\cdot \mid s_{1:h-1}^i)\big) \le (1+\varepsilon)\,\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_{1:h-1}^i)\,\|\,\bar{\pi}_h(\cdot \mid s_{1:h-1}^i)\big) + \frac{C}{\varepsilon}\,H(G+1)\,\frac{\log\big(\mathcal{N}_{\log}(\Pi_0,r)\,\delta^{-1}\big)}{n} + C'Hr,$$

for universal constants C, C'.

Proof of Lemma 2: Fix $r > 0$ and let $\Pi_r \subset \Pi$ be an r-cover in the sense of Definition A.1. For each $\pi \in \Pi$, pick a representative $\pi^r \in \Pi_r$ such that for every length-h trajectory $(u_h, s_{1:h-1})$,

$$\Big|\log\frac{\pi(u_h \mid s_{1:h-1})}{\pi^r(u_h \mid s_{1:h-1})}\Big| \le r. \tag{10}$$

We prove the result for a single step h and then sum them up. Fix $h \in [H]$ and condition on the visited prefixes $\{s_{1:h-1}^i\}_{i=1}^n$. For any $\pi \in \Pi_0$ and $i \in [n]$, define

$$Y_i(\pi) := \log\frac{\pi_h^\star(u_h^i \mid s_{1:h-1}^i)}{\pi(u_h^i \mid s_{1:h-1}^i)},$$

$$\mu_\pi := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{u_h\sim\pi_h^\star(\cdot|s_{1:h-1}^i)}\Big[\log\frac{\pi_h^\star(u_h \mid s_{1:h-1}^i)}{\pi(u_h \mid s_{1:h-1}^i)}\Big] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_{1:h-1}^i)\,\|\,\pi(\cdot \mid s_{1:h-1}^i)\big),$$

$$\hat{\mu}_\pi := \frac{1}{n}\sum_{i=1}^{n}Y_i(\pi).$$

Let $\bar{\pi}$ be any fixed policy in the class, and denote by $\bar{\pi}^r \in \Pi_r$ its representative.

We use Lemma 7 to obtain the Bernstein condition. From $\mathbb{E}[e^{-Y_i}] = 1$ and $|Y_i| \le G$, we have $\mathrm{Var}(Y_i) \le V(G)\,\mathbb{E}[Y_i]$ with $V(G) = 4 + 4G$. Hence, by the two-sided Bernstein inequality, for any

$\delta_g \in (0,1)$, with probability at least $1-\delta_g$,

$$|\mu - \hat{\mu}| \le \sqrt{\frac{2V(G)\,\mu\,\log(1/\delta_g)}{n}} + \frac{4G}{3}\,\frac{\log(1/\delta_g)}{n}. \tag{11}$$

Apply (11) simultaneously to all $\pi' \in \Pi_r$ and to

$$\hat{\pi}_r \in \arg\min_{\pi' \in \Pi_r} \hat{\mu}_{\pi'}.$$

By a union bound with $\delta_g = \delta/|\Pi_r|$, with probability at least $1-\delta$ the following holds for all $\pi' \in \Pi_r$:

$$\Delta_r(\pi') \le \widehat{\Delta}_r(\pi') + a_r\big(\sqrt{\mu_{\pi'}} + \sqrt{\mu_{\bar{\pi}^r}}\big) + 2b_r, \tag{12}$$

where

$$\Delta_r(\pi') := \mu_{\pi'} - \mu_{\bar{\pi}^r}, \qquad \widehat{\Delta}_r(\pi') := \hat{\mu}_{\pi'} - \hat{\mu}_{\bar{\pi}^r},$$

and

$$a_r := \sqrt{\frac{2V(G)\,\log(|\Pi_r|\delta^{-1})}{n}}, \qquad b_r := \frac{4G}{3}\,\frac{\log(|\Pi_r|\delta^{-1})}{n}.$$

By (10), we have

$$|Y_i(\pi) - Y_i(\pi^r)| \le r \quad \Rightarrow \quad |\mu_\pi - \mu_{\pi^r}| \le r, \qquad |\hat{\mu}_\pi - \hat{\mu}_{\pi^r}| \le r. \tag{13}$$

Since $\hat{\pi}$ is an ERM over Π_0 and $\Pi_r \subset \Pi_0$, (13) implies

$$\min_{\pi' \in \Pi_r} \hat{\mu}_{\pi'} \ge \min_{\pi \in \Pi} \hat{\mu}_\pi - r = \hat{\mu}_{\hat{\pi}} - r, \qquad \hat{\mu}_{\hat{\pi}^r} \le \hat{\mu}_{\hat{\pi}} + r,$$

and therefore

$$\widehat{\Delta}_r(\hat{\pi}^r) = \hat{\mu}_{\hat{\pi}^r} - \hat{\mu}_{\bar{\pi}^r} \le 2r. \tag{14}$$

Applying (12) with $\pi' = \hat{\pi}^r$ and using $\mu_{\hat{\pi}^r} = \Delta_r(\hat{\pi}^r) + \mu_{\bar{\pi}^r}$ together with $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, we obtain

$$\Delta_r(\hat{\pi}^r) \le a_r\big(\sqrt{\Delta_r(\hat{\pi}^r)} + 2\sqrt{\mu_{\bar{\pi}^r}}\big) + 2b_r + 2r. \tag{15}$$

Solving $x \le a_r\sqrt{x} + 2a_r\sqrt{\mu_{\bar{\pi}^r}} + 2b_r + 2r$ yields

$$\Delta_r(\hat{\pi}^r) \le 2a_r^2 + 4a_r\sqrt{\mu_{\bar{\pi}^r}} + 4b_r + 4r.$$

By (13) and $\Pi_r \subset \Pi_0$, we have $\mu_{\bar{\pi}} \le \mu_{\bar{\pi}^r} \le \mu_{\bar{\pi}} + r$. Moreover,

$$\Delta(\hat{\pi}) := \mu_{\hat{\pi}} - \mu_{\bar{\pi}} \le \mu_{\hat{\pi}^r} - \mu_{\bar{\pi}^r} + 2r = \Delta_r(\hat{\pi}^r) + 2r.$$

Combining the previous display with $|\Pi_r| = \mathcal{N}_{\log}(\Pi_0, r)$ gives

$$\Delta(\hat{\pi}) \le 4\,\frac{V(G)\,\log\big(\mathcal{N}_{\log}(\Pi_0,r)\,\delta^{-1}\big)}{n} + 4\sqrt{2}\,\sqrt{\frac{V(G)\,\mu_{\bar{\pi}^r}\,\log\big(\mathcal{N}_{\log}(\Pi_0,r)\,\delta^{-1}\big)}{n}} + \frac{16}{3}\,G\,\frac{\log\big(\mathcal{N}_{\log}(\Pi_0,r)\,\delta^{-1}\big)}{n} + 6r.$$

Finally, apply $2\sqrt{xy} \le \varepsilon x + \frac{y}{\varepsilon}$ to the square-root term with $x = \mu_{\bar{\pi}^r}$ and $y = 8V(G)\frac{\log(\mathcal{N}_{\log}(\Pi_0,r)\delta^{-1})}{n}$, and use $\mu_{\bar{\pi}^r} \le \mu_{\bar{\pi}} + r$. Thus for any $\varepsilon > 0$,

$$\frac{1}{n}\sum_{i=1}^n \mathrm{KL}\big(\pi_h^\star(\cdot \mid s^i_{1:h-1}) \,\|\, \hat{\pi}(\cdot \mid s^i_{1:h-1})\big) = \mu_{\bar{\pi}} + \Delta(\hat{\pi}) \le (1+\varepsilon)\,\frac{1}{n}\sum_{i=1}^n \mathrm{KL}\big(\pi_h^\star(\cdot \mid s^i_{1:h-1}) \,\|\, \bar{\pi}(\cdot \mid s^i_{1:h-1})\big) + \frac{C}{\varepsilon}\,(G+1)\,\frac{\log\big(\mathcal{N}_{\log}(\Pi_0,r)\,\delta^{-1}\big)}{n} + C'r,$$

□

Lemma 3 (Prefix-conditioned control for Bayesian aggregation)**.** *Let $S_1, \dots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ with $\pi^\star \notin \Pi = \Pi_0^H$ or Π_{cons}, where Π_0 has a finite log-ratio cover. Under Assumption 1, Bayesian aggregation algorithm guarantees that for any $\varepsilon > 0$, we have that,*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}(\pi_h^\star(\cdot\,|\,s_{1:h-1}^i)\|\hat{\pi}_h(\cdot\,|\,s_{1:h-1}^i)) \le \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}(\pi_h^\star(\cdot\,|\,s_{1:h-1}^i)\|\bar{\pi}_h(\cdot\,|\,s_{1:h-1}^i)) + Hr + \frac{H\log\mathcal{N}_{\log}(\Pi_0,r)}{n}.$$

Proof of Lemma 3. Let $S_i = (s_1^i, u_1^i, \ldots, s_H^i, u_H^i)$ be i.i.d. trajectories from $P^{\pi^\star}$ for $i = 1, \ldots, n$. Fix $r > 0$. Let $\Pi_{h,r} \subset \Pi_h$ be an r-cover under the log-ratio metric, i.e., for every $\pi_h \in \Pi_h$ there exists $\pi_h^r \in \Pi_{h,r}$ such that for all $(u, s_{1:h-1})$,

$$\left|\log\frac{\pi_h(u\,|\,s_{1:h-1})}{\pi_h^r(u\,|\,s_{1:h-1})}\right| \le r.$$

Define the per-step Bayesian (exponential-weights) posterior predictor on $\Pi_{h,r}$ with a uniform prior by

$$\hat{\pi}_h(\cdot\,|\,s_{1:h-1}) = \sum_{\pi_h^r\in\Pi_{h,r}} q_h(\pi_h^r)\,\pi_h^r(\cdot\,|\,s_{1:h-1}), \qquad P^{\hat{\pi}}(S) \;=\; \prod_{h=1}^{H}\hat{\pi}_h(u_h\,|\,s_{1:h-1}).$$

By log-loss mixability, for any fixed data $S_{1:n}$ and any $\pi_h^r \in \Pi_{h,r}$,

$$\sum_{i=1}^{n}\big[-\log\hat{\pi}_h(u_h^i\,|\,s_{1:h-1}^i)\big] \;\le\; \sum_{i=1}^{n}\big[-\log\pi_h^r(u_h^i\,|\,s_{1:h-1}^i)\big] \;+\; \log\mathcal{N}_{\log}(\Pi_h,r). \tag{16}$$

Summing (16) over $h = 1, \ldots, H$ and adding $\sum_{i,h}\log\pi_h^\star(u_h^i\,|\,s_{1:h-1}^i)$ to both sides gives

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\log\frac{\pi_h^\star(u_h^i\,|\,s_{1:h-1}^i)}{\hat{\pi}_h(u_h^i\,|\,s_{1:h-1}^i)} \le \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\log\frac{\pi_h^\star(u_h^i\,|\,s_{1:h-1}^i)}{\pi_h^r(u_h^i\,|\,s_{1:h-1}^i)} \;+\; \frac{1}{n}\sum_{h=1}^{H}\log\mathcal{N}_{\log}(\Pi_h,r). \tag{17}$$

For any $\pi \in \Pi$ choose its cover representatives $\{\pi_h^r\}_{h=1}^H \subset \Pi_{h,r}$. By the cover property,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\log\frac{\pi_h(u_h^i\,|\,s_{1:h-1}^i)}{\pi_h^r(u_h^i\,|\,s_{1:h-1}^i)}\right| \;\le\; Hr.$$

Since (17) works for any possible sequence, averaging both sides yields

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}(\pi_h^\star(\cdot\,|\,s_{1:h-1}^i)\|\hat{\pi}_h(\cdot\,|\,s_{1:h-1}^i)) \le \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\mathrm{KL}(\pi_h^\star(\cdot\,|\,s_{1:h-1}^i)\|\bar{\pi}_h(\cdot\,|\,s_{1:h-1}^i)) + Hr + \frac{H\log\mathcal{N}_{\log}(\Pi_0,r)}{n}. \tag{18}$$

□

A.2 Proofs for step 2: averaging over random prefixes

Lemma 4 (Averaging over random prefixes)**.** *Let Π_0 be an arbitrary class with Π_{cons} being the induced fully shared class with horizon H. Let $\Pi \in \{\Pi_0^H, \Pi_{\text{cons}}\}$ and $S_i = (u_1^i, \ldots, u_H^i)$ be i.i.d. trajectories drawn from $P^{\pi^\star}$. For any $\pi \in \Pi$ and trajectory $S = (u_1, \ldots, u_H)$, define*

$$\Delta_\pi(S) := \mathrm{KL}(P^{\pi^\star}\,\|\,P^{\pi}) - (1+\varepsilon)\sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^\star(\cdot\mid s_h)\,\|\,\pi_h(\cdot\mid s_h)\big), \qquad s_h = (u_1, \ldots, u_{h-1}). \tag{19}$$

Fix any $r > 0$, and let $\Pi_{0,r} \subset \Pi_0$ be an r-cover in the sense of Definition A.1, namely for each $\pi_0 \in \Pi_0$ there exists $\pi_0^r \in \Pi_{0,r}$ such that for all (u, s),

$$\left|\log\frac{\pi_0(u\mid s)}{\pi_0^r(u\mid s)}\right| \le r.$$

Write $\mathcal{N}_{\log}(\Pi_0, r) := |\Pi_{0,r}|$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $\pi \in \Pi$,

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_\pi(S_i) \;\le\; \frac{(1+\varepsilon)^2 HG}{\varepsilon n}\log\Big(\frac{2H\,\mathcal{N}_{\log}(\Pi_0,r)}{\delta}\Big) \;+\; (1+\varepsilon)Hr. \tag{20}$$

For fully shared class Π_{cons}, we have that for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $\pi \in \Pi_{\text{cons}}$,

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_\pi(S_i) \;\le\; \frac{(1+\varepsilon)^2 HG}{\varepsilon n}\log\Big(\frac{2\mathcal{N}_{\log}(\Pi_0,r)}{\delta}\Big) \;+\; (1+\varepsilon)Hr. \tag{21}$$

In particular, the same bound holds for any data-dependent $\hat{\pi} \in \Pi$. For the fully-shared class, the same extension holds for the bound in (21)*.*

Proof of Lemma 4: We first prove the result for decomposable class. Fix $h \in \{1,\dots,H\}$ and $\pi_h \in \Pi_0$. For $i = 1,\dots,n$ define

$$Y_{h,i}(\pi_h) := \mathrm{KL}\big(\pi_h^\star(\cdot \mid S_h^i) \,\|\, \pi_h(\cdot \mid S_h^i)\big), \qquad S_h^i = (u_1^i,\dots,u_{h-1}^i).$$

By Assumption 1, $0 \le Y_{h,i}(\pi_h) \le G$ almost surely. Let $\mathcal{F}_i$ be the natural filtration of $(S_1,\dots,S_i)$. Then $Y_{h,i}(\pi_h)$ is $\mathcal{F}_i$-measurable, and by independence together with the chain rule for KL,

$$\mathbb{E}_{i-1}\big[Y_{h,i}(\pi_h)\big] = \mathbb{E}_{S_h}\mathrm{KL}\big(\pi_h^\star(\cdot \mid S_h) \,\|\, \pi_h(\cdot \mid S_h)\big),$$

where S_h denotes a generic prefix distributed according to $P^{\pi^\star}$.

Apply Lemma 9 to the sequence $X_i := Y_{h,i}(\pi_h) \in [0,G]$. With probability at least $1-\delta_{h,\pi_h}$,

$$\sum_{i=1}^{n}\mathbb{E}_{i-1}[Y_{h,i}(\pi_h)] \le (1+\varepsilon)\sum_{i=1}^{n}Y_{h,i}(\pi_h) + \frac{(1+\varepsilon)^2 G}{\varepsilon}\log\Big(\frac{2}{\delta_{h,\pi_h}}\Big),$$

and hence after dividing by n and rearranging,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}_{i-1}[Y_{h,i}(\pi_h)] - (1+\varepsilon)Y_{h,i}(\pi_h)\Big) \le \frac{(1+\varepsilon)^2 G}{\varepsilon n}\log\Big(\frac{2}{\delta_{h,\pi_h}}\Big). \tag{22}$$

Now apply (22) to all $h \in [H]$ and all representatives $\pi_h^r \in \Pi_{0,r}$, and take a union bound with $\delta_{h,\pi_h^r} = \delta/(H\mathcal{N}_{\log}(\Pi_0,r))$. With probability at least $1-\delta$, for all h and all $\pi_h^r \in \Pi_{0,r}$,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}_{i-1}[Y_{h,i}(\pi_h^r)] - (1+\varepsilon)Y_{h,i}(\pi_h^r)\Big) \le \frac{(1+\varepsilon)^2 G}{\varepsilon n}\log\Big(\frac{2H\mathcal{N}_{\log}(\Pi_0,r)}{\delta}\Big). \tag{23}$$

Fix any $\pi = (\pi_1,\dots,\pi_H) \in \widetilde{\Pi}$ and let $\pi^r = (\pi_1^r,\dots,\pi_H^r)$ be its coordinatewise representatives in $\Pi_{0,r}$. By the log-cover property, for every prefix s,

$$\Big|\mathrm{KL}\big(\pi_h^\star(\cdot \mid s) \,\|\, \pi_h(\cdot \mid s)\big) - \mathrm{KL}\big(\pi_h^\star(\cdot \mid s) \,\|\, \pi_h^r(\cdot \mid s)\big)\Big| = \Big|\mathbb{E}_{u\sim\pi_h^\star(\cdot|s)}\log\frac{\pi_h^r(u \mid s)}{\pi_h(u \mid s)}\Big| \le r.$$

Consequently, for each trajectory S,

$$\Big|\sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h) \,\|\, \pi_h(\cdot \mid s_h)\big) - \sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h) \,\|\, \pi_h^r(\cdot \mid s_h)\big)\Big| \le Hr.$$

The same argument applies to the chain-rule expression of the joint KL:

$$\mathrm{KL}(P^{\pi^\star} \,\|\, P^{\pi}) = \sum_{h=1}^{H}\mathbb{E}_{S_h}\mathrm{KL}\big(\pi_h^\star(\cdot \mid S_h) \,\|\, \pi_h(\cdot \mid S_h)\big),$$

hence

$$\big|\mathrm{KL}(P^{\pi^\star} \,\|\, P^{\pi}) - \mathrm{KL}(P^{\pi^\star} \,\|\, P^{\pi^r})\big| \le Hr.$$

Combining the two displays with the definition (19) yields

$$\Delta_\pi(S) \le \Delta_{\pi^r}(S) + (1+\varepsilon)Hr \qquad \text{for all trajectories } S. \tag{24}$$

Finally, sum (23) over $h = 1, \dots, H$ and use $\sum_{h=1}^{H} \sum_{i=1}^{n} \mathbb{E}_{i-1}[Y_{h,i}(\pi_h^r)] = n\,\mathrm{KL}(P^{\pi^\star} \,\|\, P^{\pi^r})$ (chain rule) to obtain

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_{\pi^r}(S_i) \le \frac{(1+\varepsilon)^2 HG}{\varepsilon n} \log\Big(\frac{2H\mathcal{N}_{\log}(\Pi_0, r)}{\delta} \Big).$$

Together with (24), this proves (20). Since the bound holds uniformly over $\pi \in \widetilde{\Pi}$, it also holds for any data-dependent $\hat{\pi} \in \widetilde{\Pi}$. Finally, we explain how the uniform bound applies to the Bayesian mixture predictor. Let

$$Q = \bigotimes_{h=1}^{H} q_h$$

be any data-dependent product distribution on $\Pi = \Pi_0^H$, and let $\widetilde{\pi} \sim Q$. Since the preceding event holds simultaneously for every $\pi \in \Pi$, averaging over Q yields

$$\begin{aligned} \mathbb{E}_{\widetilde{\pi} \sim Q} \mathrm{KL}(P^{\pi^\star} \| P^{\widetilde{\pi}}) \le (1+\varepsilon) \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\widetilde{\pi} \sim Q} \left[\sum_{h=1}^{H} \mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i) \big\| \widetilde{\pi}_h(\cdot \mid s_h^i) \big) \right] \\ + \frac{(1+\varepsilon)^2 HG}{\varepsilon n} \log\left(\frac{2H\mathcal{N}_{\log}(\Pi_0, r)}{\delta} \right) + (1+\varepsilon) Hr. \end{aligned}$$

Now define the posterior mixture predictor

$$\widehat{\pi}_h(\cdot \mid s) := \mathbb{E}_{\widetilde{\pi}_h \sim q_h} \big[\widetilde{\pi}_h(\cdot \mid s) \big].$$

By convexity of KL divergence in its second argument,

$$\mathrm{KL}(\pi_h^\star(\cdot \mid s) \| \widehat{\pi}_h(\cdot \mid s)) \le \mathbb{E}_{\widetilde{\pi}_h \sim q_h} \mathrm{KL}(\pi_h^\star(\cdot \mid s) \| \widetilde{\pi}_h(\cdot \mid s)) \,.$$

Applying the KL chain rule and averaging over the random prefix therefore gives

$$\mathrm{KL}(P^{\pi^\star} \| P^{\widehat{\pi}}) \le \mathbb{E}_{\widetilde{\pi} \sim Q} \mathrm{KL}(P^{\pi^\star} \| P^{\widetilde{\pi}}).$$

Thus the same concentration event controls the improper Bayesian mixture through the posterior average of its proper components.

For fully shared class, just take $\delta_{h,\pi_h^r} = \delta / (\mathcal{N}_{\log}(\Pi_0, r))$ in the same argument. The resulting event is again uniform over the shared class, so the same posterior-averaging argument yields the corresponding extension of (20). □

A.3 Proofs of Theorem 3.1

Proposition 1 (Theorem 3.1, generalization to infinite classes)**.** *Let $S_1, \dots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ with $\pi^\star \notin \Pi$, where $\Pi = \Pi_0^H$ is a decomposable model class and Π_0 has a finite log-ratio cover at scale $r \ge 0$. Under Assumption 1, ERM algorithm* (3) *and Bayesian aggregation algorithm* (4) *guarantee that for any $\delta \in (0,1)$ and any $\varepsilon > 0$, with probability at least $1 - \delta$ and for a universal constant $C > 0$,*

$$\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \le \underbrace{(1+\varepsilon) \min_{\pi \in \Pi} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})}_{\textit{approximation error}} + \underbrace{C \, \frac{(1+\varepsilon)^{4/3}}{\varepsilon} \, \frac{(HG+1) \log(H\mathcal{N}_{\log}(\Pi_0, r)\delta^{-1})}{n} + (1+\varepsilon) Hr}_{\textit{estimation error}}. \tag{25}$$

For fully shared model class Π_{cons}, the result would be

$$\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \le \underbrace{(1+\varepsilon) \min_{\pi \in \Pi_{\text{cons}}} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})}_{\textit{approximation error}} + \underbrace{C \, \frac{(1+\varepsilon)^{4/3}}{\varepsilon} \, \frac{(HG+1) \log(\mathcal{N}_{\log}(\Pi_0, r)\delta^{-1})}{n} + (1+\varepsilon) Hr}_{\textit{estimation error}}. \tag{26}$$

Proof of Proposition 1: We focus on the ERM algorithm under the decomposable class, and the proof for the Bayesian aggregation or the fully shared class are similar. Having had Lemma 2 and

Lemma 4, the only term we need to control is

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i)\,\|\,\bar{\pi}_h(\cdot \mid s_h^i)\big) \;-\; (1+\varepsilon)\mathbb{E}_{S\sim P^{\pi^\star}}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\bar{\pi}_h(\cdot \mid s_h)\big). \tag{27}$$

Fix $h \in [H]$ and $r > 0$, and let $\Pi_{0,r} \subset \Pi_0$ be an r-log-ratio cover. Pick $\tilde{\pi}_h \in \Pi_0$ such that

$$\mathbb{E}_{s_h\sim P^{\pi^\star}}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\tilde{\pi}_h(\cdot \mid s_h)\big) \le \inf_{\pi_h\in\Pi_0}\mathbb{E}_{s_h\sim P^{\pi^\star}}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\pi_h(\cdot \mid s_h)\big) + r,$$

which exists by the definition of the infimum. By the cover property, there is $\tilde{\pi}_h^r \in \Pi_{0,r}$ such that for all (u, s), $\big|\log(\tilde{\pi}_h(u \mid s)/\tilde{\pi}_h^r(u \mid s))\big| \le r$, and hence for every prefix s,

$$\Big|\mathrm{KL}\big(\pi_h^\star(\cdot \mid s)\,\|\,\tilde{\pi}_h(\cdot \mid s)\big) - \mathrm{KL}\big(\pi_h^\star(\cdot \mid s)\,\|\,\tilde{\pi}_h^r(\cdot \mid s)\big)\Big| = \Big|\mathbb{E}_{u\sim\pi_h^\star(\cdot|s)}\log\frac{\tilde{\pi}_h^r(u \mid s)}{\tilde{\pi}_h(u \mid s)}\Big| \le r.$$

In particular,

$$\mathbb{E}_{s_h\sim P^{\pi^\star}}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\tilde{\pi}_h^r(\cdot \mid s_h)\big) \le \inf_{\pi_h\in\Pi_0}\mathbb{E}_{s_h\sim P^{\pi^\star}}\mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\pi_h(\cdot \mid s_h)\big) + 2r.$$

Define $Y_i := \mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h^i)\,\|\,\tilde{\pi}_h^r(\cdot \mid s_h^i)\big)$ and $\mathcal{F}_i := \sigma(S_1, \ldots, S_i)$. By Assumption 1, $0 \le Y_i \le G$ a.s., and $\mathbb{E}_{i-1}[Y_i] = \mathbb{E}_{s_h\sim P^{\pi^\star}}Y$ where $Y = \mathrm{KL}\big(\pi_h^\star(\cdot \mid s_h)\,\|\,\tilde{\pi}_h^r(\cdot \mid s_h)\big)$. Applying Lemma 9 yields, with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n}Y_i - (1+\varepsilon)\mathbb{E}_{s_h\sim P^{\pi^\star}}Y \le \frac{(1+\varepsilon)^2 G}{\varepsilon n}\log(\delta^{-1}).$$

Combining with the choice of $\tilde{\pi}_h^r$ gives (27) with the comparator term written as $\inf_{\pi_h\in\Pi_0}\mathbb{E}_{s_h\sim P^{\pi^\star}}\mathrm{KL}(\pi_h^\star\|\pi_h)$, up to an additional r-approximation term. Combining Lemma 2, Lemma 4 and (27), we have that

$$\mathrm{KL}(P^{\pi^\star}\|P^{\hat{\pi}}) \lesssim (1+\varepsilon)^3\min_{\pi\in\Pi}\mathrm{KL}(P^{\pi^\star}\|P^{\pi}) + \frac{(1+\varepsilon)^2}{\varepsilon}\frac{(HG+1)\log(H\mathcal{N}_{\log}(\Pi_0, r)\delta^{-1})}{n} + (1+\varepsilon)^3 Hr. \tag{28}$$

Now the preceding bound is of the form

$$(1+\varepsilon)^3 A \;+\; \frac{(1+\varepsilon)^2}{\varepsilon}\,B \;+\; (1+\varepsilon)^3 Hr,$$

where $A = \min_{\pi\in\Pi}\mathrm{KL}(P^{\pi^\star}\|P^{\pi})$ and $B = \frac{(HG+1)\log(H\mathcal{N}_{\log}(\Pi_0,r)\delta^{-1})}{n}$. To rewrite the leading factor on A as $(1+\varepsilon)$, set $\tilde{\varepsilon} := (1+\varepsilon)^{1/3} - 1$ so that $(1+\tilde{\varepsilon})^3 = 1+\varepsilon$. Substituting $\tilde{\varepsilon}$ for ε preserves the $1/\varepsilon$ order in the estimation term since

$$\frac{(1+\tilde{\varepsilon})^2}{\tilde{\varepsilon}} = \frac{(1+\varepsilon)^{2/3}}{(1+\varepsilon)^{1/3}-1} \lesssim \frac{(1+\varepsilon)^{4/3}}{\varepsilon},$$

up to universal constants. Consequently, we obtain the equivalent (up to constants) decomposition

$$\mathrm{KL}(P^{\pi^\star}\|P^{\hat{\pi}}) \lesssim \underbrace{(1+\varepsilon)\min_{\pi\in\Pi}\mathrm{KL}(P^{\pi^\star}\|P^{\pi})}_{\text{approximation error}} + \underbrace{\frac{(1+\varepsilon)^{4/3}}{\varepsilon}\frac{(HG+1)\log(H\mathcal{N}_{\log}(\Pi_0, r)\delta^{-1})}{n} + (1+\varepsilon)Hr}_{\text{estimation error}}.$$

Following the same proof procedure, the same bound can be stated with an $(1+\varepsilon)$ prefactor in the estimation term for the Bayesian aggregation, i.e., replacing $(1+\varepsilon)^{4/3}$ by $(1+\varepsilon)$ (up to universal constants). In the main paper, we state the bounds for both ERM and the Bayesian aggregation with the enlarged coefficient $(1+\varepsilon)^2$, since for practical choices of ε this is just a numerical constant and not a key ingredient of the theorem. □

B Proofs in Section 4 (Main Lower Bounds)

We establish the main misspecified lower bounds by first constructing hard instances in the realizable setting. For decomposable classes, Theorem B.1 combines a packing argument and Fano's inequality with posterior-predictive KL projection and tensorization across the horizon, yielding a lower bound even for unrestricted trajectory-law estimators. For fully shared classes, Theorem B.2 uses an

observed-baseline context construction and Fano's inequality to establish the statistical difficulty of proper model selection despite complete parameter sharing. We then leverage these realizable constructions to establish the corresponding misspecified excess-KL lower bounds for decomposable classes under convexified prediction (Theorem 4.2) and fully shared classes under proper learning (Theorem 4.3).

Theorem B.1 (Realizable lower bound for a decomposable class and unrestricted estimators)**.** *There exist absolute constants $c, c_0 > 0$ such that for any $M \geq 4$, any $n \geq c_0 \log M$, any horizon $H \in \mathbb{N}$, and any $G \in (0, 1]$, one can construct a finite decomposable policy class $\Pi = \Pi_0^H$ where $|\Pi_0| = M$ and associated distributions $\{P^\pi\}_{\pi \in \Pi}$ on a finite sample space such that for every estimator $\hat{\pi} = \hat{\pi}(S_{1:n})$ based on $S_1, \ldots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ with $\pi^\star \in \Pi$,*

$$\inf_{\hat{\pi}} \sup_{\pi^\star \in \Pi} \mathbb{E}_{P^{\pi^\star}} \left[\mathrm{KL}\Big(P^{\pi^\star} \,\Big\|\, P^{\hat{\pi}} \Big) \right] \;\geq\; c \, \frac{H \, G^2 \, \log |\Pi_0|}{n}. \tag{29}$$

Here the infimum is over all possibly randomized measurable estimators that output an arbitrary probability law on the trajectory space. In particular, $\hat{\pi}$ need not belong to the product family, factor coordinatewise, or select a candidate.

Proof of Theorem B.1: A balanced Varshamov–Gilbert lemma gives universal constants $c_1, C_1 > 0$ such that, for every $M \geq 4$, there are an even integer $d \leq C_1 \log M$ and balanced vectors

$$v^{(1)}, \ldots, v^{(M)} \in \{-1, +1\}^d, \qquad \sum_{x=1}^{d} v_x^{(j)} = 0,$$

whose pairwise Hamming distances satisfy

$$D\big(v^{(j)}, v^{(k)}\big) \geq c_1 d, \qquad j \neq k.$$

This is the usual greedy packing of the middle slice of the binary cube; the middle slice has cardinality $\binom{d}{d/2}$ and a Hamming ball of any fixed relative radius occupies at most an exponentially smaller fraction of it.

Let $a > 0$ be a sufficiently small universal constant and set

$$\alpha = aG\sqrt{\frac{\log M}{n}}.$$

Increasing the universal constant c_0 in the theorem if necessary ensures $\alpha \leq \min\{1/4, G/4\}$. On the sample space $[d]$, define

$$q_j(x) := \frac{1 + \alpha v_x^{(j)}}{d}, \qquad j \in [M], \quad x \in [d], \tag{30}$$

and let $\mathcal{Q}_0 = \{q_1, \ldots, q_M\}$. Balance makes every q_j a probability density. For all j, k, x,

$$\left| \log \frac{q_j(x)}{q_k(x)} \right| \leq \log \frac{1 + \alpha}{1 - \alpha} \leq 3\alpha \leq G,$$

so the class satisfies Assumption 1 whichever member is the truth.

If $j \neq k$, balance implies that the two types of mismatched signs occur equally often. Therefore

$$\mathrm{KL}(q_j \| q_k) = \frac{D(v^{(j)}, v^{(k)})}{d} \, \alpha \log \frac{1 + \alpha}{1 - \alpha}. \tag{31}$$

Since $2\alpha \leq \log((1 + \alpha)/(1 - \alpha)) \leq 3\alpha$ for $\alpha \leq 1/4$, it follows that

$$2c_1\alpha^2 \leq \mathrm{KL}(q_j \| q_k) \leq 3\alpha^2, \qquad j \neq k. \tag{32}$$

We first consider one coordinate. Let J be uniform on $[M]$ and, given $J = j$, observe $Y^1, \ldots, Y^n \overset{\text{iid}}{\sim} q_j$. Convexity of KL in its second argument gives the standard mutual-information bound

$$I(J; Y^{1:n}) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} \mathrm{KL}(q_j^{\otimes n} \| q_k^{\otimes n}) \leq 3n\alpha^2 \leq \frac{1}{4} \log M,$$

where the last inequality holds by choosing a sufficiently small and using $G \le 1$. Fano's inequality then implies, for every possibly randomized estimator $\widehat{J}$,

$$\Pr(\widehat{J} \neq J) \ge 1 - \frac{I(J; Y^{1:n}) + \log 2}{\log M} \ge \frac{1}{4},$$

where the last step uses $M \ge 4$.

For the H-coordinate problem, draw $J_1, \dots, J_H$ independently and uniformly from $[M]$, and put

$$P_J := \bigotimes_{h=1}^{H} q_{J_h}.$$

Write $\mathcal{D}_h = (Y_h^1, \dots, Y_h^n)$ and $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_H)$. Under the product prior and product experiment, the posterior factorizes:

$$\mathcal{L}(J \mid \mathcal{D}) = \bigotimes_{h=1}^{H} \mathcal{L}(J_h \mid \mathcal{D}_h).$$

We first lower-bound the unrestricted one-coordinate Bayes risk. Define the posterior mean codeword and posterior predictive density

$$m_h := \mathbb{E}[v^{(J_h)} \mid \mathcal{D}_h] \in [-1, 1]^d, \qquad \overline{q}_h(x) := \mathbb{E}[q_{J_h}(x) \mid \mathcal{D}_h] = \frac{1 + \alpha (m_h)_x}{d}.$$

Let $\widetilde{J}_h$ be a nearest Euclidean codeword to m_h, with measurable tie breaking. If $\widetilde{J}_h \neq J_h$, packing separation and nearestness give

$$2\sqrt{c_1 d} \le \|v^{(\widetilde{J}_h)} - v^{(J_h)}\|_2 \le 2\|m_h - v^{(J_h)}\|_2.$$

The preceding Fano bound applies to this decoder, so

$$\mathbb{E}\|v^{(J_h)} - m_h\|_2^2 \ge c_1 d \, \Pr(\widetilde{J}_h \neq J_h) \ge \frac{c_1 d}{4}. \tag{33}$$

For $f(p) = \sum_x p_x \log p_x$, every coordinate on the line segment from q_{J_h} to $\overline{q}_h$ is at most $(1+\alpha)/d \le 5/(4d)$; hence $\nabla^2 f \succeq (4d/5) I$ there. The Bregman-divergence identity for KL and strong convexity yield

$$\mathrm{KL}(q_{J_h} \| \overline{q}_h) \ge \frac{2d}{5} \|q_{J_h} - \overline{q}_h\|_2^2 = \frac{2\alpha^2}{5d} \|v^{(J_h)} - m_h\|_2^2. \tag{34}$$

Combining the last two displays gives

$$\mathbb{E}\mathrm{KL}(q_{J_h} \| \overline{q}_h) \ge \frac{c_1}{10} \alpha^2. \tag{35}$$

It remains to show that an arbitrary joint output cannot beat the sum of these posterior-predictive risks. Posterior factorization gives

$$\overline{P}_{\mathcal{D}} := \mathbb{E}[P_J \mid \mathcal{D}] = \bigotimes_{h=1}^{H} \overline{q}_h.$$

For every trajectory law Q, inserting $\overline{P}_{\mathcal{D}}$ in the log ratio and using $\mathbb{E}[P_J \mid \mathcal{D}] = \overline{P}_{\mathcal{D}}$ gives the conditional Bayes projection identity

$$\mathbb{E}[\mathrm{KL}(P_J \| Q) \mid \mathcal{D}] = \mathbb{E}[\mathrm{KL}(P_J \| \overline{P}_{\mathcal{D}}) \mid \mathcal{D}] + \mathrm{KL}(\overline{P}_{\mathcal{D}} \| Q). \tag{36}$$

The identity remains valid after averaging over estimator randomization. Therefore every unrestricted estimator $P^{\hat{\pi}}$ satisfies

$$\begin{aligned} \mathbb{E}\mathrm{KL}(P_J \| P^{\hat{\pi}}) &\ge \mathbb{E}\mathrm{KL}(P_J \| \overline{P}_{\mathcal{D}}) \\ &= \sum_{h=1}^{H} \mathbb{E}\mathrm{KL}(q_{J_h} \| \overline{q}_h) \\ &\ge \frac{c_1}{10} H \alpha^2. \end{aligned}$$

The supremum risk is at least this Bayes risk. Substituting $\alpha = aG\sqrt{\log M/n}$ and absorbing $a^2c_1/10$ into c proves (29). □

Proof of Theorem 4.2: Let $\Pi = \Pi_0^H$ and $\{P^\pi : \pi \in \Pi\}$ be the decomposable hard family supplied by Theorem B.1. Thus $|\Pi_0| = M$ and, for every possibly randomized measurable estimator $P^{\hat{\pi}}$ taking values in the set of all probability laws on the original trajectory space,

$$\sup_{\pi\in\Pi} \mathbb{E}_{P^\pi}\left[\mathrm{KL}\big(P^\pi \,\big\|\, P^{\hat{\pi}}\big)\right] \geq c\,\frac{HG^2\log M}{n}.$$

Fix two distinct probability laws R_0 and R_1 on a finite auxiliary sample space, with finite $\mathrm{KL}(R_1\|R_0)$. For example, one may take two distinct Bernoulli laws with parameters in $(0,1)$.

Augment every one-step observation with an independent auxiliary coordinate. Define the lifted model corresponding to $\pi \in \Pi$ by

$$\widetilde{P}^\pi := P^\pi \otimes R_0^{\otimes H},$$

and define

$$\widetilde{\Pi} := \{\widetilde{P}^\pi : \pi \in \Pi\}.$$

Since the same factor R_0 is appended at every coordinate, $\widetilde{\Pi}$ is again decomposable and has the same single-step cardinality M. Moreover, all likelihood ratios between members of $\widetilde{\Pi}$ are exactly the corresponding likelihood ratios between the original members, so the same G-condition is preserved.

Define the misspecified truth

$$\widetilde{P}^{\pi^\star} := P^{\bar{\pi}} \otimes R_1^{\otimes H}$$

for some $\bar{\pi} \in \Pi$. Since every element of $\mathrm{conv}(\widetilde{\Pi})$ has auxiliary marginal $R_0^{\otimes H}$ whereas $\widetilde{P}^{\pi^\star}$ has auxiliary marginal $R_1^{\otimes H}$, we have

$$\widetilde{P}^{\pi^\star} \notin \mathrm{conv}(\widetilde{\Pi}).$$

Hence, by additivity of KL divergence under products,

$$\begin{aligned}\mathrm{KL}\Big(\widetilde{P}^{\pi^\star} \,\Big\|\, \widetilde{P}^{\pi}\Big) &= \mathrm{KL}\big(P^{\bar{\pi}} \otimes R_1^{\otimes H} \,\big\|\, P^\pi \otimes R_0^{\otimes H}\big)\\ &= \mathrm{KL}\big(P^{\bar{\pi}} \,\big\|\, P^\pi\big) + H\mathrm{KL}(R_1\|R_0).\end{aligned}$$

Since $\bar{\pi} \in \Pi$, taking $P^\pi = P^{\bar{\pi}}$ gives

$$\inf_{\widetilde{P}^\pi\in\widetilde{\Pi}} \mathrm{KL}\Big(\widetilde{P}^{\pi^\star} \,\Big\|\, \widetilde{P}^{\pi}\Big) = H\mathrm{KL}(R_1\|R_0).$$

Now let

$$\widetilde{P}^{\hat{\pi}} \in \mathrm{conv}(\widetilde{\Pi})$$

be any possibly randomized estimator based on n samples from $\widetilde{P}^{\pi^\star}$. Write

$$\widetilde{P}^{\hat{\pi}} = P^{\hat{\pi}} \otimes R_0^{\otimes H}.$$

Then its excess KL risk is exactly

$$\mathrm{KL}\Big(\widetilde{P}^{\pi^\star} \,\Big\|\, \widetilde{P}^{\hat{\pi}}\Big) - \inf_{\widetilde{P}^\pi\in\widetilde{\Pi}} \mathrm{KL}\Big(\widetilde{P}^{\pi^\star} \,\Big\|\, \widetilde{P}^{\pi}\Big) = \mathrm{KL}\big(P^{\bar{\pi}} \,\big\|\, P^{\hat{\pi}}\big)\,. \tag{37}$$

The auxiliary observations are drawn from $R_1^{\otimes H}$ independently of $\bar{\pi}$. Hence they contain no information about the identity of the original hard instance $P^{\bar{\pi}}$.

Fix any possibly randomized estimator $\widetilde{P}^{\hat{\pi}} \in \mathrm{conv}(\widetilde{\Pi})$ based on n samples from $\widetilde{P}^{\pi^\star}$, and write $\widetilde{P}^{\hat{\pi}} = P^{\hat{\pi}} \otimes R_0^{\otimes H}$. The law $P^{\hat{\pi}}$ is supported on the original trajectory space. Its random mixture weights may nevertheless depend on the auxiliary observations through the estimator $\widetilde{P}^{\hat{\pi}}$. Since those observations are independent of $\bar{\pi}$, they act only as additional randomization conditional on the original data. Thus $P^{\hat{\pi}}$ is an admissible randomized estimator in Theorem B.1. Consequently,

$$\inf_{\hat{\pi}} \sup_{\pi^\star\in\Pi} \mathbb{E}_{P^{\pi^\star}}\left[\mathrm{KL}\Big(P^{\pi^\star} \,\Big\|\, P^{\hat{\pi}}\Big)\right] \;\geq\; c\,\frac{H\,G^2\,\log M}{n}. \tag{38}$$

Combining (37) and (38), we obtain

$$\inf_{\widetilde{P}^{\hat{\pi}} \in \mathrm{conv}(\widetilde{\Pi})} \sup_{\bar{\pi} \in \Pi} \mathbb{E}_{\widetilde{P}^{\pi^\star}} \left[\mathrm{KL}\Big(\widetilde{P}^{\pi^\star} \,\Big\|\, \widetilde{P}^{\hat{\pi}} \Big) - \inf_{\widetilde{P}^{\pi} \in \widetilde{\Pi}} \mathrm{KL}\Big(\widetilde{P}^{\pi^\star} \,\Big\|\, \widetilde{P}^{\pi} \Big) \right] \geq c \, \frac{HG^2 \log M}{n}. \tag{39}$$

This proves the claim. □

Theorem B.2 (Realizable lower bound for a fully shared class and proper learners)**.** *There exist absolute constants $c, c_0 > 0$ such that for any $M \geq 4$, any $n \geq c_0 \log M$, any horizon $H \in \mathbb{N}$, and any $G \in (0, 1]$, one can construct a finite fully shared policy class $\Pi_{\text{cons}} = \{\pi^{(1)}, \ldots, \pi^{(M)}\}$ with $|\Pi_{\text{cons}}| = M$ and associated distributions $\{P^\pi\}_{\pi \in \Pi_{\text{cons}}}$ on a finite sample space, satisfying Assumption 1, such that*

$$\inf_{\hat{\pi}} \sup_{\pi^\star \in \Pi_{\text{cons}}} \mathbb{E}_{P^{\pi^\star}} \left[\mathrm{KL}\left(P^{\pi^\star} \| P^{\hat{\pi}} \right) \right] \geq c \frac{HG^2 \log M}{n}. \tag{40}$$

Here each policy uses the same conditional kernel across all H steps in an observed-baseline context experiment. The infimum is over all possibly randomized measurable estimators $\hat{\pi} = \hat{\pi}(S_{1:n}) \in \Pi_{\text{cons}}$ based on $S_1, \ldots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$.

Proof of Theorem B.2: Fix $M \geq 4$. A binary Gilbert–Varshamov packing gives a universal constant C_0 and an integer $d \leq C_0 \log M$ for which there are exactly M codewords

$$v^{(1)}, \ldots, v^{(M)} \in \{-1, +1\}^d, \qquad D\big(v^{(k)}, v^{(\ell)}\big) \geq \frac{d}{4} \quad (k \neq \ell),$$

where D is Hamming distance. Each trajectory begins with a common, model-independent observable context $X = (J, R)$, where

$$J \sim \mathrm{Unif}[d], \qquad R \sim \mathrm{Bernoulli}(\rho), \qquad \rho := \frac{\log M}{16n \log 2}.$$

The variables J and R are independent. Choosing the universal c_0 in the theorem sufficiently large ensures $\rho \leq 1$. There is a singleton action space, so no action factor enters any likelihood ratio.

Set

$$\alpha := \frac{1}{2} \tanh\left(\frac{G}{2} \right).$$

For each codeword $v^{(k)}$, define one history kernel q_k and use that *same kernel at every round*:

$$q_k(Y_h = 1 \mid X, Y_{<h}) = \frac{1}{2} + \alpha R v_J^{(k)}, \qquad h \in [H].$$

Thus $Y_1, \ldots, Y_H$ are conditionally independent given X, and the kernel ignores previous outcomes. This is a genuinely fully shared class of exactly M elements, not a product class with independently selected stagewise indices. The context is an initial observable with a common marginal and hence cancels from every KL ratio.

Let $p_\pm = 1/2 \pm \alpha$. By the definition of α,

$$\log \frac{p_+}{p_-} = G.$$

Consequently, all conditional log ratios between members of the class have absolute value at most G. The KL divergence between the opposite-sign Bernoulli laws, in either direction, is

$$\kappa_G := \mathrm{KL}(\mathrm{Bern}(p_+) \,\|\, \mathrm{Bern}(p_-)) = 2\alpha G = G \tanh\left(\frac{G}{2} \right) \geq \frac{G^2}{4},$$

where the final inequality uses $G \le 1$. Let P_k denote the resulting joint law of $(X, Y_{1:H})$. Since the marginal law of X is the same under every P_k, the chain rule for KL divergence gives

$$\begin{aligned}
\mathrm{KL}(P_k \| P_\ell) &= \mathbb{E}_X \left[\mathrm{KL}(P_k(Y_{1:H} \mid X) \| P_\ell(Y_{1:H} \mid X))\right] \\
&= \mathbb{E}_{J,R} \left[\mathrm{KL}(P_k(Y_{1:H} \mid J, R) \| P_\ell(Y_{1:H} \mid J, R))\right] \\
&= H\kappa_G \mathbb{E}\left[\mathbf{1}\{R = 1\}\mathbf{1}\{v_J^{(k)} \neq v_J^{(\ell)}\}\right] \\
&= \rho H \kappa_G \frac{1}{d} \sum_{j=1}^{d} \mathbf{1}\{v_j^{(k)} \neq v_j^{(\ell)}\} = \frac{\rho H \kappa_G}{d} D\Big(v^{(k)}, v^{(\ell)}\Big).
\end{aligned}$$

Using the packing property,

$$\mathrm{KL}(P_k \| P_\ell) = \frac{\rho H \kappa_G}{d} D\Big(v^{(k)}, v^{(\ell)}\Big) \ge \frac{\rho H \kappa_G}{4}, \qquad k \neq \ell. \tag{41}$$

It remains to show that the training sample cannot reliably identify the codeword. Put the uniform prior on $K \in [M]$, and let $\mathcal{D}_n = (X_i, Y_{i,1:H})_{i=1}^n$. The contexts $X_i = (J_i, R_i)$ are independent of K. If $R_i = 0$, the conditional law of $Y_{i,1:H}$ is independent of K; if $R_i = 1$, it depends on K only through the single binary code bit $v_{J_i}^{(K)}$. We first bound the information carried by one trajectory:

$$\begin{aligned}
I(K; X_i, Y_{i,1:H}) &= I(K; X_i) + I(K; Y_{i,1:H} \mid X_i) \\
&= I(K; Y_{i,1:H} \mid J_i, R_i) \\
&= (1 - \rho) I(K; Y_{i,1:H} \mid J_i, R_i = 0) \\
&\quad + \rho I(K; Y_{i,1:H} \mid J_i, R_i = 1) \\
&\le \rho \log 2.
\end{aligned}$$

The first equality is the chain rule for mutual information, and the second uses the independence of K and X_i. For the final inequality, the term corresponding to $R_i = 0$ vanishes since $Y_{i,1:H}$ is conditionally independent of K given J_i and $R_i = 0$. For the term corresponding to $R_i = 1$, define

$$B_i := v_{J_i}^{(K)} \in \{-1, +1\}.$$

Conditional on J_i and $R_i = 1$, the Markov relation

$$K \longrightarrow B_i \longrightarrow Y_{i,1:H}$$

holds. Hence, by the data-processing inequality,

$$\begin{aligned}
I(K; Y_{i,1:H} \mid J_i, R_i = 1) &\le I(K; B_i \mid J_i, R_i = 1) \\
&\le H(B_i \mid J_i, R_i = 1) \\
&\le \log 2.
\end{aligned}$$

Now define $Z_i := (X_i, Y_{i,1:H})$. By the chain rule,

$$\begin{aligned}
I(K; \mathcal{D}_n) &= I(K; Z_1, \ldots, Z_n) \\
&= \sum_{i=1}^{n} I(K; Z_i \mid Z_{<i}) \\
&\le \sum_{i=1}^{n} I(K; Z_i) \\
&\le n\rho \log 2 \\
&= \frac{1}{16} \log M.
\end{aligned}$$

This argument remains valid even when H is arbitrarily large: the H repeated observations can reveal the activated code bit accurately, but one activated context carries at most one bit of information about the codeword. Fano's inequality now implies, for every possibly randomized estimator $\widehat{K}$,

$$\Pr(\widehat{K} \neq K) \ge 1 - \frac{I(K; \mathcal{D}_n) + \log 2}{\log M} \ge \frac{7}{16}.$$

Every proper Π_{cons}-valued estimator has the form $\hat{q} = q_{\widehat{K}}$. Combining the last display with (41) yields the Bayes-risk bound

$$\mathbb{E}\mathrm{KL}(P_K \| P_{\widehat{K}}) \geq \frac{7}{64}\rho H \kappa_G \geq c\,\frac{HG^2 \log M}{n}$$

for a universal $c > 0$. The supremum risk is at least the Bayes risk, proving (40). The least-favorable activation probability $\rho \asymp \log M/n$ depends on (M, n), as is standard for a local minimax construction. □

Proof of Theorem 4.3: The argument is identical to the decomposable lifting above, with Theorem B.2 replacing Theorem B.1. Let $\{P_k : k \in [M]\}$ be the fully shared hard family from Theorem B.2, and define

$$\widetilde{P}_k := P_k \otimes R_0^{\otimes H},$$
$$P_k^\star := P_k \otimes R_1^{\otimes H}.$$

The lifted class remains fully shared. Moreover, for every $j, k \in [M]$,

$$\mathrm{KL}\Big(P_k^\star \Big\| \widetilde{P}_j\Big) = \mathrm{KL}(P_k \| P_j) + H\mathrm{KL}(R_1 \| R_0).$$

Therefore,

$$\inf_{j \in [M]} \mathrm{KL}\Big(P_k^\star \Big\| \widetilde{P}_j\Big) = H\mathrm{KL}(R_1 \| R_0),$$

with the infimum attained at $j = k$. Hence, for every possibly randomized proper estimator $\widehat{K}$,

$$\begin{aligned}\mathrm{KL}\Big(P_k^\star \Big\| \widetilde{P}_{\widehat{K}}\Big) - \inf_{j \in [M]} \mathrm{KL}\Big(P_k^\star \Big\| \widetilde{P}_j\Big) \\ = \mathrm{KL}\big(P_k \big\| P_{\widehat{K}}\big)\,.\end{aligned}$$

Since the auxiliary variables are independent of k, any dependence of $\widehat{K}$ on them amounts only to additional estimator randomization. Theorem B.2 therefore applies to the induced randomized proper estimator, yielding

$$\inf_{\widehat{K}} \sup_{k \in [M]} \mathbb{E}_{(P_k^\star)^{\otimes n}} \left[\mathrm{KL}\Big(P_k^\star \Big\| \widetilde{P}_{\widehat{K}}\Big) - \inf_{j \in [M]} \mathrm{KL}\Big(P_k^\star \Big\| \widetilde{P}_j\Big)\right] \geq c\frac{HG^2 \log M}{n}.$$

This proves the claim. □

We next establish the statistical lower bound for general dependent policy classes discussed in Section 3.2. Unlike Theorem B.1, which allows unrestricted trajectory-law estimators over a decomposable class, the following result concerns proper learning over a class with cross-horizon coupling. Reusing the balanced packing from Theorem B.1, we construct a rare-activation experiment in which each informative trajectory reveals at most one codeword bit, while an incorrect policy incurs KL loss accumulating across the horizon. Fano's inequality then yields a linear-in-H lower bound even in the realizable setting.

Corollary 1 (Realizable lower bound for a dependent class and proper learners)**.** *There exist absolute constants $c, c_0 > 0$ such that for any $|\Pi| \geq 4$, any $n \geq c_0 \log |\Pi|$, any horizon $H \in \mathbb{N}$, and any $G \in (0, 1]$, one can construct a finite general dependent policy class Π and associated distributions $\{P^\pi\}_{\pi \in \Pi}$ on a finite sample space such that for every selector $\hat{\pi} = \hat{\pi}(S_{1:n}) \in \Pi$ based on $S_1, \ldots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ where $\pi^\star \in \Pi$,*

$$\inf_{\hat{\pi} \in \Pi} \sup_{\pi^\star \in \Pi} \mathbb{E}_{P^{\pi^\star}}\Big[\mathrm{KL}\Big(P^{\pi^\star} \Big\| P^{\hat{\pi}}\Big)\Big] \;\geq\; c\,\frac{HG^2\,\log|\Pi|}{n}. \tag{42}$$

Proof of Corollary 1: The case $H = 1$ follows from Theorem B.1. For $H \geq 2$, reuse the balanced codewords $v^{(1)}, \ldots, v^{(M)} \in \{-1, +1\}^d$ from its proof, satisfying $D(v^{(j)}, v^{(k)}) \geq c_1 d$ for $j \neq k$.

Set

$$p = \frac{\log M}{16n}, \qquad \alpha = \frac{G}{4}.$$

For each $j \in [M]$, define a trajectory law P_j on $S = (X, Y_2, \ldots, Y_H)$. The first coordinate is independent of j, with

$$\Pr(X = 0) = 1 - p, \qquad \Pr(X = x) = \frac{p}{d}, \quad x \in [d].$$

Conditional on X, the remaining coordinates are independent fair signs if $X = 0$; otherwise,

$$P_j(Y_h = y \mid X = x) = \frac{1 + \alpha v_x^{(j)} y}{2}, \qquad h = 2, \ldots, H.$$

Since

$$\log \frac{1+\alpha}{1-\alpha} \le 3\alpha \le G,$$

the class $\Pi = \{\pi_1, \ldots, \pi_M\}$ satisfies Assumption 1.

For $j \neq k$, the KL chain rule and packing separation yield

$$\begin{aligned} \mathrm{KL}(P_j \| P_k) &= p(H-1)\frac{D(v^{(j)}, v^{(k)})}{d} \alpha \log \frac{1+\alpha}{1-\alpha} \\ &\ge 2c_1 p(H-1)\alpha^2 \\ &= \frac{c_1}{8} p(H-1)G^2. \end{aligned}$$

Let J be uniform on $[M]$ and observe $S_{1:n} \overset{\text{iid}}{\sim} P_J$. Conditional on $X = x \neq 0$, the trajectory depends on J only through the single bit $v_x^{(J)}$. Hence

$$I(J; S_{1:n}) \le np \log 2 = \frac{\log 2}{16} \log M.$$

Fano's inequality gives, for every selector $\hat{J}$,

$$\Pr(\hat{J} \neq J) \ge 1 - \frac{\log 2}{16} - \frac{\log 2}{\log M} \ge \frac{1}{4}.$$

Consequently,

$$\begin{aligned} &\inf_{\hat{\pi} \in \Pi} \sup_{\pi^\star \in \Pi} \mathbb{E}_{P^{\pi^\star}} \left[\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \right] \\ &\quad \ge \frac{c_1}{32} p(H-1)G^2 \\ &\quad \ge \frac{c_1}{1024} \frac{HG^2 \log M}{n}. \end{aligned}$$

Adjusting the universal constant completes the proof. □

We next clarify the relationship between Bayesian prediction and convexification of autoregressive policy classes. Lemma 5 shows that Bayesian aggregation predictive laws belong to the trajectory-level convex hull, while their stepwise conditionals involve prefix-dependent mixture weights. Lemma 6 further establishes the equivalence between trajectory-level convexification and the convex hull of stepwise mixtures for decomposable classes. These results justify the connection between Bayesian prediction and the convexified class considered in Theorem 4.2.

Lemma 5 (Bayesian prediction and convexification)**.** *Let $\mathcal{P}_\Pi = \{P^\pi : \pi \in \Pi\}$ be a finite family of trajectory laws. For any posterior distribution $\rho(\cdot \mid D)$ over Π given training data D, the Bayesian posterior predictive law satisfies*

$$Q_D = \sum_{\pi \in \Pi} \rho(\pi \mid D) P^\pi \in \mathrm{conv}(\mathcal{P}_\Pi).$$

The same identity holds when predictions are computed sequentially using Bayesian aggregation updates along the trajectory prefix.

Proof. Write $w_\pi = \rho(\pi \mid D)$ and define

$$Z_h(s_{1:h}) = \sum_{\pi \in \Pi} w_\pi P^\pi(s_{1:h}), \qquad Z_0 = 1.$$

By Bayes' rule, the posterior after observing a prefix $s_{<h}$ assigns weight

$$w_{\pi,h} = \frac{w_\pi P^\pi(s_{<h})}{Z_{h-1}(s_{<h})}.$$

Thus the sequential Bayesian predictive conditional satisfies

$$Q_D(s_h \mid s_{<h}) = \sum_{\pi\in\Pi} w_{\pi,h}\pi_h(s_h \mid s_{<h}) = \frac{Z_h(s_{1:h})}{Z_{h-1}(s_{<h})}.$$

Multiplying over h telescopes, yielding

$$Q_D(s_{1:H}) = Z_H(s_{1:H}) = \sum_{\pi\in\Pi} w_\pi P^\pi(s_{1:H}),$$

as claimed. □

Lemma 6 (Convexification of decomposable classes)**.** *For any decomposable policy class* $\Pi = \prod_{h=1}^H \Pi_h$*, we have*

$$\operatorname{conv}\left(\prod_{h=1}^H \operatorname{conv}(\Pi_h)\right) = \operatorname{conv}(\Pi),$$

where convex hulls of policy classes are understood in terms of their induced trajectory laws.

Remark. The outer convex hull on the left-hand side is essential: a trajectory-level mixture may induce prefix-dependent mixture weights and therefore need not belong to $\prod_{h=1}^H \operatorname{conv}(\Pi_h)$ as a class of conditional kernels.

Proof. Since $\Pi_h \subseteq \operatorname{conv}(\Pi_h)$, the inclusion from right to left is immediate.

For the reverse inclusion, take any $\widetilde{\pi} \in \prod_{h=1}^H \operatorname{conv}(\Pi_h)$ and write

$$\widetilde{\pi}_h = \sum_{j=1}^{m_h} w_{h,j}\pi_{h,j}, \qquad \pi_{h,j} \in \Pi_h,$$

where the weights are nonnegative and sum to one. By autoregressive factorization,

$$\begin{aligned} P^{\widetilde{\pi}}(s_{1:H}) &= \prod_{h=1}^H \sum_{j=1}^{m_h} w_{h,j}\pi_{h,j}(s_h \mid s_{<h}) \\ &= \sum_{j_1,\dots,j_H} \left(\prod_{h=1}^H w_{h,j_h}\right) P^{(\pi_{1,j_1},\dots,\pi_{H,j_H})}(s_{1:H}). \end{aligned}$$

Since every $(\pi_{1,j_1},\dots,\pi_{H,j_H})$ belongs to Π, the resulting trajectory law lies in $\operatorname{conv}(\Pi)$. Taking convex hulls completes the proof. □

Proposition 2 (Trajectory-level aggregation of finite trajectory classes)**.** *Let* $\mathcal{P} = \{P_1,\dots,P_N\}$ *be any finite collection of trajectory laws. For every* $n \geq 0$*, there is an estimator* $\widehat{P}_{\mathrm{mix}} \in \operatorname{conv}(\mathcal{P})$ *based on* n *iid trajectories such that, for every trajectory law* $P^\star$*,*

$$\mathbb{E}_{(P^\star)^{\otimes n}} \mathrm{KL}(P^\star \| \widehat{P}_{\mathrm{mix}}) \leq \min_{k\in[N]} \mathrm{KL}(P^\star \| P_k) + \frac{\log N}{n+1}. \tag{43}$$

No horizon, boundedness, realizability, or autoregressive-independence assumption is required.

Proof. Let $p_1,\dots,p_N$ be densities with respect to a common dominating measure. Given training trajectories $Z_1,\dots,Z_n$, define, for $t = 0,\dots,n$,

$$w_{k,t} := \frac{\prod_{i=1}^t p_k(Z_i)}{\sum_{\ell=1}^N \prod_{i=1}^t p_\ell(Z_i)}, \qquad Q_t := \sum_{k=1}^N w_{k,t}P_k,$$

with $w_{k,0} = 1/N$, and set

$$\widehat{P}_{\mathrm{mix}} := \frac{1}{n+1} \sum_{t=0}^{n} Q_t.$$

Thus Q_t is the uniform-prior posterior predictive law after the first t training trajectories, and the displayed average is a deterministic estimator based on the full sample. On a zero-denominator event the weights may be defined arbitrarily; whenever the oracle term is finite, such events are $P^\star$-null.

Write q_t for the density of Q_t. The predictive likelihoods telescope:

$$\prod_{t=0}^{n} q_t(Z_{t+1}) = \frac{1}{N} \sum_{\ell=1}^{N} \prod_{i=1}^{n+1} p_\ell(Z_i).$$

Consequently, for every fixed $k \in [N]$ and every realization of $Z_{1:n+1}$,

$$\sum_{t=0}^{n} \log \frac{p_k(Z_{t+1})}{q_t(Z_{t+1})} \le \log N, \tag{44}$$

because the mixture on the right contains the kth likelihood with weight $1/N$.

Now let $Z_1, \dots, Z_{n+1}$ be iid from an arbitrary $P^\star$. Conditional expectation given $Z_{1:t}$ turns the tth expected log ratio in (44) into

$$\mathbb{E}\mathrm{KL}(P^\star \| Q_t) - \mathrm{KL}(P^\star \| P_k).$$

Taking expectations, summing, and using convexity of KL in its second argument therefore gives

$$\begin{aligned} \mathbb{E}\mathrm{KL}(P^\star \| \widehat{P}_{\mathrm{mix}}) &\le \frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E}\mathrm{KL}(P^\star \| Q_t) \\ &\le \mathrm{KL}(P^\star \| P_k) + \frac{\log N}{n+1}. \end{aligned}$$

Minimizing over k proves (43). For a realizable decomposable family $N = M^H$, the remainder is $H \log M/(n+1)$; for a realizable fully shared family $N = M$, it is $\log M/(n+1)$. □

C Missing proofs in Section 3

Proof of Theorem 3.2 It suffices to prove the claim for the special case $H = 1$. Indeed, the general H statement follows by taking a product construction across steps so that the joint KL adds over h and the lower bound scales linearly in H.

Let the token space be $\mathcal{X} = \{0, 1\}$. Fix a constant $a \in (0, 1/4]$, and for each $n \ge 1$ set

$$b := \frac{1}{10\sqrt{n}}.$$

Consider a base class $\Pi_0 = \{\pi^{(+)}, \pi^{(-)}\}$ consisting of two (history-independent) conditional models,

$$\pi^{(+)}(1) := \frac{1}{2} + a, \qquad \pi^{(-)}(1) := \frac{1}{2} - a, \qquad \pi^{(+)}(0) := 1 - \pi^{(+)}(1), \quad \pi^{(-)}(0) := 1 - \pi^{(-)}(1).$$

For the data-generating policy, define $\pi^\star \notin \Pi_0$ by

$$\pi^\star(1) := \frac{1}{2} + b, \qquad \pi^\star(0) := 1 - \pi^\star(1).$$

Then $P^{\pi^\star}$ is the induced distribution on $u \in \mathcal{X}$ with $\Pr(u = 1) = \frac{1}{2} + b$, and for any $\pi \in \Pi_0$, P^π is the induced distribution with $\Pr(u = 1) = \frac{1}{2} \pm a$. Since $b \ne a$, we have $\pi^\star \notin \Pi_0$, i.e. the model is misspecified. Moreover, Assumption 1 holds with absolute constants because all Bernoulli parameters lie in $[\frac{1}{2} - a, \frac{1}{2} + a] \subset [1/4, 3/4]$, so the log-loss range is uniformly bounded. When $H = 1$, both choices in the theorem reduce to the same class $\Pi = \Pi_0$ (since $\Pi_0^H = \Pi_0$ and $\Pi_{\mathrm{cons}} = \Pi_0$).

A direct computation yields

$$\mathrm{KL}\big(P^{\pi^\star} \| P^{\pi^{(-)}}\big) - \mathrm{KL}\big(P^{\pi^\star} \| P^{\pi^{(+)}}\big) = (2b) \log \frac{\frac{1}{2} + a}{\frac{1}{2} - a}.$$

Choosing $a = 1/4$ gives the explicit gap $2b \log 3$. In particular,

$$\inf_{\pi \in \Pi_0} \mathrm{KL}\big(P^{\pi^\star} \| P^{\pi}\big) = \min\Big\{\mathrm{KL}\big(P^{\pi^\star} \| P^{\pi^{(+)}}\big), \mathrm{KL}\big(P^{\pi^\star} \| P^{\pi^{(-)}}\big)\Big\},$$

and selecting the "wrong" element of Π_0 increases the KL by exactly $2b \log 3$.

Let $\hat{\pi}$ be any (possibly randomized) learner that outputs $\hat{\pi} \in \Pi_0$ from n i.i.d. samples $u_1, \ldots, u_n \sim P^{\pi^\star}$. Consider the associated test $\varphi := \mathbf{1}\{\hat{\pi} = \pi^{(+)}\}$. If instead the data-generating policy is

$$\pi^\star_+(1) = \frac{1}{2} + b \qquad \text{or} \qquad \pi^\star_-(1) = \frac{1}{2} - b,$$

then Le Cam's inequality gives

$$\Pr_{P^{\pi^\star_+}}(\hat{\pi} = \pi^{(-)}) + \Pr_{P^{\pi^\star_-}}(\hat{\pi} = \pi^{(+)}) \;\geq\; 1 - \mathrm{TV}\big((P^{\pi^\star_+})^{\otimes n}, (P^{\pi^\star_-})^{\otimes n}\big).$$

By Pinsker's inequality and tensorization,

$$\mathrm{TV}\big((P^{\pi^\star_+})^{\otimes n}, (P^{\pi^\star_-})^{\otimes n}\big) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}\big((P^{\pi^\star_+})^{\otimes n} \| (P^{\pi^\star_-})^{\otimes n}\big)} = \sqrt{\frac{n}{2}\mathrm{KL}\big(P^{\pi^\star_+} \| P^{\pi^\star_-}\big)}.$$

For Bernoulli parameters $\frac{1}{2} \pm b$ with $b \leq 1/10$, one has the standard bound

$$\mathrm{KL}\big(P^{\pi^\star_+} \| P^{\pi^\star_-}\big) \leq \frac{(2b)^2}{(\frac{1}{2} - b)(\frac{1}{2} + b)} \leq 17b^2,$$

which implies

$$\mathrm{TV}\big((P^{\pi^\star_+})^{\otimes n}, (P^{\pi^\star_-})^{\otimes n}\big) \leq \sqrt{\frac{n}{2} \cdot 17b^2} = \frac{\sqrt{8.5}}{10} < \frac{1}{3}.$$

Therefore,

$$\Pr_{P^{\pi^\star_+}}(\hat{\pi} = \pi^{(-)}) + \Pr_{P^{\pi^\star_-}}(\hat{\pi} = \pi^{(+)}) \geq \frac{2}{3},$$

so for at least one choice of sign $\sigma \in \{+, -\}$ we have

$$\Pr_{P^{\pi^\star_\sigma}}(\hat{\pi} \neq \pi^{(\sigma)}) \geq \frac{1}{3}.$$

Fix such a sign and take the corresponding data-generating policy $\pi^\star := \pi^\star_\sigma$.

Now we can conclude the lower bound. On the event $\{\hat{\pi} \neq \pi^{(\sigma)}\}$, the excess joint KL over the oracle in Π_0 is exactly $2b \log 3$. Hence, with probability at least $\delta_0 := 1/3$,

$$\mathrm{KL}\big(P^{\pi^\star} \| P^{\hat{\pi}}\big) \;\geq\; \inf_{\pi \in \Pi_0} \mathrm{KL}\big(P^{\pi^\star} \| P^{\pi}\big) \;+\; 2b \log 3 \;=\; \inf_{\pi \in \Pi} \mathrm{KL}\big(P^{\pi^\star} \| P^{\pi}\big) \;+\; c \cdot \frac{1}{\sqrt{n}},$$

where $c := \frac{\log 3}{5}$ since $b = 1/(10\sqrt{n})$. This proves Theorem 3.2 for $H = 1$.

As noted at the beginning, the general H statement follows by taking the product construction across $h \in [H]$, in which case the excess KL accumulates additively, yielding the term $cH/\sqrt{n}$ while the same constant-probability event persists. □

Proof of Lemma 1: By definition, $\Pi_0 = \bigcup_{h=1}^H \Pi_h$, hence

$$|\Pi_0| \leq \sum_{h=1}^H |\Pi_h|.$$

For each h, the coordinate projection map $\mathrm{proj}_h : \Pi \to \Pi_h$ given by $\mathrm{proj}_h(\pi) = \pi_h$ is surjective by definition of Π_h, and therefore $|\Pi_h| \leq |\Pi|$. Combining these two bounds yields

$$|\Pi_0| \leq \sum_{h=1}^H |\Pi_h| \leq \sum_{h=1}^H |\Pi| = H|\Pi|.$$

Taking logarithms gives the first inequality.

We next prove the second one. Since $\Pi_h \subseteq \Pi_0$ for all h, we have

$$\Pi \subseteq \Pi_1 \times \cdots \times \Pi_H \subseteq \Pi_0^H.$$

Therefore $|\Pi| \le |\Pi_0^H| = |\Pi_0|^H$. Taking logarithms gives the result. □

D More Discussion on Squared Hellinger Distance

Theorem 4.7 of Rohatgi et al. (2025) establishes an $\Omega(H)$ approximation-error amplification under squared Hellinger distance for iterative learners over policy classes without parameter sharing, corresponding to our decomposable regime. We complement their result by showing that the same linear amplification can persist even under full parameter sharing, for proper iterative learners. Thus, the Hellinger approximation barrier is not specific to horizon-dependent parameterizations. Together with our joint-KL results, this provides a more complete comparison of horizon dependence across the two divergences and the two canonical policy-class structures.

Theorem D.1 (Fully-shared lower bound for iterative learners)**.** *Fix any horizon $H \ge 2$. There exists a two-element fully-shared policy class Π_{cons} such that, for any potentially randomized iterative learner in the sense of Definition 4.6 of Rohatgi et al. (2025), there exists an expert policy $\pi^\star$ satisfying*

$$\mathbb{E}\Big[D_{\mathrm{H}}^2\Big(P^{\widehat{\pi}}, P^{\pi^\star}\Big)\Big] \ge cH \min_{\pi \in \Pi_{\text{cons}}} D_{\mathrm{H}}^2\Big(P^{\pi}, P^{\pi^\star}\Big),$$

for some universal constant $c > 0$. The expectation is over the internal randomness of the iterative learner.

Proof. Consider an H-step autoregressive model with action space $\mathcal{A} = \{0, 1\}$ and no nontrivial context. Let

$$\Delta := \frac{1}{8\sqrt{H}}, \qquad p_- := \frac{1}{2} - \Delta, \qquad p_+ := \frac{1}{2} + \Delta.$$

Define two fully-shared policies π^-, π^+ by

$$\pi_h^{\pm}(1 \mid a_{<h}) = p_\pm, \qquad h \in [H],$$

independently of the history $a_{<h}$, and let

$$\Pi_{\text{cons}} := \{\pi^-, \pi^+\}.$$

Thus

$$P^{\pi^\pm} = \mathrm{Ber}(p_\pm)^{\otimes H}.$$

We construct two expert policies $\pi^{\star,+}$ and $\pi^{\star,-}$. At the first step, both experts use the same conditional distribution,

$$\pi_1^{\star,+} = \pi_1^{\star,-} = \mathrm{Ber}\left(\frac{1}{2}\right),$$

while for every $h \ge 2$,

$$\pi_h^{\star,\pm} = \mathrm{Ber}(p_\pm).$$

Hence

$$P^{\pi^{\star,\pm}} = \mathrm{Ber}\left(\frac{1}{2}\right) \otimes \mathrm{Ber}(p_\pm)^{\otimes(H-1)}.$$

We first show that any iterative learner must make an incorrect global choice with constant probability on at least one of these two experts. By Definition 4.6 of Rohatgi et al. (2025), the first output $\widehat{\pi}_1$ is a potentially randomized function only of $\pi_1^\star$. Since

$$\pi_1^{\star,+} = \pi_1^{\star,-},$$

the distribution of $\widehat{\pi}_1$ is identical under the two experts. Moreover, since the final output satisfies $\widehat{\pi} \in \Pi_{\text{cons}}$ and $\pi_1^- \ne \pi_1^+$, the first component uniquely determines whether the final output is π^- or π^+.

Let

$$q := \mathbb{P}\big(\widehat{\pi} = \pi^+\big)$$

under this common first-step input. Then

$$\mathbb{P}_{\pi^{\star,-}}\big(\widehat{\pi} = \pi^+\big) = q, \qquad \mathbb{P}_{\pi^{\star,+}}\big(\widehat{\pi} = \pi^-\big) = 1 - q.$$

Therefore,

$$\max\left\{\mathbb{P}_{\pi^{\star,-}}(\widehat{\pi}=\pi^{+}),\mathbb{P}_{\pi^{\star,+}}(\widehat{\pi}=\pi^{-})\right\}\geq\frac{1}{2}.$$

Fix $j\in\{-,+\}$ for which the corresponding error probability is at least $1/2$.

It remains to compare the Hellinger loss of the correct and incorrect policies. Let

$$\alpha:=\int\sqrt{d\,\mathrm{Ber}(1/2)\,d\,\mathrm{Ber}(p_{+})}=\int\sqrt{d\,\mathrm{Ber}(1/2)\,d\,\mathrm{Ber}(p_{-})},$$

and

$$\beta:=\int\sqrt{d\,\mathrm{Ber}(p_{+})\,d\,\mathrm{Ber}(p_{-})}.$$

By symmetry,

$$\alpha=\frac{\sqrt{1+2\Delta}+\sqrt{1-2\Delta}}{2},\qquad\beta=\sqrt{1-4\Delta^{2}}.$$

Using multiplicativity of the Hellinger affinity for product distributions,

$$D_{\mathrm{H}}^{2}\Big(P^{\pi^{\star,j}},P^{\pi^{j}}\Big)=1-\alpha,$$

whereas

$$D_{\mathrm{H}}^{2}\Big(P^{\pi^{\star,j}},P^{\pi^{-j}}\Big)=1-\alpha\beta^{H-1}.$$

Since $\beta<1$, the matching policy π^{j} is the unique best-in-class policy. Hence

$$\min_{\pi\in\Pi_{\mathrm{cons}}}D_{\mathrm{H}}^{2}\Big(P^{\pi},P^{\pi^{\star,j}}\Big)=1-\alpha.$$

We next bound these two quantities. Since $\Delta\leq1/8$,

$$\begin{aligned}1-\alpha&=\frac{1}{2}\left[\left(\sqrt{\frac{1}{2}+\Delta}-\sqrt{\frac{1}{2}}\right)^{2}+\left(\sqrt{\frac{1}{2}-\Delta}-\sqrt{\frac{1}{2}}\right)^{2}\right]\\&\leq\Delta^{2}=\frac{1}{64H}.\end{aligned}$$

On the other hand,

$$\beta=\sqrt{1-4\Delta^{2}}\leq\exp(-2\Delta^{2}),$$

and therefore

$$\begin{aligned}D_{\mathrm{H}}^{2}\Big(P^{\pi^{\star,j}},P^{\pi^{-j}}\Big)&=1-\alpha\beta^{H-1}\\&\geq1-\beta^{H-1}\\&\geq1-\exp\big(-2(H-1)\Delta^{2}\big)\\&\geq1-e^{-1/64},\end{aligned}$$

where the last inequality uses $H\geq2$ and $\Delta^{2}=1/(64H)$.

Since the learner outputs the wrong policy with probability at least $1/2$ on the selected expert $\pi^{\star,j}$,

$$\begin{aligned}\mathbb{E}\Big[D_{\mathrm{H}}^{2}\Big(P^{\widehat{\pi}},P^{\pi^{\star,j}}\Big)\Big]&\geq\frac{1}{2}D_{\mathrm{H}}^{2}\Big(P^{\pi^{-j}},P^{\pi^{\star,j}}\Big)\\&\geq\frac{1}{2}(1-e^{-1/64}).\end{aligned}$$

Combining this with

$$\min_{\pi\in\Pi_{\mathrm{cons}}}D_{\mathrm{H}}^{2}\Big(P^{\pi},P^{\pi^{\star,j}}\Big)\leq\frac{1}{64H},$$

we obtain

$$\mathbb{E}\Big[D_{\mathrm{H}}^{2}\Big(P^{\widehat{\pi}},P^{\pi^{\star,j}}\Big)\Big]\geq32(1-e^{-1/64})\,H\min_{\pi\in\Pi_{\mathrm{cons}}}D_{\mathrm{H}}^{2}\Big(P^{\pi},P^{\pi^{\star,j}}\Big).$$

The claim follows with $c=32(1-e^{-1/64})>0$. □

E Technical Tools

Lemma 7 (Central to Bernstein (Mehta, 2017))**.** *Let X be a random variable taking values in $[-B, B]$. Assume that $\mathbb{E}\big[e^{-\eta X}\big] \le 1$, then*

$$\mathbb{E}[X^2] \le 4\left(\frac{1}{\eta} + B\right)\mathbb{E}[X].$$

Lemma 8 (Freedman's inequality (Beygelzimer et al., 2011))**.** *Let $(X_t)_{t\le T}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathcal{F}_t)_{t\le T}$. If $|X_t| \le R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \;\le\; \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}\big[X_t^2\big] \;+\; \frac{\log(\delta^{-1})}{\eta}\,. \tag{45}$$

The following result is a consequence of Lemma 8.

Lemma 9 (One-sided martingale concentration inequality)**.** *Let $(X_t)_{t\le T}$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_t)_{t\le T}$, and let $\varepsilon > 0$ and $\delta \in (0,1)$. If $0 \le X_t \le R$ almost surely, then with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} X_t \;\le\; (1+\varepsilon)\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \;+\; \frac{(1+\varepsilon)^2 R}{\varepsilon}\log(\delta^{-1}), \tag{46}$$

and also with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \;\le\; (1+\varepsilon)\sum_{t=1}^{T} X_t \;+\; \frac{(1+\varepsilon)^2 R}{\varepsilon}\log(\delta^{-1}). \tag{47}$$

Proof of Lemma 9: Define the martingale difference sequence

$$Z_t := X_t - \mathbb{E}_{t-1}[X_t], \qquad t = 1, \dots, T.$$

Since $0 \le X_t \le R$, we have $|Z_t| \le R$ almost surely. Moreover,

$$\mathbb{E}_{t-1}[Z_t^2] = \mathrm{Var}_{t-1}(X_t) \le \mathbb{E}_{t-1}[X_t^2] \le R\,\mathbb{E}_{t-1}[X_t],$$

where the last inequality uses $X_t^2 \le RX_t$ for $X_t \in [0, R]$.

We apply Freedman's inequality (Lemma 8) to $(Z_t)_{t\le T}$. For any $\eta \in (0, 1/R)$, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T} Z_t &\le \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_t^2] + \frac{\log(\delta^{-1})}{\eta} \\ &\le \eta R \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + \frac{\log(\delta^{-1})}{\eta}.\end{aligned}$$

Recalling $\sum_{t=1}^{T} Z_t = \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t]$, we obtain

$$\sum_{t=1}^{T} X_t \le (1+\eta R)\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + \frac{\log(\delta^{-1})}{\eta}.$$

Choosing $\eta = \frac{\varepsilon}{(1+\varepsilon)R} \in (0, 1/R)$ gives $1 + \eta R = \frac{1+2\varepsilon}{1+\varepsilon} \le 1+\varepsilon$ and $\frac{1}{\eta} = \frac{(1+\varepsilon)R}{\varepsilon} \le \frac{(1+\varepsilon)^2 R}{\varepsilon}$, which yields (46).

For (47), apply Freedman's inequality to the martingale difference sequence $-Z_t$. Equivalently, with probability at least $1-\delta$,

$$-\sum_{t=1}^{T} Z_t \le \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_t^2] + \frac{\log(\delta^{-1})}{\eta} \le \eta R \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] + \frac{\log(\delta^{-1})}{\eta}.$$

Since $-\sum_{t=1}^{T} Z_t = \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] - \sum_{t=1}^{T} X_t$, this implies

$$(1-\eta R)\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \le \sum_{t=1}^{T} X_t + \frac{\log(\delta^{-1})}{\eta}.$$

Choose $\eta = \frac{\varepsilon}{(1+\varepsilon)R} \in (0, 1/R)$, so that $(1-\eta R)^{-1} = 1+\varepsilon$. Then

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t] \le (1+\varepsilon)\sum_{t=1}^{T} X_t + (1+\varepsilon)\frac{\log(\delta^{-1})}{\eta} = (1+\varepsilon)\sum_{t=1}^{T} X_t + \frac{(1+\varepsilon)^2 R}{\varepsilon}\log(\delta^{-1}),$$

which is (47). □

F Impact Statement and Limitations

F.1 Broader Impacts

This paper is primarily theoretical: it studies autoregressive learning under joint KL evaluation and clarifies how log-loss training, misspecification, and horizon dependence interact at the sequence-distribution level. Its potential positive impact is methodological reliability. Many sequential prediction and decision systems are used through the full trajectories they induce, so evaluating only conditional accuracy can obscure long-horizon error accumulation. By aligning the training objective, evaluation metric, and approximation metric in joint KL, our results provide sharper tools for auditing autoregressive models and for understanding when full-sequence behavior is controlled. At the same time, these results should not be interpreted as deployment guarantees by themselves: joint-KL control on the data-generating distribution does not ensure validity under unsupported interventions, distribution shift, hidden confounding, or misaligned objectives. If sequential data encode historical bias, strategic behavior, or missing variables, sharper sequence modeling may amplify rather than remove such risks. Responsible use therefore requires support/identification checks, robustness analysis, constraint and fairness auditing, and appropriate human oversight in high-stakes applications.

F.2 Limitations

Our theoretical framework characterizes horizon dependence in autoregressive learning under joint KL and connects sequence-level statistical guarantees to downstream decision control. The present work does not provide large-scale empirical validation or domain-specific deployment studies. In particular, applying the theory to prediction and decision systems requires problem-specific modeling choices, sufficient support for the relevant history-action pairs, and careful treatment of distribution shift and identification. We view applications to language modeling, dynamical systems, and decision-making as important follow-up directions, and are actively developing empirical and domain-driven extensions of the framework.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims in the abstract and Section 1.1 are established in Sections 3–5 (Theorems 3.1, 4.2 and 4.3), with complete proofs in Appendices A–C.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes, they are discussed in Appendix F.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: All assumptions are stated in Assumption 1 and in each theorem statement; complete proofs are given in Appendices A–D, with the technical tools collected in Appendix E.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: This work currently does not contain experiments.

   Guidelines:

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

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

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

Answer: [N/A]

Justification: This work currently does not contain data and code.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: This paper currently does not contain experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: This work currently does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: This work currently does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Yes, they are discussed in Appendix F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The current paper is primarily theoretical and does not release data or product-level models.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The current paper is primarily theoretical and is not built upon existing code, data, or product-level models.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The current paper is primarily theoretical and does not release new code, data, or product-level models.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The current paper is primarily theoretical, and there are no crowdsourcing experiments or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

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

    Guidelines:

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