Autoregressive Learning under Joint KL Analysis: Horizon-Free Approximation and Computational-Statistical Tradeoffs

Yunbei Xu[1] **Yuzhe Yuan**[1] **Ruohan Zhan**[2]
[1]National University of Singapore [2]University College London
`yunbei@nus.edu.sg` `e1374511@u.nus.edu` `ruohan.zhan@ucl.ac.uk`

Abstract

We study autoregressive generative modeling under misspecification measured by the *joint* Kullback–Leibler (KL) divergence. For approximation, we show that joint KL admits a horizon-free barrier independent of the sequence length H, unlike prior Hellinger-based analyses that imply an $\Omega(H)$ dependence. For estimation, we prove a finite-sample lower bound showing that any proper learner, including empirical risk minimization, suffers $\Omega(H^2)$ error. We then propose an improper Bayesian posterior learner that leverages a lifted policy space for computational efficiency, achieving horizon-free approximation and an $O(H)$ estimation rate. Our results identify divergence choice as the source of horizon dependence in approximation and establish a genuine computational-statistical tradeoff for estimation, motivating new algorithmic designs.

1 Introduction

We study *autoregressive (AR) modeling* of long sequences, the workhorse behind modern large language models.

Let $\pi^\star = (\pi_1^\star, \dots, \pi_H^\star)$ be the data-generating policy for a length-H sequence $S = (u_1, \dots, u_H)$, where $\pi_h^*(u_h|\cdot)$ is the conditional distribution of the h-th token given its history. We write $P^{\pi^\star}$ for the induced joint distribution over S. For any candidate policy $\pi = (\pi_1, \dots, \pi_H) \in \Pi$, the induced distribution is written as P^π. The standard maximum-likelihood (cross-entropy) objective (Goodfellow et al., 2016, §20), used explicitly since GPT-1 (Radford et al., 2018), is equivalent to minimizing the *joint* Kullback–Leibler (KL) divergence:

$$\min_{\pi \in \Pi} \mathrm{KL}\big(P^{\pi^\star}, P^\pi\big). \tag{1}$$

AR modeling exploits the sequential structure via the autoregressive factorization

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

We study this learning problem under *misspecification*, where the true policy $\pi^\star$ may not belong to the policy class Π. A standard decomposition of the generalization risk of an estimator $\hat{\pi}$ is

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

where n denotes the number of *i.i.d.* trajectory samples. Recent work Rohatgi et al. (2025) studied this problem under the squared Hellinger distance $d(P,Q) = D_H^2(P,Q)$. They show that any computationally efficient algorithm must incur an approximation ratio $C_{\text{apx}} = \Omega(H)$. However, their analysis does not precisely characterize the horizon dependence of C_{est}: The stated computational–statistical trade-off is based solely on C_{apx}, while C_{est} is left unspecified; an oversimplified decomposability assumption on Π implicitly obscures its H-dependence and algorithmic tractability.

This motivates us to investigate whether adopting a more canonical, practice-aligned criterion, which evaluates models by their *joint* KL to the data-generating process, can fundamentally improve the misspecification barrier. We focus on the more general and realistic *dependent policy class* setting, which allows parameter sharing across horizons as well as non-convex or combinatorial constraints. This setting reflects how large-scale language models are trained in practice, while posing new challenges for designing computationally efficient algorithms and for precisely characterizing C_{est}.

With these considerations in place, we ask the following central questions:

Whether the horizon barrier in C_{apx} is fundamental under KL? how to develop a more precise characterization of the computational–statistical tradeoff, including potential barriers for C_{est} *as well as the design of new efficient algorithms?*

1.1 Contributions

We provide concise answers to the main questions posed above.

- **Misspecified guarantees and error amplification.** We establish a *sharp generalization guarantee* for log-loss, proving that $C_{\text{apx}} = O(1)$ under joint KL divergence.This contrasts the $\Omega(H)$ horizon-dependent growth under squared Hellinger distance (Rohatgi et al., 2025), which should be regarded not as a fundamental barrier but a metric artifact induced by the choice of divergence. (See Theorem 3.1.)
- **Limitations of *proper* learning with parameter sharing.** For dependent policy class Π, offline ERM is computationally intractable under non-convexity or combinatorial constraints, and even ignoring computation it suffers an estimation barrier of $C_{\text{est}} = \Omega(H^2)$ due to per-trajectory loss range and coupling. Hence, proper learning with shared parameters faces both computational and statistical limitations. (See Theorem 3.2.)
- **Breaking the statistical barrier efficiently with improper Bayesian learning**. We propose a simple improper Bayesian posterior learner that achieves a fast-rate excess joint KL of order $\widetilde{O}(H \log |\Pi|/n)$ over the misspecification error. This yields an $O(H)$ estimation term, improving upon the proper ERM baseline while remaining computationally tractable. (See Theorem 3.3)

Proofs, constants, and extensions are deferred to the appendix.

2 Problem setting

We specify a model class Π to learn $\pi^\star$, allowing for model misspecification ($\pi^\star \notin \Pi$). Without making restrictive assumptions on the true distribution $P^{\pi^\star}$, we analyze autoregressive modeling by characterizing the approximation and estimation errors of learning with Π.

We are given n i.i.d. sequences $(S_1, \cdots, S_n)$, each S_i drawn from $P^{\pi^\star}$ and consists of H tokens $\{u_1^i, \cdots, u_H^i\}$. We aim to use this dataset for the generative modeling objective (1). We denote $s_h = u_{<h}$ and $\pi = (\pi_1, ..., \pi_H)$.

We made a widely used assumption of bounded log density (Xie et al., 2022; Rohatgi et al., 2025).

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

$$\sup_{\pi\in\Pi} \sup_{h\in[H]} \sup_{s_h} \log \frac{\pi_h^*(x_h|s_h)}{\pi_h(x_h|s_h)} \leq G.$$

Throughout this paper, we use the notation $f \lesssim g$ to mean that there exists an absolute constant $c > 0$ such that $f \leq c \cdot g$.

We distinguish between two structural regimes of the policy class. In the *decomposable policy class* case, also referred to as *no parameter sharing* in prior work such as Foster et al. (2024a), each step admits an independent policy component, so that $\Pi = \Pi_1 \times \cdots \times \Pi_H$, with Π_h denoting the set of admissible conditional distributions at step h. In contrast, in the more general *dependent policy class* case, which is the focus of this work, Π is a subset of such a product space, $\Pi \subseteq \Pi_1 \times \cdots \times \Pi_H$ but does not decompose across steps. Dependencies couple the decision rules and often introduce non-convex or combinatorial constraints, complicating both analysis and computation.

3 Main Results

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

We first discuss the traditional offline ERM algorithm, which outputs any minimizer:

$$\hat{\pi} \in \arg\min_{\pi \in \Pi} -\frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \log \pi_h(u_h^i \mid u_1^i, \ldots, u_{h-1}^i). \tag{2}$$

The below theorem shows that empirical log-loss minimization controls the misspecification error under joint KL divergence up to a constant factor.

Theorem 3.1 (ERM: approximation under joint KL)**.** *Let $S_1, \ldots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ with $\pi^\star \notin \Pi$, where Π is a finite model class. Under Assumption 1, offline ERM algorithm* (2) *guarantees that for any $\delta \in (0,1)$ and any $\varepsilon > 0$, with probability at least $1-\delta$,*

$$\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \lesssim \underbrace{(1+\varepsilon) \min_{\pi \in \Pi} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})}_{\textit{approximation error}} + \underbrace{\frac{1+\varepsilon}{\varepsilon} \frac{(H^2G^2+1)\log(|\Pi|\delta^{-1})}{n}}_{\textit{estimation error}}. \tag{3}$$

The first term in (3) shows that the approximation error does not amplify with the horizon (i.e., sequence length): the approximation ratio C_{apx} on $\min\limits_{\pi \in \Pi} \mathrm{KL}(P^{\pi^\star} \| P^{\pi})$ is $(1+\varepsilon)$, independent of H. This follows 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

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

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.

Although the approximation term benefits from the exact chain rule, the estimation error in (3) behaves differently—it exhibits an undesirable H^2-dependence. The H^2 barrier in C_{est} arises because typical fast-rate proofs via a Bernstein-type condition require bounding the *joint* log-likelihood, whose magnitude scales with H. The subsequent theorem shows that this dependence is in fact fundamental with proper learning, even under the easier realizable case.

Theorem 3.2 (Lower bound for proper learning)**.** *There exist absolute constants $c, c_0 > 0$ such that for any $M \ge 4$, any $n \ge c_0 \log |\Pi|$, any horizon $H \in \mathbb{N}$, and any $G \in (0,1]$, one can construct a finite 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 $\pi^\star \in \Pi$,*

$$\sup_{\pi^\star \in \Pi} \mathbb{E}_{P^{\pi^\star}}\Big[\mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\big)\Big] \;\ge\; c\, \frac{H^2G^2 \,\log |\Pi|}{n}. \tag{4}$$

Consequently, an $O\big(\frac{H^2G^2 \log |\Pi|}{n}\big)$ upper bound for ERM is unimprovable in the worst case.

Beyond statistical limitations, ERM also faces computational challenges. When the policy class Π is nonconvex or subject to intricate geometric/combinatorial constraints, optimization and sampling over Π are, in general, computationally inefficient. This motivates us to consider improper learning algorithms, which relax the requirement of staying within Π and thereby offer both statistical and computational advantages.

3.2 Improper Learning via Lifting the Policy Space

Improper Bayesian posterior over the original class Π. To address the statistical barrier of proper learning, we consider a Bayesian posterior predictor directly over the dependent class Π. Under log-loss, the mixability of KL yields horizon-free estimation rates and removes the $\Omega(H^2)$ dependence even under misspecification. Intuitively, the predictor maintains a weighted combination of dependent policies and thus avoids horizon amplification. However, operating at the joint level of Π serves as an idealized benchmark for statistical performance rather than an implementable algorithm. The precise construction and its guarantees are deferred to Appendix B.

Towards computational efficiency. For computational tractability, we enlarge the hypothesis space to the lifted class (mentioned only in passing by Rohatgi et al. (2025))

$$\widetilde{\Pi} := \Pi_1 \times \cdots \times \Pi_H,$$

where each Π_h is defined as the collection of conditional distributions $\pi_h(\cdot \mid s_h)$ brown induced by any policy π in Π at step h. Formally, $\Pi_h := \{\, \pi_h \;\mid\; \pi \in \Pi \,\}$.

A policy in the lifted class is thus represented as a tuple $(\pi_1, \ldots, \pi_H)$, with each $\pi_h \in \Pi_h$ chosen independently. This removes parameter sharing across steps and makes the class stepwise decomposable, allowing for tractable computation. In this lifted space, we apply the Bayesian posterior separately at each step rather than as a whole. Starting from a prior over Π_h, the posterior after observing n trajectories is updated according to the empirical log-likelihood, yielding a mixture

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

where q_h denotes the updated posterior weight over Π_h. This yields a computationally efficient algorithm, and the full pseudocode is deferred to Appendix B. The statistical guarantee of this procedure is given in the following theorem.

Theorem 3.3 (Posterior fast rate under misspecification in the lifted space)**.** *Let $S_1, \ldots, S_n \sim P^{\pi^\star}$ i.i.d. Under Assumption 1, if $\hat{\pi}$ is the per-step Bayesian posterior (exponential-weights) predictor with the uniform prior on $\widetilde{\Pi}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{KL}\Big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\Big) \;\lesssim\; \underbrace{\min_{\pi \in \Pi} \mathrm{KL}\Big(P^{\pi^\star} \,\|\, P^{\pi}\Big)}_{\textit{approximation error}} \;+\; \underbrace{\tilde{O}\Big(\tfrac{HG\,(\log|\Pi| + \log(1/\delta))}{n}\Big)}_{\textit{estimation error}}.$$

Theorem 3.3 shows that by lifting to the product class $\tilde{\Pi}$ and applying per-step Bayesian posteriors, we obtain a computationally efficient procedure whose approximation error remains horizon-free under KL, while the estimation error scales linearly in H. This stands in contrast to the lower bound (4) for proper algorithms (e.g., ERM) as well as the infeasible posterior over Π. The result thus makes the computation–statistics tradeoff explicit: statistical optimality requires an intractable posterior over the dependent class, whereas efficient algorithms necessarily incur an $O(H)$ estimation factor.

While prior analyses (e.g., Rohatgi et al. (2025)) provide valuable insights on approximation error and horizon dependence, our results show that under misspecification characterized by joint KL, the estimation term C_{est} is equally fundamental. Making this explicit reveals that in the dependent-policy regime, feasibility is ultimately governed by the estimation barrier, not approximation.

Discussion. It is important to note that our analysis does not exploit any particular structure of the policy class Π. For certain structured policy classes with at most K switches, the horizon dependence of C_{est} can drop from H to $O(K)$ as shown in Appendix A.2. Therefore, there is the potential to design more sophisticated algorithms that fully leverage the geometry of Π to reduce the horizon dependence of C_{est}, thereby narrowing the gap between computational tractability and statistical efficiency.

References

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.

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.

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

A Proofs of Main Results

A.1 Proof from subsection 3.1

We prove a generalization of Theorem 3.1 that allows for infinite policy class. To recover the finite class case, just let $r = 0$.

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

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

Theorem A.1 (generalization of Theorem 3.1)**.** *Let* $S_1, \ldots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$ *with* $\pi^\star \notin \Pi$*, where* Π *has a finite log-ratio cover. Under assumption 1, offline ERM algorithm guarantees that for any* $\delta \in (0, 1)$ *and any* $\varepsilon > 0$*, with probability at least* $1 - \delta$*,*

$$\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \lesssim \mathcal{O}\left((1+\varepsilon) \min_{\pi \in \Pi} \mathrm{KL}(P^{\pi^\star} \| P^{\pi}) + \frac{1+\varepsilon}{\varepsilon} \frac{(H^2G^2+1)\log(|\mathcal{N}_{\log}(\Pi, r)|\delta^{-1})}{n} + Hr\right).$$

Proof of A.1. 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 one representative $\pi^r \in \Pi_r$ such that for every length-H trajectory S,

$$\left|\log \frac{P^{\pi}(S)}{P^{\pi^r}(S)}\right| \;\leq\; Hr. \tag{5}$$

For any π, we set

$$Y(\pi) := \log \frac{P^{\pi^\star}(S)}{P^{\pi}(S)}, \qquad \mu_\pi := \mathbb{E}_{P^{\pi^\star}}[Y(\pi)] = \mathrm{KL}(P^{\pi^\star} \| P^{\pi}), \qquad \hat{\mu}_\pi := \tfrac{1}{n}\sum_{i=1}^{n} Y_i(\pi),$$

and let $\mu_\star := \min_{\pi \in \Pi} \mu_\pi$. From $\mathbb{E}_{P^{\pi^\star}}[e^{-Y}] = 1$, $|Y| \leq HG$, and the third-order Taylor lower bound for e^{-y},

$$e^{-Y} \geq 1 - Y + \frac{Y^2}{2} - \frac{Y^3}{6},$$

one obtains the Bernstein-type condition $\mathrm{Var}(Y) \leq V(G)\,\mu$ with $V(G) = 2 + \frac{H^2G^2}{3}$, since $\mathbb{E}Y^3 \leq H^2G^2\,\mathbb{E}Y$. Hence the one-sided Bernstein inequality implies that, for any $\delta_g \in (0, 1)$, with probability at least $1 - \delta_g$,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2V(G)\,\mu\,\log(1/\delta_g)}{n}} + \frac{4G}{3}\,\frac{\log(1/\delta_g)}{n}. \tag{6}$$

Apply (6) simultaneously to all $\pi' \in \Pi_r$ and to $\hat{\pi}_r \in \arg\min_{\pi' \in \Pi_r} \hat{\mu}_{\pi'}$; after a union bound with $\delta_g = \delta/|\Pi_r|$ we get, with probability at least $1 - \delta$ and for all $\pi' \in \Pi_r$,

$$\Delta_r(\pi') \;\leq\; \widehat{\Delta}_r(\pi') \;+\; a_r\big(\sqrt{\mu_{\pi'}} + \sqrt{\mu_\star^r}\big) \;+\; 2b_r \tag{7}$$

where

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

and $\mu_\star^r := \min_{\pi' \in \Pi_r} \mu_{\pi'}$, $\Delta_r(\pi') := \mu_{\pi'} - \mu_\star^r$, and $\widehat{\Delta}_r(\pi') := \hat{\mu}_{\pi'} - \hat{\mu}_\star^r$.

By (5) we have that,

$$|Y(\pi) - Y(\pi^r)| \leq Hr \quad \Rightarrow \quad |\mu_\pi - \mu_{\pi^r}| \leq Hr, \qquad |\hat{\mu}_\pi - \hat{\mu}_{\pi^r}| \leq Hr. \tag{8}$$

Since $\hat{\pi}$ is an ERM on Π and $\Pi_r \subset \Pi$, combining (8) with $\min_{\pi' \in \Pi_r} \hat{\mu}_{\pi'} \geq \min_{\pi \in \Pi} \hat{\mu}_\pi - Hr = \hat{\mu}_{\hat{\pi}} - Hr$ and $\hat{\mu}_{\hat{\pi}^r} \leq \hat{\mu}_{\hat{\pi}} + Hr$ yields

$$\widehat{\Delta}_r(\hat{\pi}^r) = \hat{\mu}_{\hat{\pi}^r} - \hat{\mu}_\star^r \leq 2Hr. \tag{9}$$

Applying (7) to $\pi' = \hat{\pi}^r$ and using $\mu_{\hat{\pi}^r} = \Delta_r(\hat{\pi}^r) + \mu_\star^r$ and $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$,

$$\Delta_r(\hat{\pi}^r) \;\le\; a_r\Big(\sqrt{\Delta_r(\hat{\pi}^r)} + 2\sqrt{\mu_\star^r}\Big) + 2b_r + 2Hr. \tag{10}$$

Solving $x \le a_r\sqrt{x} + 2a_r\sqrt{\mu_\star^r} + 2b_r + 2Hr$ gives

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

By (9) and $\Pi_r \subset \Pi$, $\mu_\star \le \mu_\star^r \le \mu_\star + Hr$ and

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

Combining with (10) and writing $|\Pi_r| = \mathcal{N}_{\log}(\Pi, r)$,

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

Finally, apply the A–G inequality $2\sqrt{xy} \le \varepsilon x + \frac{y}{\varepsilon}$ to the square-root term (with $x = \mu_\star^r$ and $y = 8V(G)\frac{\log(\mathcal{N}_{\log}(\Pi,r)\delta^{-1})}{n}$) and use $\mu_\star^r \le \mu_\star + Hr$ to obtain, for any $\varepsilon > 0$,

$$\mathrm{KL}\big(P^{\pi^\star}\|P^{\hat{\pi}}\big) = \mu_\star + \Delta(\hat{\pi}) \;\le\; (1+\varepsilon)\,\mu_\star \;+\; \frac{C}{\varepsilon}\,(H^2G^2+1)\,\frac{\log\!\big(\mathcal{N}_{\log}(\Pi,r)\,\delta^{-1}\big)}{n} \;+\; C'Hr,$$

for universal constants C, C'. This proves the high-probability bound for infinite Π. □

Now we turn to the proof overview for Theorem 3.2. We establish the lower bound by an information–theoretic *packing* argument combined with Fano's inequality. We build a small family of policies that are well separated yet individually close to the class average, so the data reveal only limited information about which one is true. Under proper learning the estimator must pick a single policy; when the information is too small, Fano forces a constant chance of picking the wrong one, which yields the desired lower bound combining with the separation. The proof below instantiates this outline.

Proof of Theorem 3.2: Let the sample space be $\mathcal{S} = \{1,\dots,d\}$ with $d \ge |\Pi|$. Choose $|\Pi|$ vectors $v^{(1)},\dots,v^{(|\Pi|)} \in \{\pm1\}^d$ that are pairwise orthogonal and balanced (e.g., columns of a Hadamard matrix, padded if needed). For $\varepsilon > 0$, define

$$P^{\pi_v}(s) = \frac{\exp(\varepsilon v_s)}{\sum_{k=1}^d \exp(\varepsilon v_k)} = \frac{e^{\varepsilon v_s}}{d\cosh\varepsilon} \qquad (s \in \mathcal{S}).$$

Then, for any v, w and any s,

$$\log\frac{P^{\pi_v}(s)}{P^{\pi_w}(s)} = \varepsilon\,(v_s - w_s) \in \{-2\varepsilon, 0, 2\varepsilon\}.$$

Taking $\varepsilon \le HG/2$ gives the required two–sided ratio bound.

Let the average (uniform) distribution be

$$\bar{P}(s) = \frac{1}{|\Pi|}\sum_v P^{\pi_v}(s) = \frac{1}{d}.$$

A direct calculation using orthogonality yields, for $v \ne w$,

$$\mathrm{KL}(P^{\pi_v}\,\|\,P^{\pi_w}) = \varepsilon\tanh\varepsilon, \qquad \mathrm{KL}\big(P^{\pi_v}\,\|\,\bar{P}\big) = \varepsilon\tanh\varepsilon - \log\cosh\varepsilon.$$

Using $\tanh x \ge x/2$ and $\log\cosh x \le x^2/2$ for $x \in (0,1]$, we have

$$\mathrm{KL}(P^{\pi_v}\,\|\,P^{\pi_w}) \ge \tfrac{1}{2}\varepsilon^2, \qquad \mathrm{KL}\big(P^{\pi_v}\,\|\,\bar{P}\big) \le \tfrac{1}{2}\varepsilon^2. \tag{11}$$

We use Fano's information inequality here. Let π be uniform on $\{1,\dots,|\Pi|\}$ and observe $S_{1:n} \sim (P^\pi)^{\otimes n}$. For any estimator $\hat{\pi}$, Fano's inequality gives that

$$\Pr(\hat{\pi} \ne \pi) \;\ge\; 1 - \frac{I(\pi; S_{1:n}) + \log 2}{\log|\Pi|}. \tag{12}$$

By the chain rule and (11),

$$I(\pi; S_{1:n}) \le \sum_{i=1}^{n} \max_{v} \mathrm{KL}\big(P^{\pi_v} \,\|\, \bar{P}\big) \le \frac{n}{2}\varepsilon^2. \tag{13}$$

Choose

$$\varepsilon := \frac{HG}{4}\sqrt{\frac{\log|\Pi|}{n}},$$

and assume $n \ge c_0 \log|\Pi|$ so that $\varepsilon \le \min\{HG/2, 1\}$. Then (12)–(13) imply, for $|\Pi|$ large enough,

$$\Pr(\hat{\pi} \ne \pi) \;\ge\; 1 - \frac{H^2G^2}{32} - \frac{\log 2}{\log|\Pi|} \;\ge\; \frac{1}{2}. \tag{14}$$

Notice that whenever $\hat{\pi} \ne \pi$, by (11),

$$\mathrm{KL}(P^{\pi} \,\|\, P^{\pi_{\hat{\pi}}}) \;\ge\; \tfrac{1}{2}\varepsilon^2.$$

Therefore,

$$\sup_{\pi^\star \in \Pi} \; \mathbb{E}_{P^{\pi^\star}}\Big[\mathrm{KL}\Big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\Big)\Big] \;\ge\; \Pr(\hat{\pi} \ne \pi) \cdot \frac{\varepsilon^2}{2} \;\overset{(14)}{\ge}\; \frac{1}{4} \cdot \frac{1}{2}\varepsilon^2 \;=\; \frac{1}{64}\,\frac{H^2G^2 \log|\Pi|}{n},$$

which proves the claim with $c = 1/64$. □

A.2 Proof from subsection 3.2

As before, we consider the more general case of an *infinite* policy class Π.

Theorem A.2 (generalization of Theorem 3.3)**.** *Let $S_1, \ldots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$. Fix $r > 0$ and, for each step $h \in [H]$, let $\Pi_{h,r} \subset \Pi_h$ be an r-cover under the log–ratio metric with $\mathcal{N}_{\log}(\Pi_h, r) := |\Pi_{h,r}|$ and $\mathcal{N}_{\log}(\Pi, r) := \max_{h\in[H]} \mathcal{N}_{\log}(\Pi_h, r)$. If $\hat{\pi}$ is the per-step Bayesian posterior (exponential-weights) predictor with the uniform prior on $\Pi_{h,r}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\big) \;\lesssim\; \underbrace{\min_{\pi\in\Pi} \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi}\big)}_{\textit{approximation error}} \;+\; \underbrace{Hr \;+\; \tilde{O}\bigg(\frac{HG\{\log \mathcal{N}_{\log}(\Pi, r) + \log(1/\delta)\}}{n}\bigg)}_{\textit{estimation error}}.$$

Proof. 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 $i = 1, \ldots, n$. For each $h \in [H]$, let $\Pi_h = \{\pi_h(\cdot \,|\, s_{1:h-1}) : \pi \in \Pi\}$ and 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 $(u, s_{1:h-1})$,

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

Write $\mathcal{N}_{\log}(\Pi_h, r) := |\Pi_{h,r}|$ and $\mathcal{N}_{\log}(\Pi, r) := \max_{h\in[H]} \mathcal{N}_{\log}(\Pi_h, r)$. 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 $\pi_h^r \in \Pi_{h,r}$,

$$\sum_{i=1}^{n} \big[-\log \hat{\pi}_h(u_h^i \,|\, s_{1:h-1}^i)\big] \;\le\; \sum_{i=1}^{n} \big[-\log \pi_h^r(u_h^i \,|\, s_{1:h-1}^i)\big] \;+\; \log \mathcal{N}_{\log}(\Pi_h, r). \tag{15}$$

Summing (15) 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

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

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

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

Hence, minimizing the right-hand side of (16) over $\pi \in \Pi$ yields

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

Let $Z_{i,h}(\pi_h) = \log\!\big(\pi_h^\star(u_h^i \mid s_{1:h-1}^i)/\hat\pi_h(u_h^i \mid s_{1:h-1}^i)\big) - \mathbb{E}[\,\cdot\,]$ be the centered log-likelihood ratio under $P^{\pi^\star}$. By Assumption 1 (bounded log-density), the range/variance of $Z_{i,h}(\cdot)$ is controlled by G, hence Bernstein/Freedman inequality and a union bound over h imply that with probability at least $1-\delta$,

$$\left| \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H} \log \frac{\pi_h^\star(u_h^i \mid s_{1:h-1}^i)}{\hat\pi_h(u_h^i \mid s_{1:h-1}^i)} - \sum_{h=1}^{H} \mathbb{E}\!\left[\log \frac{\pi_h^\star(U_h \mid S_{1:h-1})}{\hat\pi_h(U_h \mid S_{1:h-1})} \right] \right| \;\lesssim\; \tilde O\!\left(\frac{HG \log(1/\delta)}{n} \right).$$

Taking expectations in (17) and using the KL chain rule, the left-hand side becomes $\mathrm{KL}(P^{\pi^\star} \,\|\, P^{\hat\pi})$. Therefore,

$$\mathrm{KL}(P^{\pi^\star} \,\|\, P^{\hat\pi}) \;\le\; \min_{\pi\in\Pi} \mathrm{KL}(P^{\pi^\star} \,\|\, P^{\pi}) \;+\; Hr \;+\; \frac{H \log \mathcal{N}_{\log}(\Pi, r)}{n} \;+\; \tilde O\!\left(\frac{HG \log(1/\delta)}{n} \right),$$

and absorbing constants/logs into $\tilde O(\cdot)$ gives the claim. □

Theorem A.3 (generalization of Theorem 3.3 with K-switch and Fixed-Share)**.** *Let $S_1, \dots, S_n \overset{\text{iid}}{\sim} P^{\pi^\star}$. Fix $K \in \{0, 1, \dots, H-1\}$ and let Π be a finite base policy class. Define the S-switch sequence class*

$$\Pi_{\text{seq}}^{\le K} = \Big\{ \pi = (\pi_1, \dots, \pi_H) : \; \pi_h \in \Pi, \; \#\{h \in [H-1] : \pi_{h+1} \ne \pi_h\} \le K \Big\}.$$

Let q_α be the Markov prior on $\Pi_{\text{seq}}^{\le K}$ with share-rate $\alpha \in [0,1]$, given by

$$q_\alpha(\pi) = \tfrac{1}{|\Pi|}\,(1-\alpha)^{H-1-K} \Big(\alpha/(|\Pi|-1) \Big)^{K}, \qquad \pi \in \Pi_{\text{seq}}^{\le K}.$$

If $\hat\pi = (\hat\pi_1, \dots, \hat\pi_H)$ is the per-step Bayesian posterior (exponential-weights) predictor generated by the Fixed-Share update in Algorithm 3, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\hat\pi}\big) \;\lesssim\; \underbrace{\min_{\pi \in \Pi_{\text{seq}}^{\le K}} \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi}\big)}_{\textit{approximation error}} \;+\; \underbrace{\tilde O\!\left(\frac{G\{-\log q_{\alpha^\star}(\pi^\star) + \log(1/\delta)\}}{n} \right)}_{\textit{estimation error}},$$

where $\pi^\star$ denotes any minimizer of the approximation term and $\alpha^\star = K/(H-1)$. Moreover,

$$-\log q_{\alpha^\star}(\pi^\star) \;\le\; \log|\Pi| + K\log(|\Pi|-1) + K \log\tfrac{eH}{K}.$$

Proof. Let $S_i = (s_1^i, u_1^i, \dots, s_H^i, u_H^i)$ be i.i.d. trajectories from $P^{\pi^\star}$ for $i = 1, \dots, n$. For any sequence $\pi = (\pi_1, \dots, \pi_H) \in \Pi_{\text{seq}}^{\le K}$ write $P^\pi(S) = \prod_{h=1}^H \pi_h(u_h \mid s_{1:h-1})$. Let q_α be the Fixed-Share prior above, and let $P^{\hat\pi}$ be the per-step predictor returned by Algorithm 3.

(1) Mixability inequality at the trajectory level. By log-loss mixability (the Bayes inequality for exponential weights), for any fixed dataset $S_{1:n}$ and any $\pi \in \Pi_{\text{seq}}^{\le K}$,

$$\sum_{i=1}^{n} \big[-\log P^{\hat\pi}(S_i) \big] \;\le\; \sum_{i=1}^{n} \big[-\log P^{\pi}(S_i) \big] \;+\; \big[-\log q_\alpha(\pi) \big]. \tag{18}$$

Adding $\sum_i \log P^{\pi^\star}(S_i)$ to both sides and dividing by n yields

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{P^{\pi^\star}(S_i)}{P^{\hat\pi}(S_i)} \;\le\; \min_{\pi \in \Pi_{\text{seq}}^{\le K}} \frac{1}{n}\sum_{i=1}^{n} \log \frac{P^{\pi^\star}(S_i)}{P^{\pi}(S_i)} \;+\; \frac{-\log q_\alpha(\pi)}{n}. \tag{19}$$

(2) Concentration from empirical to expected log-likelihood ratio. Let

$$L_i(\pi) = \log \frac{P^{\pi^\star}(S_i)}{P^{\pi}(S_i)} - \mathbb{E}\left[\log \frac{P^{\pi^\star}(S)}{P^{\pi}(S)}\right]$$

be the centered log-likelihood ratio under $P^{\pi^\star}$. By Assumption 1 (bounded log-density), the range and variance of $L_i(\pi)$ are controlled by G. Hence, by Freedman's inequality and a union bound over $\pi \in \Pi_{\text{seq}}^{\le K}$, with probability at least $1-\delta$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{P^{\pi^\star}(S_i)}{P^{\pi}(S_i)} - \mathbb{E}\left[\log \frac{P^{\pi^\star}(S)}{P^{\pi}(S)}\right] \right| \lesssim \tilde{O}\left(\frac{G \log(1/\delta)}{n}\right). \tag{20}$$

Combining (20) with (19) and minimizing over $\pi \in \Pi_{\text{seq}}^{\le K}$ gives

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{P^{\pi^\star}(S_i)}{P^{\hat{\pi}}(S_i)} \le \min_{\pi \in \Pi_{\text{seq}}^{\le S}} \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi}\big) + \frac{-\log q_\alpha(\pi)}{n} + \tilde{O}\left(\frac{G \log(1/\delta)}{n}\right).$$

(3) Taking expectations and optimizing the share rate. Taking expectations in the left-hand side and using the KL chain rule, the left-hand side becomes $\mathrm{KL}(P^{\pi^\star} \,\|\, P^{\hat{\pi}})$. Choosing $\alpha^\star = K/(H-1)$ minimizes the prior complexity, giving

$$-\log q_{\alpha^\star}(\pi) \le \log|\Pi| + K\log(|\Pi|-1) + K\log\tfrac{eH}{K},$$

and absorbing constants/logs into $\tilde{O}(\cdot)$ yields the claim. □

Remark (global posterior on Π removes the factor H in the model-count term). If, instead of per-step mixing, we place a *single* uniform prior on the whole class Π and form the (trajectory-level) Bayesian predictor over Π, the mixability step is applied once at the sequence level. The analogue of (15) then yields a single complexity term $\log|\Pi|$ (or $\log \mathcal{N}_{\log}(\Pi, r)$ in the infinite case), so that the counterpart of (17) replaces $\sum_{h=1}^{H} \log|\Pi_h|$ by $\log|\Pi|$. Consequently, the generalization bound contains *only* $\frac{1}{n}\log|\Pi|$ (rather than $\frac{H}{n}\log|\Pi|$), while the deviation term remains $\tilde{O}\big(\frac{HG\log(1/\delta)}{n}\big)$.

B Algorithms and Computation–Statistics Tradeoffs

This section collects the algorithms used in our analysis and aligns it with the two axes considered in the paper: (i) *learning rule* (offline ERM vs. Bayesian posterior sampling) and (ii) *model base* (the dependent class Π vs. the lifted product class $\tilde{\Pi} = \Pi_1 \times \cdots \times \Pi_H$).

Algorithm 1 offline ERM for Autoregressive Modeling (joint vs. per-step)

```
Require: Dataset $\mathcal{D} = \{S_i\}_{i=1}^n$, where $S_i = (u_1^i, \dots, u_H^i)$
Require: Base $B \in \{\Pi, \tilde{\Pi}\}$ with $\tilde{\Pi} = \Pi_1 \times \cdots \times \Pi_H$ and $\Pi_h = \{\pi_h(\cdot \mid s_{1:h-1}) : \pi \in \Pi\}$
For each $i$ and $h$, set $s_{1:h-1}^i \leftarrow (u_1^i, \dots, u_{h-1}^i)$
if $B = \Pi$ then                                          ▷ proper; joint ERM on the dependent class
    $L_n(\pi) \leftarrow -\frac{1}{n}\sum_{i=1}^n \sum_{h=1}^H \log \pi_h(u_h^i \mid s_{1:h-1}^i)$
    return $\hat{\pi} \in \arg\min_{\pi \in \Pi} L_n(\pi)$
else                                                       ▷ $B = \tilde{\Pi}$; improper per-step ERM on the lifted class
    for $h = 1$ to $H$ do
        $L_{n,h}(\pi_h) \leftarrow -\frac{1}{n}\sum_{i=1}^n \log \pi_h(u_h^i \mid s_{1:h-1}^i)$
        $\hat{\pi}_h \in \arg\min_{\pi_h \in \Pi_h} L_{n,h}(\pi_h)$
    end for
    return $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_H)$
end if
```

ERM: joint vs. per-step. Algorithm 1 offers two training routes: a *joint* ERM on the dependent class Π (proper) and a *per-step* ERM on the lifted class $\tilde{\Pi}$ (improper). Both achieve constant-level approximation error, but they differ in estimation complexity. The joint route pays the capacity penalty once, yielding an estimation term of the form $C_{\text{est}}^{\text{joint}} \sim \frac{\log \mathcal{N}_{\log}(\Pi)}{n}$, yet it is typically computationally challenging due to cross-step coupling. The per-step route decomposes training into H independent problems, which is parallelizable and practical, but the estimation penalty accumulates across steps, $C_{\text{est}}^{\text{per-step}} \sim \frac{H\ \log \mathcal{N}_{\log}(\Pi,r)}{n}$. Thus, one trades tighter statistics for harder optimization in the joint case, versus tractable computation at the cost of an H-linear estimation term in the per-step case.

Algorithm 2 Bayesian Posterior (joint vs. per-step) for Autoregressive Modeling

```
Require: Dataset $\mathcal{D} = \{S_i\}_{i=1}^n$; base $B \in \{\Pi, \tilde{\Pi}\}$; prior $w_0$ on $B$
if $B = \Pi$ then                                  ▷ idealized joint posterior on the dependent class
    $w_n(\pi) \propto w_0(\pi) \prod_{i=1}^n P^\pi(S_i)$
    $P^{\hat{\pi}}(S) = \int_\Pi P^\pi(S)\, w_n(\pi)\, d\pi$
    return predictor $P^{\hat{\pi}}$
else                                               ▷ $B = \tilde{\Pi}$; per-step posterior on the lifted class
    for $h = 1$ to $H$ do
        Initialize $q_h^{(0)} \leftarrow \mathrm{Unif}(\Pi_h)$
        $\ell_h(\pi_h) \leftarrow \sum_{i=1}^n [-\log \pi_h(u_h^i \mid s_{1:h-1}^i)]$
        $q_h^{(n)}(\pi_h) \propto q_h^{(0)}(\pi_h) \exp\big(-\ell_h(\pi_h)\big)$
        $\hat{\pi}_h(\cdot \mid s) \leftarrow \sum_{\pi_h \in \Pi_h} q_h^{(n)}(\pi_h)\, \pi_h(\cdot \mid s)$
    end for
    return $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_H)$
end if
```

Algorithm 3 Fixed-Share Bayesian Posterior for K-switch Policies (extension of Alg. 2)

```
Require: Dataset $\mathcal{D} = \{S_i\}_{i=1}^n$; base $B \in \{\Pi, \tilde{\Pi}\}$; prior $w_0$ on $B$; share-rate $\alpha \in [0, 1]$
if $B = \Pi$ then                                  ▷ idealized joint posterior on the dependent class
    $w_n(\pi) \propto w_0(\pi) \prod_{i=1}^n P^\pi(S_i)$
    $P^{\hat{\pi}}(S) = \int_\Pi P^\pi(S)\, w_n(\pi)\, d\pi$
    return predictor $P^{\hat{\pi}}$
else                                               ▷ $B = \tilde{\Pi}$; per-step posterior on the lifted class with Fixed-Share
    for $h = 1$ to $H$ do
        if $h = 1$ then
            $q_h^{(0)} \leftarrow \mathrm{Unif}(\Pi_h)$
        else
            $q_h^{(0)} \leftarrow (1 - \alpha)\, q_{h-1}^{(n)} + \alpha\, \mathrm{Unif}(\Pi_h)$      ▷ share initialization (allows switching)
        end if
        $\ell_h(\pi_h) \leftarrow \sum_{i=1}^n [-\log \pi_h(u_h^i \mid s_{1:h-1}^i)]$
        $q_h^{(n)}(\pi_h) \propto q_h^{(0)}(\pi_h) \exp[-\ell_h(\pi_h)]$
        $\hat{\pi}_h(\cdot \mid s) \leftarrow \sum_{\pi_h \in \Pi_h} q_h^{(n)}(\pi_h)\, \pi_h(\cdot \mid s)$
    end for
    return $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_H)$
end if
```

Bayesian Posterior: joint vs. per-step. In Algorithm 2, the *joint* Bayesian posterior algorithm forms a single posterior over Π. Under mixability, it incurs only a one-shot complexity penalty, $C_{\text{est}}^{\text{joint}} \sim \frac{\log \mathcal{N}_{\log}(\Pi)}{n}$, and provides horizon-free behavior under KL, but maintaining this posterior is generally infeasible at scale. The *per-step* version updates posteriors on $\tilde{\Pi}$ step by step, dramatically lowering compute and memory, while the estimation term grows additively with the horizon, $C_{\text{est}}^{\text{per-step}} \sim \frac{H\ \log \mathcal{N}_{\log}(\Pi,r)}{n}$. This makes the computation–statistics tradeoff explicit: choose joint when a joint posterior/optimizer is computationally within reach; otherwise use per-step to gain scalability with a controlled, linear-in-H statistical cost.

C Discussion about Improper Learning: Mixing and Lifting

We explore two complementary ways to introduce *improper* learning in our autoregressive setting: (i) *mixing*—predicting with a Bayesian (exponential–weights) posterior rather than selecting a single policy; and (ii) *lifting*—expanding the hypothesis space to the product class $\tilde{\Pi} = \Pi_1 \times \cdots \times \Pi_H$ so that optimization and posterior updates can be carried out stepwise. Both routes keep the approximation component at a constant level, but they impact the estimation complexity C_{est} and computation in different ways.

Mixing over the dependent class Π. Under log-loss, the mixability of KL implies that a posterior predictor built *directly over* Π removes the statistical barrier $\Omega(H^2)$ that appears for proper ERM under misspecification: the estimation term is *horizon-free* (one-shot complexity over Π). Conceptually, mixing preserves cross-step dependence and eliminates horizon amplification. Computationally, however, maintaining a joint posterior on Π is generally infeasible: Π is non-convex, high-dimensional, and subject to cross-step constraints in the worst case. Hence this construction is best viewed as a statistical benchmark rather than a practical algorithm.

Lifting to the product class $\tilde{\Pi}$. To regain tractability, we lift to $\tilde{\Pi}$ and perform per-step policy search—either by ERM or by a per-step posterior. This decouples steps and turns the joint problem into H independent subproblems that are parallelizable and memory-friendly. The price is an $O(H)$ estimation term, e.g. $C_{\text{est}} \sim \frac{H \,\log \mathcal{N}_{\log}(\Pi, r)}{n}$ in our high-probability bounds. Thus, lifting makes the method computationally feasible while introducing a linear-in-H statistical cost.

In what follows, we highlight two considerations that favor posterior mixing over ERM in the lifted space. First, *ERM is not sharp*: even for tiny classes, the approximation ratio C_{apx} cannot decrease to 1, which inherently introduce a tradeoff between approximation error and estimation error. (See Proposition 1). In contrast, mixing avoids this selection noise. Second, Bayesian posterior potentially opens algorithmic handles to *beat the naive $O(H)$ factor* when Π has structure.

Proposition 1 (ERM is non-sharp under misspecification)**.** *There exist absolute constants $c_1, c_2 > 0$ and a two-policy class $\Pi = \{\pi_1, \pi_2\}$ such that, for n large enough and some misspecified truth $P^{\pi^\star}$,*

$$\mathbb{E}\Big[\mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\big)\Big] \;\geq\; \mu_\star \;+\; c_1\sqrt{\frac{\mu_\star}{n}}, \qquad \mu_\star := \min_{\pi \in \Pi} \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi}\big) = \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi_1}\big).$$

Consequently, no uniform oracle inequality of the form $\mathrm{KL}(P^{\pi^\star} \| P^{\hat{\pi}}) \leq \mu_\star + C/n$ can hold for ERM. Hence no sharpening of the $(1+\varepsilon)\mu_\star$ leading constant in Theorem 3.1 down to 1 is possible for ERM.

Proof. Let the sample space be $\mathcal{S} = \{0, 1\}$. Denote the logistic $\sigma(x) = \frac{e^x}{1+e^x}$ and $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Fix $a, \tau > 0$ (to be set below) and define two candidate policies (single-step Bernoulli models):

$$P^{\pi_1}(1) = \sigma(a), \qquad P^{\pi_2}(1) = \sigma(-a), \qquad \text{and the truth} \;\; P^{\pi^\star}(1) = \sigma\Big(\frac{a}{2} + \tau\Big).$$

Let

$$Z \;:=\; \log\frac{P^{\pi_1}(S)}{P^{\pi_2}(S)} \in \{\pm a\}, \qquad \bar{Z}_n := \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

$$m := \mathbb{E}[Z] \;=\; a\big(2P^{\pi^\star}(1) - 1\big) \;=\; a\,\tanh\Big(\frac{a/2 + \tau}{2}\Big).$$

By definition of ERM,

$$\hat{\pi} \;=\; \arg\max_{\pi \in \{\pi_1, \pi_2\}} \frac{1}{n}\sum_{i=1}^{n} \log P^{\pi}(S_i) \;=\; \begin{cases} \pi_1, & \bar{Z}_n > 0, \\ \pi_2, & \bar{Z}_n \leq 0. \end{cases}$$

Hence

$$\Pr(\hat{\pi} = \pi_2) = \Pr(\bar{Z}_n \leq 0) = \Pr\left(\frac{\sqrt{n}(\bar{Z}_n - m)}{\sqrt{a^2 - m^2}} \leq -\frac{\sqrt{n}\,m}{\sqrt{a^2 - m^2}}\right). \tag{21}$$

The excess KL when misselecting π_2 equals the mean log-likelihood ratio:

$$\Delta \;:=\; \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi_2}\big) - \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi_1}\big) = \mathbb{E}\left[\log\frac{P^{\pi_1}(S)}{P^{\pi_2}(S)}\right] = m.$$

Therefore

$$\mathbb{E}\Big[\mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\big)\Big] = \mathrm{KL}\big(P^{\pi^\star} \,\|\, P^{\pi_1}\big) + \Pr(\hat{\pi} = \pi_2) \cdot \Delta = \mu_\star + \Pr(\hat{\pi} = \pi_2)\, m. \tag{22}$$

Set

$$a = \frac{\kappa}{\sqrt{n}}, \qquad \tau = \frac{\gamma}{\sqrt{n}}, \quad \text{with fixed } \kappa, \gamma \in (0, 1/5].$$

Then $m = a \tanh\big((a/2 + \tau)/2\big)$ and $\sqrt{n}\, m/\sqrt{a^2 - m^2}$ is bounded by an absolute constant depending only on κ, γ. Applying Berry–Esseen (or Slud's inequality for binomial tails) to (21) yields: there exist absolute constants $c_0, \beta \in (0, 1)$ such that whenever $\tau\sqrt{n} \le c_0$ (which holds by construction) and n is large,

$$\Pr(\hat{\pi} = \pi_2) \ \ge \ \beta. \tag{23}$$

Because $P^{\pi_1}(1), P^{\pi^\star}(1) \in [1/4, 3/4]$ for our choice of κ, γ and large n, Bernoulli KL enjoys quadratic bounds (standard Taylor/Lipschitz curvature): there exist absolute constants $c_{\text{low}}, c_{\text{up}} > 0$ such that, for $\delta := P^{\pi^\star}(1) - P^{\pi_1}(1)$,

$$c_{\text{low}}\, \delta^2 \ \le \ \text{KL}\big(P^{\pi^\star} \,\|\, P^{\pi_1}\big) \ \le \ c_{\text{up}}\, \delta^2. \tag{24}$$

By the mean value theorem,

$$\delta = \sigma\Big(\frac{a}{2} + \tau\Big) - \sigma(a) = \sigma'(\xi)\Big(\tau - \frac{a}{2}\Big) \quad \text{for some } \xi \in \Big[a, \frac{a}{2} + \tau\Big] \subset \Big[-\frac{1}{5}, \frac{1}{5}\Big],$$

hence $\sigma'(\xi) \in [\sigma_0, 1/4]$. Therefore, from (24),

$$\mu_\star = \text{KL}\big(P^{\pi^\star} \,\|\, P^{\pi_1}\big) \ \asymp \ \Big(\tau - \frac{a}{2}\Big)^2 \ \asymp \ \frac{(\gamma - \kappa/2)^2}{n}. \tag{25}$$

Since $x \mapsto \tanh x$ is concave and $\tanh x \ge x/2$ for $x \in [0, 1]$, and here $x = (a/2 + \tau)/2 \le \kappa/(2\sqrt{n}) + \gamma/(2\sqrt{n}) \le 1$ for n large, we obtain

$$m = a \tanh\Big(\frac{a/2 + \tau}{2}\Big) \ \ge \ \frac{a}{2}\Big(\frac{a}{2} + \tau\Big) \ \ge \ \frac{a\tau}{4} = \frac{\kappa\gamma}{4n}.$$

Plugging (23) and the last display into (22) gives

$$\mathbb{E}\Big[\text{KL}\big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\big)\Big] \ \ge \ \mu_\star + \beta\, \frac{\kappa\gamma}{4n}.$$

On the other hand, by (25) we have $\sqrt{\mu_\star/n} \le C_1 \frac{\max\{\kappa,\gamma\}}{n}$ for an absolute $C_1 > 0$ (taking $\kappa \le \gamma$). Hence

$$\mathbb{E}\Big[\text{KL}\big(P^{\pi^\star} \,\|\, P^{\hat{\pi}}\big)\Big] \ \ge \ \mu_\star \ + \ \Big(\frac{\beta\,\kappa}{4C_1}\Big)\sqrt{\frac{\mu_\star}{n}} \ \ge \ \mu_\star \ + \ c_1\sqrt{\frac{\mu_\star}{n}},$$

which proves the claim with an absolute constant $c_1 > 0$. □