A Unified Perspective on Task Retrieval and Learning in In-Context Learning on Markov Data*

Wendong Li*
University of Macau & University of Bonn

Huikang Liu*
Shanghai Jiao Tong University

Zhiyuan Huang
Shenzhen Loop Area Institute

Po Chen
Fudan University

Rujun Jiang
Fudan University

Peng Wang†
University of Macau

Abstract

In-context learning enables pretrained transformers to adapt from a prompt without parameter updates, but it remains unclear when this behavior arises from retrieving memorized pretraining tasks (i.e., *task retrieval*) versus learning a rule from in-context data (i.e., *task learning*). We study this distinction through a finite-mixture Markov data model, where each latent Markov chain represents a pretraining task. We characterize the finite-prior Bayes predictor induced by next-token pretraining and use it to explain when pretrained transformers exhibit retrieval or learning. In low-diversity regimes, this predictor retrieves the pretraining Markov chain most compatible with the prompt, with error decaying exponentially in the prompt length. This mechanism is realizable by a two-layer single-head causal transformer. In high-diversity regimes, the finite-prior Bayes predictor approaches the empirical Markov predictor inferred from the prompt, and architectural constraints can bias globally trained transformers toward this capacity-efficient task-learning mechanism, with error decreasing as the prompt length increases. Finally, we extend our analysis to multi-state higher-order Markov chains.

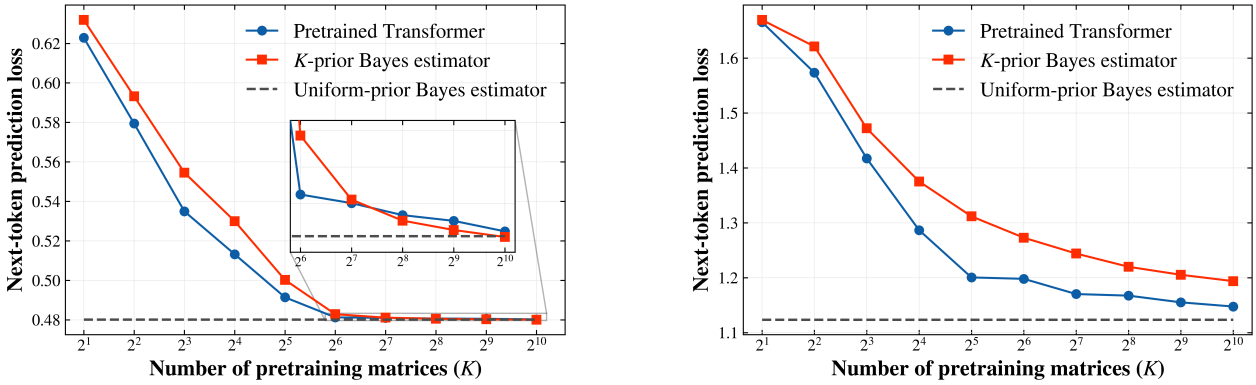
1. Introduction

Over the past few years, transformers (Vaswani et al., 2017), powered by the attention mechanism, have become one of the most widely used architectures in modern deep learning and achieved remarkable empirical success across a wide range of domains, including computer vision (Bao et al., 2023; Chen et al., 2020), natural language processing (Brown et al., 2020b; Devlin et al., 2019), and scientific machine learning (Geneva & Zabaras, 2022). A hallmark emergent capability of transformers is *in-context learning* (ICL) (Brown et al., 2020a; Olsson et al., 2022), through which pretrained transformers (PTs) adapt to new tasks without parameter updates by leveraging information provided in the input context, for example in translation, question answering, and arithmetic or logical reasoning. In particular, prior works (Lin & Lee, 2024; Pan et al., 2023) have identified two distinct modes of ICL behavior: *task retrieval*, where the model identifies and invokes a relevant pretrained task, and *task learning*, where the model infers an underlying rule from an in-context prompt and applies it to new queries.

Recently, a growing body of work has explored transformers using Markov data models to study their ICL behavior and underlying mechanisms; see, e.g., (Edelman et al., 2024; Ekbote et al., 2025b; Lepage et al., 2025; Makkuva et al., 2025; Rajaraman et al., 2024b). In these models, each token is generated according to a transition rule that depends only on a finite number of preceding tokens, providing a simple yet expressive abstraction of autoregressive sequential prediction (Li et al., 2024; Yüksel & Flammarion, 2025). This Markovian perspective also provides a natural testbed for studying task retrieval and task learning in ICL. For example, Lepage et al. (Lepage et al., 2025) observed a transition from memorizing a training chain to inferring from the in-context prompt as the number of Markov chains used to generate pretraining sequences increases; see Figure 1. More specifically, when the number of Markov chains is small, they found that transformers tend to generate new tokens according to a memorized training chain, regardless of the in-context prompt. In contrast, when the number of Markov chains is large, transformers instead generate new tokens

*The first two authors contributed equally. Correspondence to Peng Wang (**Email:** pengw@um.edu.mo).

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).



(a) Number of states $S = 2$ *(b)* Number of states $S = 4$

Figure 1. **Transition from task retrieval to task learning in ICL.** We report the next-token prediction loss on new test sequences for the K-prior Bayes predictor (see Lemma 1), the uniform-prior Bayes predictor (see Lemma 2), and the pretrained transformer (PT). When the number of pretraining matrices is small, the PT performs similarly to the K-prior Bayes predictor, suggesting that it relies on memorized pretraining tasks. In contrast, when the number of pretraining matrices is large, its performance approaches that of the uniform-prior Bayes predictor, indicating that it infers the transition rule from the in-context prompt. Please refer to Section 5.1 for more details.

according to the transition rule inferred from the in-context prompt. Here, if each pretraining Markov chain is viewed as a task, then memorizing a training chain can be interpreted as task retrieval in ICL, whereas inferring from the in-context prompts corresponds to task learning. This observation is confirmed by recent work (Park et al., 2025) on competition dynamics and algorithmic phases in ICL on a mixture of Markov chains, which identifies a similar transition between retrieval and learning behaviors.

The above discussion naturally motivates the central question of this paper:

Can we characterize when and why PTs on Markov data exhibit task retrieval versus task learning in ICL?

Although recent works (Lepage et al., 2025; Park et al., 2025) provide empirical evidence and mechanistic interpretations of these phenomena, a precise theory remains lacking. We aim to develop a framework that characterizes when PTs exhibit each behavior and how the two regimes depend on task diversity and prompt length.

1.1. Main Contributions

We develop a theoretical framework explaining when PTs exhibit task retrieval versus task learning in ICL under a mixture of K Markov transition matrices. A central object is the finite-prior Bayes predictor induced by next-token pretraining: for small K, it performs posterior retrieval over pretrained Markov tasks, whereas for large K, it approaches prompt-based empirical transition estimation. Combining this Bayes characterization with architectural constraints, we explain how PTs can realize retrieval in low-diversity regimes and are biased toward task learning in high-diversity regimes. Our main contributions are summarized as follows:

- We explain task retrieval in the low-diversity regime. We derive the finite-prior Bayes predictor in closed form and show that it concentrates on the pretrained task most compatible with the prompt, with retrieval error decaying exponentially in the prompt length (see Theorem 1). Moreover, we show that this retrieval-based predictor is realizable by a two-layer single-head causal transformer whose hidden dimension scales linearly with K (see Theorem 2).
- We explain task learning in the high-diversity regime. As K grows, the finite-prior Bayes predictor approaches the uniform-prior Bayes predictor and hence the empirical Markov predictor inferred from the prompt, yielding the task-learning guarantee (see Theorem 3). Under architectural constraints, explicit retrieval over all K pretrained kernels becomes capacity-inefficient, biasing PTs trained to global optimality toward prompt-based transition estimation.
- We extend the framework beyond binary first-order chains to general multi-state, higher-order Markov data models, demonstrating that the same perspective applies to task retrieval and learning in broader sequential prediction settings.

1.2. Related Work

Studies of transformers on Markov data. Extensive works have studied transformers on Markov data models as tractable testbeds for understanding ICL. One line of work (Bietti et al., 2023; Chen et al., 2024; Ekbote et al., 2025a;

Nichani et al., 2024; Rajaraman et al., 2024a) shows that two-layer single-head transformers can learn Markov chains under suitable architectural constructions. They also provide a training-dynamics analysis, showing that gradient descent can find such solutions under appropriate conditions. Another line of work (Edelman et al., 2024; Makkuva et al., 2025; Varre et al., 2025) characterizes the loss landscape of transformers, showing how data distribution and architecture influence global optima, bad local minima, and transitions from simple unigram predictors to context-dependent bigram predictors. In addition, some works have studied other aspects of transformers on Markov data, including generalization bounds (Yüksel & Flammarion, 2025; Zekri et al., 2024), induction heads (Chen et al., 2024; Edelman et al., 2024), and Bayesian inference (Hao et al., 2025; Xie et al., 2022; Zhou et al., 2026).

Task retrieval and learning in ICL. (Min et al., 2022; Pan et al., 2023) demonstrate that the effectiveness of ICL depends on two distinct mechanisms through which transformers exploit in-context information, namely task retrieval and task learning. Later, (Lin & Lee, 2024) provides the first explanation of task retrieval and task learning in ICL through a linear regression framework. Recently, (Nafar et al., 2024) studies ICL in real-world regression settings and argues that its behavior ranges from retrieving pretrained task information to learning from in-context examples, depending on task familiarity and the richness of the provided examples. More recently, (Yang et al., 2026) provides a unified mechanistic framework for understanding how LLMs perform ICL by identifying two specialized types of attention heads, namely task-recognition heads and task-learning heads. Moreover, our work is closely related to (Lepage et al., 2025; Park et al., 2025), which study ICL in a finite mixture of Markov chains and demonstrate that transformers exhibit retrieval and learning behaviors, with transitions governed by data diversity and prompt length.

Notation. Given a matrix $\boldsymbol{A}$, we denote its spectral norm, i-th largest singular value, (i,j)-th entry, and Frobenius norm by $\|\boldsymbol{A}\|$, $\sigma_i(\boldsymbol{A})$, a_{ij}, and $\|\boldsymbol{A}\|_F$, respectively. Given a vector $\boldsymbol{a}$, we use $\|\boldsymbol{a}\|$ to denote its Euclidean norm and a_i to denote its i-th entry. We use $[n]$ to denote the set $\{1,\dots,n\}$, and $\boldsymbol{e}_i \in \mathbb{R}^d$ to denote the i-th standard basis vector in $\mathbb{R}^d$. We use $\mathrm{Unif}(0,1)$ to denote the uniform distribution on the interval $(0,1)$, and $\mathbb{I}\{A\}$ to denote the indicator function of an event A, which equals 1 if A occurs and 0 otherwise. Moreover, δ_z denotes the Dirac measure at z.

2. Problem Setup

In this section, we focus on the simplest Markov data model, namely binary first-order Markov chains. We discuss extensions of our analysis to general Markov chains in Section 4.

Markov chain model for token sequences. Let $\{0,1\}$ denote the binary alphabet, so that the number of states is $S=2$. We model the token sequence $s_{1:T} := (s_1,\dots,s_T)$ as a Markov chain, so that each next token depends only on the current token, i.e., $\mathbb{P}(s_{n+1}=j \mid s_{1:n}) = \mathbb{P}(s_{n+1}=j \mid s_n)$ for any $n \geq 1$ and $j \in \{0,1\}$. The transition dynamics are specified by a transition matrix $\boldsymbol{P} = (\boldsymbol{P}_{i,j})$, where $\boldsymbol{P}_{i,j} := \mathbb{P}(s_{n+1}=j \mid s_n = i)$. Let $s_{1:T} \in \{0,1\}^T$ be a sequence generated by a Markov chain with transition matrix

$$\boldsymbol{P}(p,q) = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}, \quad p,q \in (0,1). \tag{1}$$

We assume that the initial token s_1 is drawn from a fixed distribution $\boldsymbol{\mu} = (\mu_0,\mu_1)$, independent of (p,q), where $\mu_0,\mu_1 \in (0,1)$ and $\mu_0+\mu_1=1$.

Transformer architecture. We consider a causal transformer that maps a binary sequence $s_{1:T} \in \{0,1\}^T$ to layer-wise hidden representations $\{\boldsymbol{x}_n^{(\ell)}\}_{n=1}^T$ for $\ell = 0,\dots,L$. The final representation $\boldsymbol{x}_n^{(L)}$ is passed through a readout map to produce $f_{\boldsymbol{\theta}}(s_{1:n}) \in (0,1)$, interpreted as the predicted probability of the next token being 1. We defer the formal setup of this L-layer transformer to Section A.

Pretraining and ICL. A standard training objective is the next-token prediction loss, namely, the binary cross-entropy between the predicted probability $f_{\boldsymbol{\theta}}(s_{1:n})$ and the next token s_{n+1}, as follows:

$$\min_{\boldsymbol{\theta}} \ \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{T-\hat{T}} \mathbb{E}_{(p,q)\sim\Pi_{\mathrm{train}},\ s_{1:T}\sim\boldsymbol{P}(p,q)} \left[\sum_{n=\hat{T}}^{T-1} \ell\left(s_{n+1}, f_{\boldsymbol{\theta}}(s_{1:n})\right) \right], \tag{2}$$

where $\hat{T} \in [T-1]$ is a context warm-up length, Π_{train} is the training prior over the transition parameters (p,q), and $\ell(\cdot,\cdot)$ denotes the binary cross-entropy loss defined by

$$\ell(x,y) := -x\log(y) - (1-x)\log(1-y). \tag{3}$$

In our setting, each task is characterized by a latent Markov transition matrix. We therefore adopt the following Markov data model for generating pretraining sequences.

Definition 1 (Markov Data Generative Model)**.** *The pretraining task family consists of K latent Markov chains with transition matrices $\{\boldsymbol{P}(p^{(k)}, q^{(k)})\}_{k=1}^K$, where $\{(p^{(k)}, q^{(k)})\}_{k=1}^K \overset{i.i.d.}{\sim} \mathrm{Unif}(0,1)^2$. Conditioned on this task family, the training prior is $\Pi_{\mathrm{train}} := \sum_{k=1}^K \delta_{(p^{(k)},q^{(k)})}/K$.*

We pretrain the transformer by solving Problem (2). At inference time, given an in-context prompt $s_{1:n}^{\mathrm{ICL}} \in \{0,1\}^n$, the model generates tokens autoregressively: for each $t = n, n+1, \ldots$, it samples s_{t+1} from the predictive distribution induced by $f_{\boldsymbol{\theta}}(s_{1:t})$.

3. Main Results

Given a prefix $s_{1:n} \in \{0,1\}^n$ with $n \geq 2$, we define the transition counts

$$N_{ij}(s_{1:n}) := \sum_{t=1}^{n-1} \mathbb{I}\{s_t = i, s_{t+1} = j\}, \quad \forall i, j \in \{0,1\}. \tag{4}$$

This counts how many times the transition $i \to j$ appears in the prefix $s_{1:n}$. For simplicity, we write $N_{ij}(s_{1:n})$ as N_{ij} when the dependence on $s_{1:n}$ is clear from the context. Moreover, we define

$$N_0 := N_{00} + N_{01}, \quad N_1 := N_{10} + N_{11}, \quad \hat{c}_0 := \frac{N_0}{n-1}, \quad \hat{c}_1 := \frac{N_1}{n-1}, \quad \hat{p} := \frac{N_{01}}{N_0}, \quad \hat{q} := \frac{N_{10}}{N_1}. \tag{5}$$

Here, N_0 and N_1 respectively denote the total number of observed transitions out of state 0 and 1 along the prefix $s_{1:n}$. Accordingly, we define the empirical transition matrix

$$\hat{\boldsymbol{P}} := \boldsymbol{P}(\hat{p}, \hat{q}) = \begin{bmatrix} 1-\hat{p} & \hat{p} \\ \hat{q} & 1-\hat{q} \end{bmatrix}. \tag{6}$$

Given $\boldsymbol{P}(p,q)$ and $\ell(\cdot,\cdot)$ defined in (3), we define the weighted empirical cross-entropy as

$$D\big(\hat{\boldsymbol{P}}, \boldsymbol{P}\big) := \hat{c}_0\, \ell\big(\hat{p}, p\big) + \hat{c}_1\, \ell\big(\hat{q}, q\big). \tag{7}$$

3.1. Analysis of ICL Task Retrieval

Analysis of the Bayes predictor. Armed with the setup in Section 2, we first characterize the Bayes predictor of the pretraining next-token prediction loss (2) under the pretraining prior $\Pi_{\mathrm{train}} = \sum_{k=1}^K \delta_{(p^{(k)},q^{(k)})}/K$ as follows:

Lemma 1. *Consider the setup in Definition 1. For any prefix $s_{1:n} \in \{0,1\}^n$, define*

$$\alpha_k(s_{1:n}) := \frac{\mathbb{P}_k(s_{1:n})}{\sum_{r=1}^K \mathbb{P}_r(s_{1:n})}, \tag{8}$$

where $\mathbb{P}_k(s_{1:n})$ denotes the probability of observing $s_{1:n}$ under the k-th Markov chain with transition matrix $\boldsymbol{P}^{(k)} := \boldsymbol{P}(p^{(k)}, q^{(k)})$, as defined in Definition 1. The Bayes predictor for Problem (2) is

$$f_K^*(s_{1:n}) = \sum_{k=1}^K \alpha_k(s_{1:n}) \boldsymbol{P}^{(k)}_{s_n,1}, \tag{9}$$

where $\boldsymbol{P}^{(k)}_{s_n,1}$ denotes the transition probability from state s_n to 1 under the k-th Markov chain.

The proof is deferred to Section B.3. The predictor $f_K^*(\cdot)$ in (9) is the Bayes predictor, or equivalently, the optimal solution of the pretraining objective (2) over all measurable predictors. Based on this characterization, we analyze how the Bayes predictor (9) exhibits task-retrieval behavior in ICL.

Theorem 1. *Consider the setup in Definition 1. Let $s_{1:n}^{\mathrm{ICL}}$ be the observed in-context prompt with $N_0, N_1 > 0$ and*

$$k^* = \arg\min_{k\in[K]} D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}), \quad \Delta := \min_{k\neq k^*}\Big(D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}) - D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)})\Big), \tag{10}$$

where $\boldsymbol{P}^{(k)}$ for all $k \in [K]$ are defined in Definition 1 and $\hat{\boldsymbol{P}}$ is defined in (6). It holds that

$$\left| f_K^*(s_{1:n}^{\mathrm{ICL}}) - \boldsymbol{P}^{(k^*)}_{s_n^{\mathrm{ICL}},1} \right| \leq (K-1)\exp\left(-(n-1)\Delta\right). \tag{11}$$

The proof is deferred to Section C.1. We make the following remarks on this theorem.

- This theorem shows that if the empirical gap Δ is bounded away from zero, then the Bayes retrieval error decays exponentially with the prompt length n, up to the factor $K-1$. Hence, when K is moderate and the prompt is sufficiently long, the Bayes predictor exhibits *task-retrieval* behavior by selecting the pretrained task most compatible with the prompt and predicting according to its transition rule.
- In the above theorem, we assumed that the minimizer in (10) is unique. If the minimizer is not unique, the same argument shows that the Bayes predictor concentrates on the set of minimizing tasks rather than on a single task. This yields Corollary 1, which is deferred to Section C.2.

Analysis of pretrained transformers. Theorem 1 describes the Bayes predictor induced by the pretraining objective independently of any architectural considerations. We next show that the Bayes predictor can in fact be realized by the transformer architecture introduced in Section 2 with hidden dimension scaling linearly in the number of pretraining tasks. The proof is deferred to Section C.3.

Theorem 2. *The Bayes predictor $f_K^*(s_{1:n})$ in (9) can be represented by a two-layer single-head causal transformer defined in Section 2. In particular, there exists a choice of network parameters with hidden dimensions $d = 13 + K$, $d_h^{(0)} = 2$, $d_h^{(1)} = 13 + K$, and $d_{\mathrm{ff}} = 4$ such that in the hard-attention limit the resulting predictor recovers $f_K^*(s_{1:n})$.*

- $f_K^*(s_{1:n})$ globally minimizes the next-token prediction objective (2) over all measurable predictors, while Theorem 2 shows that this predictor can be represented by a small transformer architecture, with hidden dimensions scaling linearly in K. Thus, the Bayes predictor lies within the transformer hypothesis class considered here. Consequently, if the transformer is optimized sufficiently with respect to Problem (2), the trained model can converge to a predictor that realizes, or closely approximates, $f_K^*(s_{1:n})$, thereby exhibiting the task-retrieval behavior characterized in Theorem 1.
- The bound in (11) also extends to a trained predictor that only approximately realizes the Bayes predictor. Specifically, for any trained transformer $f_{\boldsymbol{\theta}}$, define the prompt-wise optimization error $\varepsilon_{\mathrm{opt}}(s_{1:n}^{\mathrm{ICL}}) := \left| f_{\boldsymbol{\theta}}(s_{1:n}^{\mathrm{ICL}}) - f_K^*(s_{1:n}^{\mathrm{ICL}}) \right|$. Then, by the triangle inequality and (11) in Theorem 1, we have

$$\left| f_{\boldsymbol{\theta}}(s_{1:n}^{\mathrm{ICL}}) - \boldsymbol{P}^{(k^*)}_{s_n^{\mathrm{ICL}},1} \right| \leq \varepsilon_{\mathrm{opt}}(s_{1:n}^{\mathrm{ICL}}) + (K-1)\exp\left(-(n-1)\Delta\right). \tag{12}$$

Takeaway: When pretraining task diversity is low, a well-trained transformer can behave like the Bayes predictor in (9), which exhibits task-retrieval behavior: its retrieval error scales linearly with K and decays exponentially with the prompt length n.

3.2. Analysis of ICL Task Learning

Analysis of the Bayes predictor. We now turn to analyze how the Bayes predictor (9) exhibits task-learning behavior in ICL, where the predictor infers the transition rule directly from the in-context prompt. Given a prefix $s_{1:n}$, we define the marginal likelihood of the prefix $s_{1:n}$ under the continuous uniform prior as

$$\rho_{\mathrm{unif}}(s_{1:n}) := \mathbb{E}_{(p,q)\sim\mathrm{Unif}(0,1)^2}\left[\mathbb{P}(s_{1:n} \mid p, q)\right]. \tag{13}$$

Theorem 3. *Consider the setup in Definition 1. Let $s_{1:n}^{\mathrm{ICL}}$ be the observed in-context prompt with $n \geq 2$ satisfying $\hat{c}_0 \in [\gamma, 1-\gamma]$ for some constant $\gamma \in (0, 1/2)$, where $\hat{c}_0$ is defined in (5). For any $\delta \in (0,1)$, if*

$$K \geq \frac{8(n+1)^2 2^{n-1}}{\min\{\mu_0, \mu_1\}} \log\left(\frac{4}{\delta}\right), \tag{14}$$

then, with probability at least $1-\delta$ *over the draw of the finite pretraining task family, it holds that*

$$\left| f_K^*(s_{1:n}^{\mathrm{ICL}}) - \hat{\boldsymbol{P}}_{s_n^{\mathrm{ICL}},1} \right| \le 8\sqrt{\frac{\log(4/\delta)}{K\rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}})}} + \frac{1}{\gamma(n-1)}. \tag{15}$$

The proof is deferred to Section D.1. We make the following remarks on this theorem.

- This theorem bounds the discrepancy between the Bayes predictor $f_K^*(s_{1:n}^{\mathrm{ICL}})$ and the empirical Markov predictor $\hat{\boldsymbol{P}}_{s_n^{\mathrm{ICL}},1}$ estimated from the in-context prompt. The bound has two terms. The first term $8\sqrt{\frac{\log(4/\delta)}{K\rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}})}}$ is the finite-prior approximation error: it measures how well the K sampled pretraining tasks approximate the continuous uniform-prior Bayes predictor. The second term $\frac{1}{\gamma(n-1)}$ is the smoothing bias between the uniform-prior Bayes predictor and the empirical Markov predictor. Thus, when K is sufficiently large and the prompt is sufficiently long, the finite-prior Bayes predictor approaches the empirical Markov predictor, exhibiting *task-learning* behavior.
- The high-probability statement is with respect to the random draw of the pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$. In particular, for any fixed prefix with $\rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}}) > 0$, taking $\delta = 1/K$ gives

$$\left| f_K^*(s_{1:n}^{\mathrm{ICL}}) - \hat{\boldsymbol{P}}_{s_n^{\mathrm{ICL}},1} \right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log K}{K\rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}})}} + \frac{1}{n} \right).$$

Analysis of pretrained transformers. In the large-diversity regime, it is important to account for the constraints imposed by the transformer hypothesis class. Although the finite-task Bayes predictor is globally optimal over all measurable predictors, explicitly realizing it through retrieval over K pretrained transition kernels can require model size growing with K (see Theorem 2). This motivates studying an alternative, capacity-efficient mechanism: estimating the transition rule directly from the in-context prompt, i.e., task learning.

Motivated by this perspective, we consider the following constrained optimization problem:

$$\boldsymbol{\theta}^* \in \arg\min_{f_{\boldsymbol{\theta}} \in \mathcal{F}} \mathcal{L}(\boldsymbol{\theta}), \tag{16}$$

where $\mathcal{L}(\boldsymbol{\theta})$ is defined in (2) and $\mathcal{F}$ denotes a transformer hypothesis class that contains the uniform-prior Bayes predictor f_{unif}^* (see Lemma 3) and the empirical Markov predictor $\hat{\boldsymbol{P}}_{s_n,1}$, where $\hat{\boldsymbol{P}}$ is defined in (6). Notably, Ekbote et al. (2025a, Theorem 1) shows that transformers with hidden dimension scaling as $O(T)$, where T denotes the sequence length, can realize these predictors. We next show that, under sufficient task diversity, a global minimizer over the constrained class $\mathcal{F}$ approximates the empirical Markov predictor over in-context prompts of length n.

Theorem 4. *Consider the setup in Theorem 3. Suppose that*

$$K \gtrsim \frac{T^2 2^T}{\min\{\mu_0, \mu_1\}} \log(2^T K). \tag{17}$$

Let $\mathcal{S}_n$ *be the set of non-degenerate prefixes of length* n*, and let* $s_{1:n}^{\mathrm{ICL}}$ *be sampled uniformly from* $\mathcal{S}_n$*. Suppose further that* $s_{1:n}^{\mathrm{ICL}}$ *satisfies* $\hat{c}_0 \in [\gamma, 1-\gamma]$ *for some constant* $\gamma \in (0, 1/2)$*, where* $\hat{c}_0$ *is computed from* $s_{1:n}^{\mathrm{ICL}}$ *as in* (5)*. It holds with probability at least* $1 - 1/\log K - 1/K$ *that*

$$\left| f_{\boldsymbol{\theta}^\star}(s_{1:n}^{\mathrm{ICL}}) - \hat{\boldsymbol{P}}_{s_n^{\mathrm{ICL}},1} \right| \lesssim \frac{\log K}{\sqrt{K}} + \frac{1}{n}, \tag{18}$$

where the hidden constant depends on the fixed sequence length T*, as well as on* γ *and* $\min\{\mu_0, \mu_1\}$*.*

The proof is deferred to Section D.2. We make the following remarks on this theorem.

- This theorem shows that, in the large-diversity regime, a global minimizer over the constrained class $\mathcal{F}$ can approximate the empirical Markov predictor (see (6)) estimated from the in-context prompt. Thus, the PT exhibits *task-learning* behavior: instead of explicitly retrieving one of the pretrained transition kernels, it predicts according to transition statistics inferred from the prompt. In particular, the bound (18) quantifies how task diversity and prompt length affect task learning: for fixed T, the approximation error decays as $\log K/\sqrt{K}$, while the smoothing bias scales as $1/n$.

- The exact global minimizer assumption can be relaxed. Specifically, suppose that a trained predictor $f_{\widehat{\boldsymbol{\theta}}} \in \mathcal{F}$ achieves the constrained optimum up to an optimization error ε_{opt}, i.e., $\mathcal{L}(f_{\widehat{\boldsymbol{\theta}}}) \le \inf_{f\in\mathcal{F}} \mathcal{L}(f) + \varepsilon_{\text{opt}}$. The same argument shows that the task-learning error decomposes into the statistical approximation and smoothing errors in (18), plus an additional optimization-error term:

$$\left| f_{\widehat{\boldsymbol{\theta}}}(s_{1:n}^{\text{ICL}}) - \hat{\boldsymbol{P}}_{s_n^{\text{ICL}},1} \right| \lesssim \frac{\log K}{\sqrt{K}} + \frac{1}{n} + \sqrt{\varepsilon_{\text{opt}}}. \tag{19}$$

Takeaway: When pretraining task diversity is large, explicit retrieval becomes capacity-inefficient. Under constrained optimization, the global optimum favors task learning: it predicts according to the empirical Markov rule inferred from the prompt rather than retrieving a pretrained task, with error scaling as $\log K/\sqrt{K} + 1/n$.

4. Extension to Multi-State Higher-Order Markov Chains

In this section, we extend the binary first-order Markov model in Section 2 to general finite-state and m-order Markov chains, where $m \ge 1$ denotes the Markov order.

4.1. Multi-State Higher-Order Markov Chains.

Let $[S] := \{1, \dots, S\}$ be a finite alphabet with S states. An m-th order Markov chain is a stochastic process $s_{1:T} \in [S]^T$ such that the next token depends on the past only through the most recent m tokens. Specifically, for any $n \ge m$ and $j \in [S]$, $\mathbb{P}(s_{n+1} = j \mid s_{1:n}) = \mathbb{P}(s_{n+1} = j \mid s_{n-m+1:n})$. The transition law is specified by a transition tensor

$$\boldsymbol{P} = \left(\boldsymbol{P}_{u,j}\right)_{u\in[S]^m,\ j\in[S]}, \quad \boldsymbol{P}_{u,j} := \mathbb{P}(s_{n+1} = j \mid s_{n-m+1:n} = u),$$

where $u \in [S]^m$ denotes a length-m context. For each context u, $\boldsymbol{P}_{u,\cdot}$ is a probability vector over the next state, i.e., $\boldsymbol{P}_{u,j} \ge 0$, $\sum_{j\in[S]} \boldsymbol{P}_{u,j} = 1$ for all $u \in [S]^m$. We assume that the initial block $s_{1:m}$ is drawn from a fixed distribution $\boldsymbol{\mu}$ over $[S]^m$, independent of the transition tensor.

Transition counts and empirical transition tensor. Given a prefix $s_{1:n} \in [S]^n$ with $n \ge m+1$, we define

$$N_{u,j}(s_{1:n}) := \sum_{t=m}^{n-1} \mathbb{I}\{s_{t-m+1:t} = u,\ s_{t+1} = j\},\ \forall u \in [S]^m,\ j \in [S].$$

For simplicity, we write $N_{u,j}$ when the dependence on $s_{1:n}$ is clear from the context. We define

$$N_u := \sum_{j\in[S]} N_{u,j}, \quad \hat{c}_u := \frac{N_u}{n-m}, \quad \hat{\boldsymbol{P}}_{u,j} := \frac{N_{u,j}}{N_u},\ \forall j \in [S]. \tag{20}$$

Here, $\hat{\boldsymbol{P}}_{u,\cdot}$ is the empirical next-token distribution after observing context u. Throughout this extension, we restrict attention to contexts with $N_u > 0$. Given a transition tensor $\boldsymbol{P}$, define the weighted empirical cross-entropy

$$D(\hat{\boldsymbol{P}}, \boldsymbol{P}) := \sum_{u:\, N_u>0} \hat{c}_u \sum_{j\in[S]} \left(-\hat{\boldsymbol{P}}_{u,j} \log \boldsymbol{P}_{u,j}\right). \tag{21}$$

Up to terms depending only on $\hat{\boldsymbol{P}}$, the quantity $D(\hat{\boldsymbol{P}}, \boldsymbol{P})$ is the negative normalized log-likelihood of the observed transitions in the prefix under transition tensor $\boldsymbol{P}$.

4.2. Analysis of Task Retrieval and Learning

As in the binary first-order case, we assume that pretraining sequences are generated from a finite family of latent Markov tasks.

Definition 2. *Suppose the pretraining task family consists of K latent m-th order Markov chains with transition tensors $\{\boldsymbol{P}^{(k)}\}_{k=1}^K$. For each $k \in [K]$ and each context $u \in [S]^m$, the transition probability vector $\boldsymbol{P}_{u,\cdot}^{(k)}$ is sampled independently from the uniform distribution over the probability simplex.*

By the same argument as in Lemma 1, the Bayes predictor for the next-token distribution is

$$\boldsymbol{f}_K^*(s_{1:n}) = \sum_{k=1}^{K} \alpha_k(s_{1:n})\, \boldsymbol{P}_{u_n,\cdot}^{(k)} \in \Delta^{S-1} := \left\{ \boldsymbol{p} \in \mathbb{R}_+^S : \sum_{j=1}^{S} p_j = 1 \right\}$$

where $u_n := s_{n-m+1:n}$ is the current length-m context and $\alpha_k(s_{1:n})$ is defined in (8). Here, unlike the binary case where $f_K^*(s_{1:n})$ is a scalar predicting the probability of the next token being 1, the predictor $\boldsymbol{f}_K^*(s_{1:n})$ is an S-dimensional probability vector. The j-th coordinate $f_{K,j}^*(s_{1:n})$ represents the Bayes predicted probability of the next token being j, i.e., $f_{K,j}^*(s_{1:n}) = \mathbb{P}(s_{n+1} = j \mid s_{1:n})$. With the above setup in place, we now extend our task-retrieval and task-learning characterizations of the Bayes predictor to multi-state higher-order Markov chains.

Theorem 5. *Consider the setup in Definition 2. Let $s_{1:n}^{\mathrm{ICL}}$ be the observed in-context prompt, where $n \geq m+1$. The following statements hold:*
(i) Define

$$k^* = \arg\min_{k \in [K]} D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}), \quad \Delta := \min_{k \neq k^*} \Big(D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}) - D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)}) \Big).$$

It holds for all $j \in [S]$ that

$$\left| f_{K,j}^*(s_{1:n}^{\mathrm{ICL}}) - \boldsymbol{P}_{u_n,j}^{(k^*)} \right| \leq (K-1)\exp\big(-(n-m)\Delta\big), \quad \textit{where } u_n = s_{n-m+1:n}^{\mathrm{ICL}}.$$

(ii) Suppose that the current context $u_n = s_{n-m+1:n}^{\mathrm{ICL}}$ satisfies $N_{u_n} \geq \gamma(n-m)$ for some $\gamma \in (0,1)$. Let $\rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}}) := \mathbb{E}_{\boldsymbol{P} \sim \Pi_{\mathrm{unif}}} \left[\mathbb{P}(s_{1:n}^{\mathrm{ICL}} \mid \boldsymbol{P}) \right]$. Then, for any $\delta \in (0,1)$, if $K \geq \frac{32 \log(2(S+1)/\delta)}{\rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}})}$, it holds with probability at least $1-\delta$ for every $j \in [S]$ that

$$\left| f_{K,j}^*(s_{1:n}^{\mathrm{ICL}}) - \hat{\boldsymbol{P}}_{u_n,j} \right| \leq 8\sqrt{\frac{\log(2(S+1)/\delta)}{K \rho_{\mathrm{unif}}(s_{1:n}^{\mathrm{ICL}})}} + \frac{S-1}{\gamma(n-m)+S}.$$

The proof is deferred to Section E. Although we state the extension for the Bayes predictor, the analysis of PTs can be extended in the same way as in the binary first-order case, covering both task retrieval and task learning.

5. Experimental Results

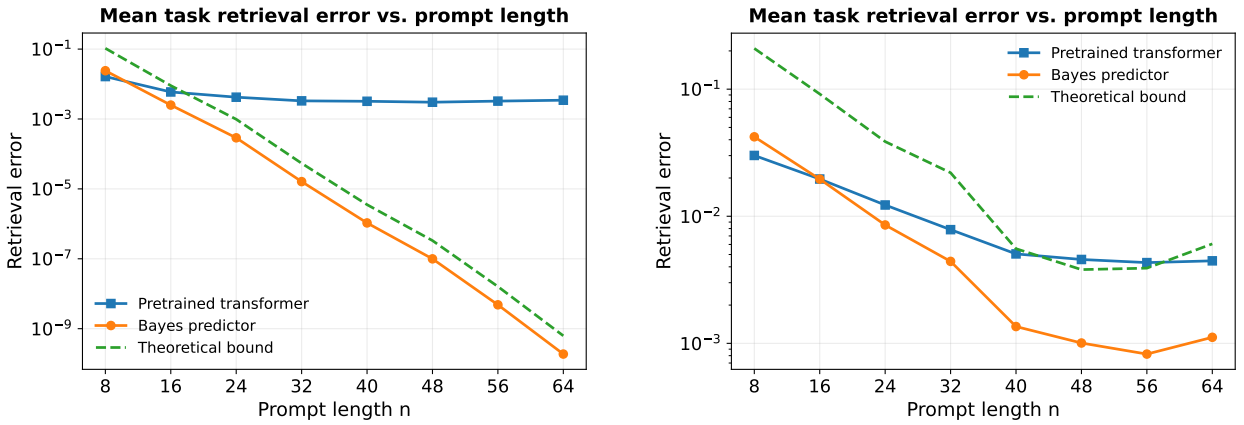
To empirically validate our theoretical findings on task retrieval and task learning in ICL, we conduct experiments on Markov data. Our experiments are organized around the two theoretical regimes studied in this paper: a low-diversity setting for task retrieval and a high-diversity setting for task learning.

Data Generation. We generate pretraining data following Definition 1. Specifically, the pretraining prior is a discrete uniform mixture over K Markov chains, whose transition parameters are sampled independently from $\mathrm{Unif}(0.01, 0.99)^2$ for numerical stability unless otherwise specified. The initial distribution is fixed as $\mu = (0.5, 0.5)$, and the sequence length is set to $T = 128$.

5.1. Transition from Task Retrieval to Task Learning

We train a 4-layer causal transformer with relative positional encoding, using $d_{\mathrm{model}} = 128$ and $d_{\mathrm{ff}} = 128$. The model is trained from scratch with AdamW, batch size 512, peak learning rate 10^{-3}, and a linear warmup over 10 epochs. To ensure sufficiently long contexts, the next-token prediction loss is computed only at positions $n \geq 20$. To study the transition from task retrieval to task learning, we vary the number of pretraining Markov chains over $K \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. For each K, we generate $N = 200K$ binary sequences, keeping the number of training sequences per task fixed.

After training, we evaluate on 50 unseen Markov tasks, for a total of 1,000 test sequences. We compute the ICL next-token prediction loss for prefixes with $n \geq 50$, and compare the PT with two theoretical baselines: the K-prior Bayes predictor in (9), which corresponds to task retrieval, and the uniform-prior Bayes predictor in (30), which corresponds to task learning. As shown in Figure 1(a), the PT closely follows the K-prior Bayes predictor in the low-diversity regime, indicating task-retrieval behavior. As K increases, its performance shifts toward the uniform-prior Bayes predictor, suggesting that the model increasingly infers the transition rule from the in-context prompt and exhibits task-learning behavior. Figure 1(b) shows an analogous transition on four-state Markov data.



(a) $(p, q) = (0.8, 0.2),\ \Delta \approx 0.45$ *(b)* $(p, q) = (0.5, 0.2),\ \Delta \approx 0.33$

Figure 2. **Verification of the theory for task retrieval.** We fix $K = 2$ with $(p, q) \in \{(0.3, 0.7), (0.6, 0.4)\}$ during pretraining, evaluate prompts generated from different (p, q), and plot the retrieval error defined by the LHS of (11) for the PT and the Bayes estimator in (9). We also plot the theoretical upper bound given by the right-hand side of (11).

5.2. Empirical Validation of the Theory

We next empirically validate the quantitative predictions of our theory. In this set of experiments, we train an 8-layer causal transformer with 8 attention heads, $d_{\text{model}} = 256$, and $d_{\text{ff}} = 512$. The model is trained for 400 epochs using AdamW with batch size 512, learning rate 10^{-3}, and weight decay 10^{-4}. The pretraining transition matrices are fixed, while the training sequences are sampled online from the corresponding mixture of Markov chains.

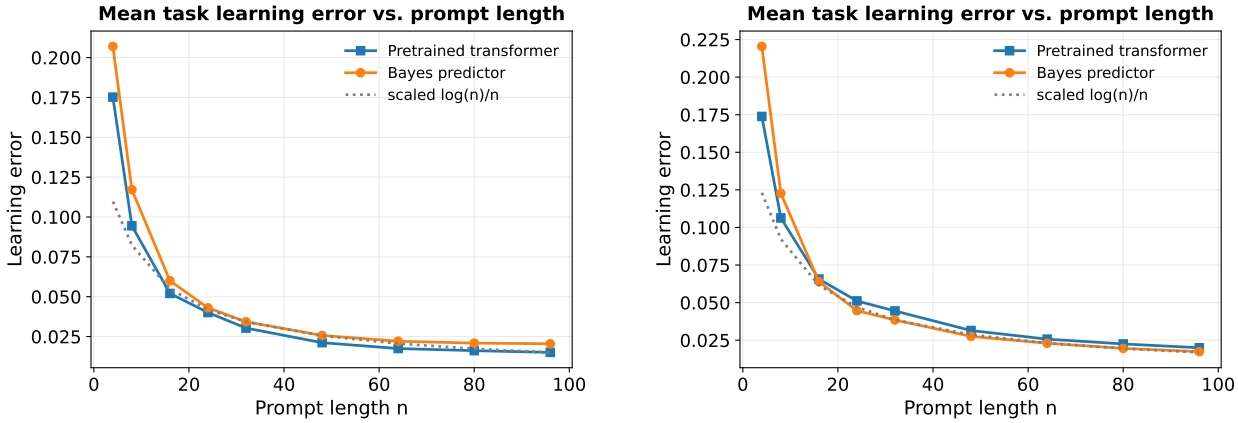
Task retrieval. To validate the task-retrieval prediction in Theorem 1, we fix $K = 2$ during pretraining, with transition parameters $(0.3, 0.7)$ and $(0.6, 0.4)$. We then evaluate on in-context prompts generated from a single test Markov chain and plot the retrieval error $|f^*_K(s^{\text{ICL}}_{1:n}) - \boldsymbol{P}^{(k^*)}_{s^{\text{ICL}}_n,1}|$ (see (11)), as a function of the prompt length n, for both the PT and the Bayes predictor in (9). We also plot the theoretical upper bound given by the right-hand side of (11).

We consider two evaluation settings with different likelihood gaps. In Case 1, the in-context prompts are generated from the Markov chain with transition parameters $(0.8, 0.2)$, yielding $\Delta \approx 0.45$, as shown in Figure 2(a). In Case 2, the prompts are generated from $(0.5, 0.2)$, yielding $\Delta \approx 0.33$, as shown in Figure 2(b). The results show that the retrieval error decays exponentially with the prompt length n, and that a larger likelihood gap Δ leads to faster decay. In particular, the Bayes predictor closely follows the theoretical upper bound in (11), consistent with Theorem 1. Notably, the gap between the PT and the Bayes predictor reflects the prompt-wise training error in (12).

Task learning. To evaluate task-learning behavior, we use the pretrained models with $K = 128$ and $K = 256$. For each K, we sample 50 unseen test Markov chains from $\text{Unif}(0.01, 0.99)^2$, and generate 20 in-context sequences from each test chain. For each prompt length n, we compute the empirical Markov predictor $\hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1}$ from the prefix $s^{\text{ICL}}_{1:n}$, and measure the learning error $|f(s^{\text{ICL}}_{1:n}) - \hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1}|$ (see (18)), where f is either the PT or the finite-prior Bayes predictor in (9). We then average this error over all test prompts and plot it as a function of n; see Figure 3. Both predictors exhibit task-learning behavior in this large-diversity regime, with the error decreasing as the prompt length and task diversity increase. Moreover, the decay with respect to n is consistent with the predicted finite-prompt rate, scaling on the order of $\log(n)/n$. This empirically supports Theorem 4.

6. Conclusions

We studied when PTs on Markov data exhibit task retrieval versus task learning in ICL. Under a finite-mixture Markov model, we showed that the Bayes predictor transitions from retrieving a compatible pretrained chain to estimating the transition rule from the prompt as task diversity and prompt length increase. We further connected this transition to transformer architectures and extended the framework to multi-state higher-order Markov chains. A key future direction is to analyze transformer training dynamics and prove that optimization algorithms can find the retrieval- and learning-based solutions characterized (Nichani et al., 2024; D'Angelo et al., 2025). Another promising direction is to extend our analysis to hidden Markov models (Dai et al., 2026; Hao et al., 2025), where the underlying Markov state is



(a) $K = 128$ *(b)* $K = 256$

Figure 3. **Verification of the theory for task learning.** We plot the task-learning errors, defined by the LHS of (18), for the PT and the Bayes predictor with $K = 128$ and $K = 256$. We also include a scaled $\log(n)/n$ reference curve. The errors decrease with prefix length and task diversity, and closely follow the predicted finite-prompt scaling, supporting Theorems 3 and 4.

only indirectly observed through emissions. This would allow us to study task retrieval and task learning under partial observability and latent sequential structure.

References

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A VIT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36:1560–1588, 2023.

Brown, T., Mann, B., Ryder, N., and others. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020a.

Brown, T., Mann, B., Ryder, N., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020b.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020.

Chen, S., Sheen, H., Wang, T., and Yang, Z. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66479–66567, 2024.

Dai, Y., Gao, Z., Sattar, Y., Dean, S., and Sun, J. J. Pre-trained large language models learn to predict hidden markov models in-context. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=btBqWTbf6q.

D'Angelo, F., Croce, F., and Flammarion, N. Selective induction heads: How transformers select causal structures in context. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bnJgzAQjWf.

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

Edelman, E., Tsilivis, N., Edelman, B., Malach, E., and Goel, S. The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems*, 37:64273–64311, 2024.

Ekbote, C., Bondaschi, M., Rajaraman, N., Lee, J. D., Gastpar, M., Makkuva, A. V., and Liang, P. P. What one cannot, two can: Two-layer transformers provably represent induction heads on any-order markov chains. *arXiv preprint arXiv:2508.07208*, 2025a.

Ekbote, C., Makkuva, A. V., Bondaschi, M., Rajaraman, N., Gastpar, M., Lee, J. D., and Liang, P. P. What one cannot, two can: Two-layer transformers provably represent induction heads on any-order markov chains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=nYg6Qzm5xS.

Geneva, N. and Zabaras, N. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, 2022.

Hao, Y., Ye, C., Han, C., and Zhang, T. Transformers as multi-task learners: Decoupling features in hidden markov models. *arXiv preprint arXiv:2506.01919*, 2025.

Lepage, S., Mary, J., and Picard, D. Markov chain estimation with in-context learning. *arXiv preprint arXiv:2508.03934*, 2025.

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S., and Oymak, S. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024.

Lin, Z. and Lee, K. Dual operating modes of in-context learning. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ElVHUWyL3n.

Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H., and Gastpar, M. Attention with markov: A curious case of single-layer transformers. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SqZ0KY4qBD.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Nafar, A., Venable, K. B., and Kordjamshidi, P. Learning vs retrieval: The role of in-context examples in regression with LLMs. *arXiv preprint arXiv:2409.04318*, 2024.

Nichani, E., Damian, A., and Lee, J. D. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, pp. 38018–38070, 2024.

Olsson, C., Elhage, N., Nanda, N., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Pan, J., Gao, T., Chen, H., and Chen, D. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, 2023.

Park, C. F., Lubana, E. S., and Tanaka, H. Competition dynamics shape algorithmic phases of in-context learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XgH1wfHSX8.

Rajaraman, N., Bondaschi, M., Makkuva, A. V., Ramchandran, K., and Gastpar, M. Transformers on markov data: Constant depth suffices. *Advances in Neural Information Processing Systems*, 37:137521–137556, 2024a.

Rajaraman, N., Jiao, J., and Ramchandran, K. An analysis of tokenization: Transformers under markov data. *Advances in Neural Information Processing Systems*, 37:62503–62556, 2024b.

Varre, A., Yüce, G., and Flammarion, N. Learning in-context n-grams with transformers: Sub-n-grams are near-stationary points. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=OMwdvGDeHL.

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

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Yang, H., Cho, H., and Inoue, N. Localizing task recognition and task learning in in-context learning via attention head analysis. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gdvOF1OMa7.

Yüksel, O. K. and Flammarion, N. On the sample complexity of next-token prediction. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 694–702. PMLR, 2025.

Zekri, O., Odonnat, A., Benechehab, A., Bleistein, L., Boullé, N., and Redko, I. Large language models as markov chains. *arXiv preprint arXiv:2410.02724*, 2024.

Zhou, R., Tian, C., and Diggavi, S. An information-theoretic approach to understanding transformers' in-context learning of variable-order markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=N6DOObrBqq.

Appendices

A. Transformer Architecture

We consider a causal transformer that maps a binary sequence $s_{1:T} \in \{0,1\}^T$ to layer-wise hidden representations $\{\boldsymbol{x}_n^{(\ell)}\}_{n=1}^T$ for $\ell = 0, \ldots, L$. The final representation $\boldsymbol{x}_n^{(L)}$ is used to predict the next token s_{n+1}. The L-layer transformer is defined as follows:

- *Embedding layer:* For each $n \in [T]$,

$$\boldsymbol{x}_n^{(0)} = \boldsymbol{\Phi} \boldsymbol{e}_{s_n} \in \mathbb{R}^d, \tag{22}$$

  where $\boldsymbol{\Phi} \in \mathbb{R}^{d \times S}$ is the token embedding matrix.

- *Causal self-attention and feedforward layers:* For each layer $\ell = 0, \ldots, L-1$ and position $n \in [T]$, we first compute the attention update

$$\widetilde{\boldsymbol{x}}_n^{(\ell+1)} = \boldsymbol{x}_n^{(\ell)} + \boldsymbol{W}_O^{(\ell)} \sum_{i=1}^{n} \mathrm{att}_{n,i}^{(\ell)} \boldsymbol{W}_V^{(\ell)} \big(\boldsymbol{x}_i^{(\ell)} + \boldsymbol{p}_{n-i}^{(\ell),V} \big) \in \mathbb{R}^d, \tag{23}$$

  where the attention weights are given by

$$\mathrm{att}_{n,i}^{(\ell)} = \frac{\exp\Big(\Big\langle \boldsymbol{W}_K^{(\ell)} \big(\boldsymbol{x}_i^{(\ell)} + \boldsymbol{p}_{n-i}^{(\ell),K} \big), \boldsymbol{W}_Q^{(\ell)} \boldsymbol{x}_n^{(\ell)} \Big\rangle \Big)}{\sum_{j=1}^{n} \exp\Big(\Big\langle \boldsymbol{W}_K^{(\ell)} \big(\boldsymbol{x}_j^{(\ell)} + \boldsymbol{p}_{n-j}^{(\ell),K} \big), \boldsymbol{W}_Q^{(\ell)} \boldsymbol{x}_n^{(\ell)} \Big\rangle \Big)}, \ \forall i \in [n]. \tag{24}$$

  Here, $\boldsymbol{W}_Q^{(\ell)}, \boldsymbol{W}_K^{(\ell)}, \boldsymbol{W}_V^{(\ell)} \in \mathbb{R}^{d_h^{(\ell)} \times d}$ are the query, key, and value matrices in layer ℓ, $\boldsymbol{W}_O^{(\ell)} \in \mathbb{R}^{d \times d_h^{(\ell)}}$ is the output projection matrix, and $\boldsymbol{p}_{n-i}^{(\ell),K}, \boldsymbol{p}_{n-i}^{(\ell),V} \in \mathbb{R}^{d_h^{(\ell)}}$ are the relative positional encodings for the key and value representations, respectively. We then apply a position-wise feedforward update:

$$\boldsymbol{x}_n^{(\ell+1)} = \widetilde{\boldsymbol{x}}_n^{(\ell+1)} + \boldsymbol{W}_2^{(\ell)} \, \mathrm{ReLU}\big(\boldsymbol{W}_1^{(\ell)} \widetilde{\boldsymbol{x}}_n^{(\ell+1)} \big), \tag{25}$$

  where $\boldsymbol{W}_1^{(\ell)} \in \mathbb{R}^{d_{\mathrm{ff}} \times d}$ and $\boldsymbol{W}_2^{(\ell)} \in \mathbb{R}^{d \times d_{\mathrm{ff}}}$ are learnable weight matrices, and d_{ff} is the hidden dimension of the feedforward network.

- *Output layer:* The final representation at position n is mapped to a scalar prediction through a readout function

$$f_{\boldsymbol{\theta}}(s_{1:n}) := \mathbb{P}\left(s_{n+1} = 1 \mid s_{1:n} \right) = \Psi_{\boldsymbol{\theta}}\big(\boldsymbol{x}_n^{(L)} \big) \in [0,1], \tag{26}$$

  where $\Psi_{\boldsymbol{\theta}} : \mathbb{R}^d \to [0,1]$ is a parameterized readout map.

B. Characterization of the Bayes Predictor

Throughout this section, we consider the binary first-order Markov model and assume that

$$(p,q) \sim \Pi, \qquad s_{1:T} \mid (p,q) \sim (\boldsymbol{\mu}, \boldsymbol{P}(p,q)),$$

where Π is a prior distribution on $(p,q) \in (0,1)^2$, $\boldsymbol{\mu}$ is an initial distribution independent of (p,q), and

$$\boldsymbol{P}(p,q) := \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

B.1. Bayes Predictor with a General Prior

We first establish the following characterization of the Bayes predictor under a general prior distribution Π.

Lemma 2. *Let Π be a prior distribution on $(p,q) \in (0,1)^2$. For a given prefix $s_{1:n} \in \{0,1\}^n$, consider the one-step log-loss*

$$\ell\big(s_{n+1}, f(s_{1:n})\big) = -s_{n+1} \log f(s_{1:n}) - (1 - s_{n+1}) \log\big(1 - f(s_{1:n})\big),$$

where $f(s_{1:n}) \in (0,1)$.
(i) The Bayes estimator of minimizing the posterior expected log-loss, i.e.,

$$f^*(s_{1:n}) \in \arg\min_{x \in (0,1)} \mathbb{E}\left[\ell\big(s_{n+1}, x\big) \mid s_{1:n}\right], \tag{27}$$

is

$$f^*(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \begin{cases} \mathbb{E}[p \mid s_{1:n}], & s_n = 0, \\ \mathbb{E}[1 - q \mid s_{1:n}], & s_n = 1. \end{cases} \tag{28}$$

(ii) It holds that

$$f^*(s_{1:n}) = \frac{\displaystyle\int \boldsymbol{P}_{s_n,1}(p,q)\, \mathbb{P}(s_{1:n} \mid p,q)\, \Pi(dp\, dq)}{\displaystyle\int \mathbb{P}(s_{1:n} \mid p,q)\, \Pi(dp\, dq)}, \tag{29}$$

where $\mathbb{P}(s_{1:n} \mid p,q)$ denotes the likelihood of the observed prefix $s_{1:n}$ under the Markov chain with transition parameters (p,q) and $\Pi(dp\, dq)$ denotes integration with respect to the prior distribution of (p,q).

Proof. (i) Note that $s_{1:n} \in \{0,1\}^n$ is a prefix. For any predictor $f(s_{1:n}) \in (0,1)$, it follows from the definition of the log-loss and $s_{n+1} \in \{0,1\}$ that

$$\begin{aligned} \mathbb{E}\big[\ell\big(s_{n+1}, f(s_{1:n})\big) \,\big|\, s_{1:n}\big] &= \mathbb{E}\big[-s_{n+1} \log f(s_{1:n}) - (1 - s_{n+1}) \log\big(1 - f(s_{1:n})\big) \,\big|\, s_{1:n}\big] \\ &= -\mathbb{E}[s_{n+1} \mid s_{1:n}] \log f(s_{1:n}) - \big(1 - \mathbb{E}[s_{n+1} \mid s_{1:n}]\big) \log\big(1 - f(s_{1:n})\big). \end{aligned}$$

Define $\beta(s_{1:n}) := \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \mathbb{E}[s_{n+1} \mid s_{1:n}]$. This implies

$$\mathbb{E}\big[\ell\big(s_{n+1}, f(s_{1:n})\big) \,\big|\, s_{1:n}\big] = -\beta(s_{1:n}) \log f(s_{1:n}) - \big(1 - \beta(s_{1:n})\big) \log\big(1 - f(s_{1:n})\big).$$

Now regard the right-hand side as a function of the scalar variable $u := f(s_{1:n}) \in (0,1)$. Define

$$\varphi(u) := -\beta \log u - (1 - \beta) \log(1 - u),$$

where $u \in (0,1)$ and $\beta := \beta(s_{1:n}) \in [0,1]$ is fixed. A direct differentiation gives

$$\varphi'(u) = -\frac{\beta}{u} + \frac{1 - \beta}{1 - u} = \frac{u - \beta}{u(1 - u)}.$$

Hence, the unique critical point is $u = \beta$. Moreover,

$$\varphi''(u) = \frac{\beta}{u^2} + \frac{1 - \beta}{(1 - u)^2} > 0, \quad \forall u \in (0,1),$$

so φ is strictly convex on $(0,1)$. Therefore, the unique minimizer of Problem (27) is

$$f^*(s_{1:n}) = \beta(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}).$$

Taking conditional expectation with respect to the posterior distribution of (p,q) given $s_{1:n}$, we obtain

$$\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \mathbb{E}[\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, p, q) \,|\, s_{1:n}] = \mathbb{E}[\boldsymbol{P}_{s_n,1}(p,q) \mid s_{1:n}],$$

where the second equality use Markov property. This, together with (1), implies

$$f^*(s_{1:n}) = \begin{cases} \mathbb{E}[p \mid s_{1:n}], & s_n = 0, \\ \mathbb{E}[1-q \mid s_{1:n}], & s_n = 1. \end{cases}$$

(ii) Conditional on the parameter pair (p, q), the chain is first-order Markov. Therefore,

$$\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, p, q) = \mathbb{P}(s_{n+1} = 1 \mid s_n, p, q) = \boldsymbol{P}_{s_n,1}(p, q).$$

Hence, by the tower property,

$$f^*(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \mathbb{E}[\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, p, q) \mid s_{1:n}] = \mathbb{E}[\boldsymbol{P}_{s_n,1}(p, q) \mid s_{1:n}]\,.$$

Applying Bayes' rule, the posterior distribution of (p, q) given $s_{1:n}$ is

$$\Pi(dp\,dq \mid s_{1:n}) = \frac{\mathbb{P}(s_{1:n} \mid p, q)\,\Pi(dp\,dq)}{\displaystyle\int \mathbb{P}(s_{1:n} \mid p, q)\,\Pi(dp\,dq)}.$$

Substituting this posterior representation into the conditional expectation above yields

$$f^*(s_{1:n}) = \frac{\displaystyle\int \boldsymbol{P}_{s_n,1}(p, q)\,\mathbb{P}(s_{1:n} \mid p, q)\,\Pi(dp\,dq)}{\displaystyle\int \mathbb{P}(s_{1:n} \mid p, q)\,\Pi(dp\,dq)}.$$

□

B.2. Bayes Predictor under the Uniform Prior

In this subsection, we consider the special case in which the prior distribution is uniform on $(0, 1)^2$, i.e., $\Pi = \mathrm{Unif}(0, 1)^2$.

Lemma 3. *Suppose that the prior distribution* $\Pi = \mathrm{Unif}(0, 1)^2$. *For any prefix* $s_{1:n} \in \{0, 1\}^n$*, it holds that the Bayes estimator*

$$f^*_{\mathrm{unif}}(s_{1:n}) = \begin{cases} \dfrac{N_{01}+1}{N_{00}+N_{01}+2}, & \text{if } s_n = 0, \\ \dfrac{N_{11}+1}{N_{10}+N_{11}+2}, & \text{if } s_n = 1. \end{cases} \tag{30}$$

where $N_{ij} := N_{ij}(s_{1:n})$ *is defined in* (4).

Proof. Since the initial distribution $\boldsymbol{\pi}$ is fixed and does not depend on (p, q), the likelihood of the observed prefix $s_{1:n}$ is, up to a multiplicative constant independent of (p, q),

$$\mathbb{P}(s_{1:n} \mid p, q) \propto (1-p)^{N_{00}} p^{N_{01}} q^{N_{10}} (1-q)^{N_{11}}.$$

Because the prior is $\Pi = \mathrm{Unif}(0, 1)^2$, its density with respect to Lebesgue measure on $(0, 1)^2$ is constant. Hence, by Bayes' rule, the posterior density of (p, q) given $s_{1:n}$ satisfies

$$\pi(p, q \mid s_{1:n}) \propto (1-p)^{N_{00}} p^{N_{01}} q^{N_{10}} (1-q)^{N_{11}},\ \forall (p, q) \in (0, 1)^2.$$

This expression factorizes as $\pi(p, q \mid s_{1:n}) \propto \left[p^{N_{01}}(1-p)^{N_{00}}\right]\left[q^{N_{10}}(1-q)^{N_{11}}\right]$. Therefore, conditioned on $s_{1:n}$, the random variables p and q remain independent, with

$$p \mid s_{1:n} \sim \mathrm{Beta}(N_{01}+1,\ N_{00}+1), \quad q \mid s_{1:n} \sim \mathrm{Beta}(N_{10}+1,\ N_{11}+1). \tag{31}$$

According to (28) in Lemma 2, we have

$$f^*_{\text{unif}}(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \begin{cases} \mathbb{E}[p \mid s_{1:n}], & s_n = 0, \\ \mathbb{E}[1-q \mid s_{1:n}], & s_n = 1. \end{cases}$$

Using the mean of a Beta distribution $\mathbb{E}[\text{Beta}(a,b)] = a/(a+b)$ and (31), we obtain

$$\mathbb{E}[p \mid s_{1:n}] = \frac{N_{01}+1}{N_{01}+N_{00}+2}, \quad \mathbb{E}[q \mid s_{1:n}] = \frac{N_{10}+1}{N_{10}+N_{11}+2}.$$

Hence, we have

$$\mathbb{E}[1-q \mid s_{1:n}] = 1 - \mathbb{E}[q \mid s_{1:n}] = \frac{N_{11}+1}{N_{10}+N_{11}+2}.$$

Combining the two cases yields (30). □

B.3. Proof of Lemma 1

Proof. By (28) in Lemma 2, the Bayes predictor satisfies

$$f^*_K(s_{1:n}) = \begin{cases} \mathbb{E}[p \mid s_{1:n}], & s_n = 0, \\ \mathbb{E}[1-q \mid s_{1:n}], & s_n = 1. \end{cases} \tag{32}$$

Since the prior distribution of (p,q) is $\Pi_{\text{train}} := \sum_{k=1}^K \delta_{(p^{(k)},q^{(k)})}/K$, Bayes' rule gives

$$\mathbb{P}\big((p,q) = (p^{(k)},q^{(k)}) \mid s_{1:n}\big) = \frac{\mathbb{P}_k(s_{1:n})\,\mathbb{P}\big((p,q) = (p^{(k)},q^{(k)})\big)}{\sum_{r=1}^K \mathbb{P}_r(s_{1:n})\,\mathbb{P}\big((p,q) = (p^{(r)},q^{(r)})\big)}.$$

Using $\mathbb{P}\big((p,q) = (p^{(k)},q^{(k)})\big) = 1/K$ for each $k \in [K]$, we obtain

$$\mathbb{P}\big((p,q) = (p^{(k)},q^{(k)}) \mid s_{1:n}\big) = \frac{\mathbb{P}_k(s_{1:n})}{\sum_{r=1}^K \mathbb{P}_r(s_{1:n})} = \alpha_k(s_{1:n}).$$

Therefore,

$$\mathbb{E}[p \mid s_{1:n}] = \sum_{k=1}^K \alpha_k(s_{1:n})\,p^{(k)}, \quad \mathbb{E}[1-q \mid s_{1:n}] = \sum_{k=1}^K \alpha_k(s_{1:n})\,(1-q^{(k)}).$$

This, together with (32) and

$$\boldsymbol{P}^{(k)}_{s_n,1} = \begin{cases} p^{(k)}, & s_n = 0, \\ 1-q^{(k)}, & s_n = 1, \end{cases}$$

yields

$$f^*_K(s_{1:n}) = \sum_{k=1}^K \alpha_k(s_{1:n})\,\boldsymbol{P}^{(k)}_{s_n,1}.$$

□

Remark 1. *For any prefix $s_{1:n} \in \{0,1\}^n$, because the initial distribution $\boldsymbol{\mu} = (\mu_0,\mu_1)$ is fixed and independent of k, the likelihood of observing $s_{1:n}$ under the k-th Markov chain is*

$$\mathbb{P}_k(s_{1:n}) = \mu_{s_1}(1-p^{(k)})^{N_{00}}(p^{(k)})^{N_{01}}(q^{(k)})^{N_{10}}(1-q^{(k)})^{N_{11}},$$

where $N_{ij} := N_{ij}(s_{1:n})$ is defined in (4). Therefore, $\alpha_k(s_{1:n})$ in (8) admits the explicit form

$$\alpha_k(s_{1:n}) = \frac{(1-p^{(k)})^{N_{00}}(p^{(k)})^{N_{01}}(q^{(k)})^{N_{10}}(1-q^{(k)})^{N_{11}}}{\sum_{r=1}^K (1-p^{(r)})^{N_{00}}(p^{(r)})^{N_{01}}(q^{(r)})^{N_{10}}(1-q^{(r)})^{N_{11}}}. \tag{33}$$

C. Proofs in Section 3.1

C.1. Proof of Theorem 1

Proof of Theorem 1. To simplify notation, we write $s_{1:n}^{\mathrm{ICL}}$ as $s_{1:n}$. According to Lemma 1, we have

$$f_K^*(s_{1:n}) = \sum_{k=1}^{K} \alpha_k(s_{1:n}) \boldsymbol{P}_{s_n,1}^{(k)}, \tag{34}$$

where $\alpha_k(s_{1:n})$ takes the form of (33). Note that

$$\begin{aligned}
&\log(1-p^{(k)})^{N_{00}}(p^{(k)})^{N_{01}}(q^{(k)})^{N_{10}}(1-q^{(k)})^{N_{11}} \\
&= N_{01}\log p^{(k)} + N_{00}\log\left(1-p^{(k)}\right) + N_{10}\log q^{(k)} + N_{11}\log\left(1-q^{(k)}\right) \\
&= N_0\left(\hat{p}\log p^{(k)} + (1-\hat{p})\log\left(1-p^{(k)}\right)\right) + N_1\left(\hat{q}\log q^{(k)} + (1-\hat{q})\log\left(1-q^{(k)}\right)\right) \\
&= -N_0\ell(\hat{p}, p^{(k)}) - N_1\ell(\hat{q}, q^{(k)}) = -(n-1)D\big(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}\big),
\end{aligned}$$

where the second equality follows from (5), the third equality uses (3), and the last equality is due to (7). This implies

$$(1-p^{(k)})^{N_{00}}(p^{(k)})^{N_{01}}(q^{(k)})^{N_{10}}(1-q^{(k)})^{N_{11}} = \exp\left(-(n-1)D\left(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}\right)\right). \tag{35}$$

For simplicity, let $w_k := \exp\left(-(n-1)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\right)$ for each $k \in [K]$. Using (33), (34), and (35), we have

$$f_K^*(s_{1:n}) = \frac{\sum_{k=1}^{K} w_k \boldsymbol{P}_{s_n,1}^{(k)}}{\sum_{k=1}^{K} w_k}. \tag{36}$$

Using the definitions of k^* and Δ in (10), we have for any $k \neq k^*$,

$$\begin{aligned}
w_k &= \exp\left(-(n-1)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\right) \\
&= \exp\left(-(n-1)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)})\right)\exp\left(-(n-1)\left(D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}) - D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)})\right)\right) \\
&\leq \exp\left(-(n-1)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)})\right)\exp\left(-(n-1)\Delta\right) = w_{k^*}\exp\left(-(n-1)\Delta\right).
\end{aligned} \tag{37}$$

Using (36), we have

$$f_K^*(s_{1:n}) = \frac{w_{k^*}\boldsymbol{P}_{s_n,1}^{(k^*)} + \sum_{k\neq k^*} w_k \boldsymbol{P}_{s_n,1}^{(k)}}{w_{k^*} + \sum_{k\neq k^*} w_k}. \tag{38}$$

This, together with $\boldsymbol{P}_{s_n,1}^{(k^*)} \leq 1$ for all $k \in [K]$, implies

$$\begin{aligned}
f_K^*(s_{1:n}) &\leq \frac{w_{k^*}\boldsymbol{P}_{s_n,1}^{(k^*)} + \sum_{k\neq k^*} w_k}{w_{k^*}} = \boldsymbol{P}_{s_n,1}^{(k^*)} + \frac{\sum_{k\neq k^*} w_k}{w_{k^*}} \\
&\leq \boldsymbol{P}_{s_n,1}^{(k^*)} + (K-1)\exp\left(-(n-1)\Delta\right),
\end{aligned} \tag{39}$$

where the last inequality follows from (37). Using (38) again, we have

$$f_K^*(s_{1:n}) \geq \frac{w_{k^*}\boldsymbol{P}_{s_n,1}^{(k^*)}}{w_{k^*} + \sum_{k\neq k^*} w_k} = \frac{\boldsymbol{P}_{s_n,1}^{(k^*)}}{1 + \sum_{k\neq k^*} w_k/w_{k^*}} \geq \frac{\boldsymbol{P}_{s_n,1}^{(k^*)}}{1 + (K-1)\exp\left(-(n-1)\Delta\right)},$$

where the last inequality uses (37). This, together with (39), yields

$$\frac{\boldsymbol{P}_{s_n,1}^{(k^*)}}{1 + (K-1)\exp\left(-(n-1)\Delta\right)} \leq f_K^*(s_{1:n}) \leq \boldsymbol{P}_{s_n,1}^{(k^*)} + (K-1)\exp\left(-(n-1)\Delta\right),$$

which implies (11).

□

C.2. Proof of Corollary 1

Corollary 1. *Consider the setup in Theorem 1. Let*

$$\Omega = \arg\min_{k\in[K]} D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}), \quad \Delta := \min_{k\notin\Omega, k^*\in\Omega} \Big(D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}) - D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)}) \Big). \tag{40}$$

Then, it holds that

$$\left| f_K^*(s_{1:n}^{\mathrm{ICL}}) - \frac{1}{|\Omega|} \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n^{\mathrm{ICL}},1} \right| \le (K - |\Omega|) \exp\left(-(n-1)\Delta\right). \tag{41}$$

Proof of Corollary 1. To simplify notation, write $s_{1:n}^{\mathrm{ICL}}$ as $s_{1:n}$. By the same likelihood calculation as in the proof of Theorem 1, the Bayes predictor can be written as

$$f_K^*(s_{1:n}) = \frac{\sum_{k=1}^K w_k \boldsymbol{P}^{(k)}_{s_n,1}}{\sum_{k=1}^K w_k}, \quad w_k := \exp\left(-(n-1) D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\right). \tag{42}$$

For all $k^* \in \Omega$, the weights w_{k^*} are equal and denote their common value by w_*, i.e., $w_* = w_{k^*}$ for each $k^* \in \Omega$. Then (42) gives

$$f_K^*(s_{1:n}) - \frac{1}{|\Omega|} \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n,1}. = \frac{\sum_{k\notin\Omega} w_k \left(\boldsymbol{P}^{(k)}_{s_n,1} - \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n,1}/|\Omega|. \right)}{|\Omega| w_* + \sum_{k\notin\Omega} w_k}.$$

Since $\boldsymbol{P}^{(k)}_{s_n,1}, \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n,1}/|\Omega| \in [0,1]$, we have

$$\left| \boldsymbol{P}^{(k)}_{s_n,1} - \frac{1}{|\Omega|} \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n,1} \right| \le 1.$$

Therefore,

$$\left| f_K^*(s_{1:n}) - \frac{1}{|\Omega|} \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n,1} \right| \le \frac{\sum_{k\notin\Omega} w_k}{|\Omega| w_* + \sum_{k\notin\Omega} w_k} \le \sum_{k\notin\Omega} \frac{w_k}{w_*}.$$

By the definition of Δ in (40), we have $w_k/w_* \le \exp\left(-(n-1)\Delta\right)$ for each $k \notin \Omega$. Therefore, we have

$$\left| f_K^*(s_{1:n}) - \frac{1}{|\Omega|} \sum_{k^*\in\Omega} \boldsymbol{P}^{(k^*)}_{s_n,1} \right| \le (K - |\Omega|) \exp\left(-(n-1)\Delta\right).$$

□

C.3. Proof of Theorem 2

Throughout this section, we use $\boldsymbol{x}_t^{(\ell)}[j]$ to denote the j-th coordinate of the vector $\boldsymbol{x}_t^{(\ell)} \in \mathbb{R}^d$.

Proof of Theorem 2. We augment the input sequence with a fixed start-of-sequence symbol $\mathrm{BOS} \notin \{0,1\}$, and consider the sequence

$$\bar{s}_{1:n+1} := (\mathrm{BOS}, s_1, \ldots, s_n).$$

We construct a two-layer causal masked transformer, as defined in Section 2, whose output at position $n+1$ equals the Bayes predictor $f_K^*(s_{1:n})$ in (9). To do so, we use hidden dimension

$$d = 13 + K.$$

For each layer representation $\boldsymbol{x}_t^{(\ell)}$ with $\ell \in \{0,1,2\}$, the hidden coordinates are interpreted as follows:

$$\begin{aligned}
\boldsymbol{x}_t^{(\ell)}[1] &: \text{ current binary token value,}\\
\boldsymbol{x}_t^{(\ell)}[2] &: \text{ constant } 1,\\
\boldsymbol{x}_t^{(\ell)}[3] &: \text{ predecessor binary token value,}\\
\boldsymbol{x}_t^{(\ell)}[4] &: \text{ current-token validity flag,}\\
\boldsymbol{x}_t^{(\ell)}[5] &: \text{ predecessor-token validity flag,}\\
\boldsymbol{x}_t^{(\ell)}[6:9] &: \text{ local transition indicators } (z_t^{00}, z_t^{01}, z_t^{10}, z_t^{11}) \text{ (see (45)),}\\
\boldsymbol{x}_t^{(\ell)}[10:13] &: \text{ aggregated transition statistics,}\\
\boldsymbol{x}_t^{(\ell)}[14:13+K] &: \text{ logits } (g_t^{(1)}, \dots, g_t^{(K)}).
\end{aligned}$$

Step 1: Embedding layer. We use token set $\{\mathrm{BOS}, 0, 1\}$, so $S = 3$. Let

$$\boldsymbol{e}_{\mathrm{BOS}} := \begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \boldsymbol{e}_0 := \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \boldsymbol{e}_1 := \begin{bmatrix}0\\0\\1\end{bmatrix} \in \mathbb{R}^3$$

denote the standard basis vectors corresponding to the tokens BOS, 0, and 1, respectively. We choose the embedding matrix $\boldsymbol{\Phi} \in \mathbb{R}^{d\times 3}$ as

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & 0 & 1\\ 1 & 1 & 1\\ 0 & 0 & 0\\ 0 & 1 & 1\\ \mathbf{0}_{(d-4)\times 1} & \mathbf{0}_{(d-4)\times 1} & \mathbf{0}_{(d-4)\times 1}\end{bmatrix}.$$

Hence, for $t \in [n+1]$, the embedding $\boldsymbol{x}_t^{(0)} = \boldsymbol{\Phi}\boldsymbol{e}_{\bar{s}_t}$ satisfies

$$\boldsymbol{x}_t^{(0)}[1] = \mathbb{I}\{\bar{s}_t = 1\} = \begin{cases}0, & \bar{s}_t \in \{\mathrm{BOS}, 0\},\\ 1, & \bar{s}_t = 1,\end{cases} \quad \boldsymbol{x}_t^{(0)}[2] = 1, \ \boldsymbol{x}_t^{(0)}[4] = \mathbb{I}\{\bar{s}_t \in \{0,1\}\} = \begin{cases}0, & \bar{s}_t = \mathrm{BOS},\\ 1, & \bar{s}_t \in \{0,1\}.\end{cases}$$

All remaining coordinates are zero.

Step 2: The first attention layer copies predecessor token validity. The first layer uses one attention head with head dimension $d_h^{(0)} = 2$. For each $t \in [n+1]$, define

$$\boldsymbol{q}_t^{(0)} := \boldsymbol{W}_Q^{(0)}\boldsymbol{x}_t^{(0)}, \quad \boldsymbol{k}_i^{(0)} := \boldsymbol{W}_K^{(0)}\boldsymbol{x}_i^{(0)} + \boldsymbol{p}_{t-i}^{(0),K}, \quad \boldsymbol{v}_i^{(0)} := \boldsymbol{W}_V^{(0)}\boldsymbol{x}_i^{(0)} + \boldsymbol{p}_{t-i}^{(0),V}, \tag{43}$$

where

$$\boldsymbol{W}_Q^{(0)} = \begin{bmatrix}0 & 1 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & 0 & \cdots & 0\end{bmatrix} \in \mathbb{R}^{2\times d}, \quad \boldsymbol{W}_K^{(0)} = \mathbf{0} \in \mathbb{R}^{2\times d},$$

$$\boldsymbol{W}_V^{(0)} = \begin{bmatrix}1 & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & 1 & \cdots & 0\end{bmatrix} \in \mathbb{R}^{2\times d}, \quad \boldsymbol{W}_O^{(0)} = \begin{bmatrix} 0 & 0\\ 0 & 0\\ 1 & 0\\ 0 & 0\\ 0 & 1\\ \mathbf{0}_{(d-5)\times 1} & \mathbf{0}_{(d-5)\times 1}\end{bmatrix} \in \mathbb{R}^{d\times 2},$$

and relative positional encodings

$$\boldsymbol{p}_r^{(0),V} = \mathbf{0} \in \mathbb{R}^2, \quad \boldsymbol{p}_r^{(0),K} = \begin{cases}(\beta, 0), & r = 1,\\ (0,0), & r \neq 1,\end{cases} \tag{44}$$

where $\beta > 0$ is a large scalar. By construction, we have $\boldsymbol{q}_t^{(0)} = (1, 0)$. For every $t \geq 2$, it follows from (43)

$$\boldsymbol{k}_{t-1}^{(0)} = (\beta, 0), \quad \boldsymbol{k}_i^{(0)} = (0, 0),\ \forall i \neq t-1.$$

Hence, by (24), as $\beta \to \infty$

$$\mathrm{att}_{t,t-1}^{(0)} = \frac{e^\beta}{e^\beta + (t-1)} \to 1, \quad \mathrm{att}_{t,i}^{(0)} = \frac{1}{e^\beta + (t-1)} \to 0,\ \forall i \neq t-1.$$

Therefore, for every $t \geq 2$, we have

$$\sum_{i=1}^{t} \mathrm{att}_{t,i}^{(0)}\, \boldsymbol{v}_i^{(0)} \to \boldsymbol{v}_{t-1}^{(0)} = \boldsymbol{W}_V^{(0)} \boldsymbol{x}_{t-1}^{(0)} = \begin{bmatrix} \boldsymbol{x}_{t-1}^{(0)}[1] \\ \boldsymbol{x}_{t-1}^{(0)}[4] \end{bmatrix} = \begin{bmatrix} \mathbb{I}\{\bar{s}_{t-1} = 1\} \\ \mathbb{I}\{\bar{s}_{t-1} \in \{0,1\}\} \end{bmatrix}.$$

At $t = 1$, the predecessor is irrelevant because $\bar{s}_1 = \mathrm{BOS}$, so $\boldsymbol{x}_1^{(0)}[4] = 0$, and the local transition indicators defined below vanish automatically. Using this and the definition of $\boldsymbol{W}_O^{(0)}$, in the hard-attention limit, the relevant coordinates of $\widetilde{\boldsymbol{x}}_t^{(1)}$ satisfy

$$\widetilde{\boldsymbol{x}}_t^{(1)}[1] = \mathbb{I}\{\bar{s}_t = 1\}, \qquad \widetilde{\boldsymbol{x}}_t^{(1)}[2] = 1, \qquad \widetilde{\boldsymbol{x}}_t^{(1)}[4] = \mathbb{I}\{\bar{s}_t \in \{0,1\}\},$$

and, for every $t \geq 2$,

$$\widetilde{\boldsymbol{x}}_t^{(1)}[3] = \mathbb{I}\{\bar{s}_{t-1} = 1\}, \qquad \widetilde{\boldsymbol{x}}_t^{(1)}[5] = \mathbb{I}\{\bar{s}_{t-1} \in \{0,1\}\}.$$

Moreover, at $t = 1$, we have $\widetilde{\boldsymbol{x}}_1^{(1)}[3] = \widetilde{\boldsymbol{x}}_1^{(1)}[5] = 0$.

Step 3: The first feedforward layer computes exact local transition indicators. We now define the four transition indicators so that they are nonzero only when both the current token and its predecessor are genuine binary tokens. Using the coordinates

$$s := \widetilde{\boldsymbol{x}}_t^{(1)}[1], \quad u := \widetilde{\boldsymbol{x}}_t^{(1)}[4], \quad r := \widetilde{\boldsymbol{x}}_t^{(1)}[3], \quad v := \widetilde{\boldsymbol{x}}_t^{(1)}[5],$$

we define

$$z_t^{00} := \mathrm{ReLU}(u + v - s - r - 1), \quad z_t^{01} := \mathrm{ReLU}(u + v + s - r - 2), \tag{45}$$

$$z_t^{10} := \mathrm{ReLU}(u + v - s + r - 2), \quad z_t^{11} := \mathrm{ReLU}(u + v + s + r - 3). \tag{46}$$

A direct check shows that

$$z_t^{ab} = \mathbb{I}\{(\bar{s}_{t-1}, \bar{s}_t) = (a, b)\}, \qquad a, b \in \{0, 1\},$$

for all $t \in [n+1]$. In particular, since $\bar{s}_1 = \mathrm{BOS}$ and $\bar{s}_2 = s_1$, we have

$$z_1^{ab} = z_2^{ab} = 0, \qquad \forall\, a, b \in \{0, 1\}.$$

Choose $\boldsymbol{W}_1^{(0)}$ and $\boldsymbol{W}_2^{(0)}$ accordingly so that these quantities are written into coordinates $6:9$, i.e.,

$$\boldsymbol{W}_1^{(0)} = \begin{bmatrix} -1 & -1 & -1 & 1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & -1 & 1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & -2 & 1 & 1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -3 & 1 & 1 & 1 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{4 \times d}, \quad \boldsymbol{W}_2^{(0)} = \begin{bmatrix} \boldsymbol{0}_{5\times 4} \\ \boldsymbol{I}_4 \\ \boldsymbol{0}_{(d-9)\times 4} \end{bmatrix} \in \mathbb{R}^{d \times 4}.$$

Step 4: The second attention layer averages the local indicators. The second layer uses one attention head with head dimension $d_h^{(1)} = d$. We choose

$$\boldsymbol{W}_Q^{(1)} = \boldsymbol{0}, \quad \boldsymbol{W}_K^{(1)} = \boldsymbol{0}, \quad \boldsymbol{p}_r^{(1),K} = \boldsymbol{0}, \quad \boldsymbol{p}_r^{(1),V} = \boldsymbol{0},$$

so that

$$\mathrm{att}_{t,i}^{(1)} = \frac{1}{t},\ \forall i \in [t].$$

We choose $\boldsymbol{W}_V^{(1)}$ so that it reads the local indicators in coordinates $6:9$ and places them into the first four coordinates of the attention output, and choose $\boldsymbol{W}_O^{(1)}$ so that these are written into coordinates $10:13$. Then, at $t = n+1$,

$$\widetilde{\boldsymbol{x}}_{n+1}^{(2)}[10:13] = \frac{1}{n+1}\sum_{i=1}^{n+1}(z_i^{00}, z_i^{01}, z_i^{10}, z_i^{11}).$$

Since $z_1^{ab} = z_2^{ab} = 0$ for all $a, b \in \{0,1\}$ and $z_i^{ab} = \mathbb{I}\{(s_{i-2}, s_{i-1}) = (a,b)\}$ for $i = 3, \ldots, n+1$, it follows that

$$\sum_{i=1}^{n+1}(z_i^{00}, z_i^{01}, z_i^{10}, z_i^{11}) = (N_{00}, N_{01}, N_{10}, N_{11}).$$

Therefore,

$$\widetilde{\boldsymbol{x}}_{n+1}^{(2)}[10:13] = \frac{1}{n+1}(N_{00}, N_{01}, N_{10}, N_{11}).$$

Step 5: The second feedforward layer computes the logits $g_k(s_{1:n})$. We define

$$\boldsymbol{x}_{n+1}^{(2)} = \widetilde{\boldsymbol{x}}_{n+1}^{(2)} + \boldsymbol{W}_2^{(1)}\,\mathrm{ReLU}\big(\boldsymbol{W}_1^{(1)}\widetilde{\boldsymbol{x}}_{n+1}^{(2)}\big),$$

with $d_{\mathrm{ff}} = 4$. We choose $\boldsymbol{W}_1^{(1)}$ to read coordinates $10:13$. Since these coordinates are nonnegative, the ReLU acts trivially:

$$\mathrm{ReLU}\big(\boldsymbol{W}_1^{(1)}\widetilde{\boldsymbol{x}}_{n+1}^{(2)}\big) = \frac{1}{n+1}\begin{bmatrix} N_{00} \\ N_{01} \\ N_{10} \\ N_{11} \end{bmatrix}.$$

Now define $\boldsymbol{W}_2^{(1)} \in \mathbb{R}^{d\times 4}$ so that, for $k = 1, \ldots, K$,

$$\boldsymbol{W}_2^{(1)}[13+k,:] = (n+1)\cdot\begin{bmatrix}\log(1-p^{(k)}) & \log p^{(k)} & \log q^{(k)} & \log(1-q^{(k)})\end{bmatrix},$$

and all other rows are zero. Then

$$\boldsymbol{x}_{n+1}^{(2)}[13+k] = N_{00}\log(1-p^{(k)}) + N_{01}\log p^{(k)} + N_{10}\log q^{(k)} + N_{11}\log(1-q^{(k)}) =: g_k(s_{1:n}),$$

for every $k = 1, \ldots, K$.

Step 6: The readout map computes the Bayes predictor. We define the readout map $\Psi_{\boldsymbol{\theta}} : \mathbb{R}^d \to [0,1]$ by

$$\Psi_{\boldsymbol{\theta}}(\boldsymbol{x}) := \sum_{k=1}^{K}\frac{\exp(x_{13+k})}{\sum_{r=1}^{K}\exp(x_{13+r})}\big[(1-x_1)p^{(k)} + x_1(1-q^{(k)})\big].$$

Since the first coordinate is preserved through the residual blocks, we have

$$\boldsymbol{x}_{n+1}^{(2)}[1] = s_n.$$

Therefore,

$$f_{\boldsymbol{\theta}}(s_{1:n}) = \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}_{n+1}^{(2)}) = \sum_{k=1}^{K}\frac{\exp(g_k(s_{1:n}))}{\sum_{r=1}^{K}\exp(g_r(s_{1:n}))}\big[(1-s_n)p^{(k)} + s_n(1-q^{(k)})\big].$$

By the definition of $\alpha_k(s_{1:n})$, this equals

$$f_{\boldsymbol{\theta}}(s_{1:n}) = \sum_{k=1}^{K}\alpha_k(s_{1:n})\big[(1-s_n)p^{(k)} + s_n(1-q^{(k)})\big] = f_K^*(s_{1:n}).$$

□

Remark 2. *The only approximation in the above construction occurs in the first attention layer, where the relative positional encoding $p_{t-i}^{(0),K}$ is used to make the attention concentrate on the predecessor position $i = t-1$. Hence the construction is exact in the hard-attention limit $\beta \to \infty$. Once the predecessor features are recovered, all subsequent operations are exact, including the computation of the local transition indicators, the aggregation of transition counts, the formation of the logits $g_k(s_{1:n})$, and the final readout. Furthermore, for any fixed $\varepsilon > 0$, one can choose β sufficiently large so that the resulting transformer approximates the Bayes predictor $f_K^*(s_{1:n})$ within ε, uniformly over all binary sequences of length n.*

D. Proofs in Section 3.2

D.1. Proof of Theorem 3

We first show that, for any fixed prefix, the finite-prior Bayes predictor converges to the uniform-prior Bayes predictor as the number of pretraining tasks K increases.

Proposition 1. *Fix a prefix $s_{1:n} \in \{0,1\}^n$. For any $\delta \in (0,1)$, if*

$$K \geq \frac{32 \log(4/\delta)}{\rho_{\text{unif}}(s_{1:n})}, \tag{47}$$

then it holds with probability at least $1-\delta$ over the draw of the finite pretraining task family that

$$|f_K^*(s_{1:n}) - f_{\text{unif}}^*(s_{1:n})| \leq 8\sqrt{\frac{\log(4/\delta)}{K\rho_{\text{unif}}(s_{1:n})}}, \tag{48}$$

where $\rho_{\text{unif}}(s_{1:n})$ is defined in (13).

Proof. For simplicity, write $\boldsymbol{P} := \boldsymbol{P}(p,q)$ and $\rho := \rho_{\text{unif}}(s_{1:n})$. According to (9) and (29), we write

$$f_K^*(s_{1:n}) = \frac{A_K}{\rho_K}, \qquad f_{\text{unif}}^*(s_{1:n}) = \frac{A}{\rho},$$

where

$$A_K := \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}_k(s_{1:n})\boldsymbol{P}_{s_n,1}^{(k)}, \qquad \rho_K := \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}_k(s_{1:n}),$$

and

$$A := \mathbb{E}_{(p,q)\sim\text{Unif}(0,1)^2}\left[\mathbb{P}(s_{1:n} \mid p,q)\boldsymbol{P}_{s_n,1}\right], \qquad \rho := \mathbb{E}_{(p,q)\sim\text{Unif}(0,1)^2}\left[\mathbb{P}(s_{1:n} \mid p,q)\right].$$

Define $X_k := \mathbb{P}_k(s_{1:n})$ and $Y_k := \mathbb{P}_k(s_{1:n})\boldsymbol{P}_{s_n,1}^{(k)}$. Then $X_1, \ldots, X_K$ are independent copies of $\mathbb{P}(s_{1:n} \mid p,q)$, and $Y_1, \ldots, Y_K$ are independent copies of $\mathbb{P}(s_{1:n} \mid p,q)\boldsymbol{P}_{s_n,1}$, where $(p,q) \sim \text{Unif}(0,1)^2$. Hence, $0 \leq X_k \leq 1$, $0 \leq Y_k \leq 1$, and $\mathbb{E}[X_k] = \rho$, $\mathbb{E}[Y_k] = A$. Moreover, since $0 \leq X_k \leq 1$ and $0 \leq Y_k \leq X_k$, we have

$$\text{Var}(X_k) \leq \mathbb{E}[X_k^2] \leq \mathbb{E}[X_k] = \rho, \quad \text{Var}(Y_k) \leq \mathbb{E}[Y_k^2] \leq \mathbb{E}[Y_k] = A \leq \rho.$$

Applying the two-sided consequence of Corollary 2 with $u := \log(4/\delta)$, $b_K := \sqrt{\frac{2\rho u}{K}} + \frac{2u}{3K}$, and $\sigma^2 = \rho$ gives

$$\mathbb{P}\left(|\rho_K - \rho| > b_K\right) \leq 2e^{-u} = \frac{\delta}{2}, \quad \mathbb{P}\left(|A_K - A| > b_K\right) \leq 2e^{-u} = \frac{\delta}{2}.$$

Therefore, by a union bound, with probability at least $1-\delta$,

$$|\rho_K - \rho| \leq b_K, \qquad |A_K - A| \leq b_K. \tag{49}$$

We now lower bound the denominator. Using (47), we have

$$\sqrt{\frac{2\rho u}{K}} \leq \frac{\rho}{4}, \qquad \frac{2u}{3K} \leq \frac{\rho}{48}.$$

Thus $b_K \le \rho/4 + \rho/48 \le \rho/2$. On the event (49), this yields

$$\rho_K \ge \rho - b_K \ge \rho/2. \tag{50}$$

Furthermore, since $0 \le \boldsymbol{P}_{s_n,1} \le 1$, we have $A = \mathbb{E}_{(p,q)\sim\mathrm{Unif}(0,1)^2}\left[\mathbb{P}(s_{1:n} \mid p,q)\boldsymbol{P}_{s_n,1}\right] \le \rho$. On the event (49), we have

$$\begin{aligned}|f_K^*(s_{1:n}) - f_{\mathrm{unif}}^*(s_{1:n})| &= \left|\frac{A_K}{\rho_K} - \frac{A}{\rho}\right| = \left|\frac{\rho(A_K - A) - A(\rho_K - \rho)}{\rho\rho_K}\right| \\ &\le \frac{|A_K - A|}{\rho_K} + \frac{A|\rho_K - \rho|}{\rho\rho_K} \le \frac{b_K}{\rho/2} + \frac{\rho b_K}{\rho(\rho/2)} = \frac{4b_K}{\rho},\end{aligned}$$

where the second inequality follows from (49) and (50). Substituting the definition of b_K, we obtain

$$|f_K^*(s_{1:n}) - f_{\mathrm{unif}}^*(s_{1:n})| \le 4\sqrt{\frac{2u}{K\rho}} + \frac{8u}{3K\rho} \le 8\sqrt{\frac{u}{K\rho}},$$

where the last inequality follows from (47). Finally, using $u = \log(4/\delta)$, we obtain (48). □

The quantity $\rho_{\mathrm{unif}}(s_{1:n})$ is the marginal likelihood of the prefix $s_{1:n}$ under the continuous uniform prior:

$$\rho_{\mathrm{unif}}(s_{1:n}) := \mu_{s_1}\int_0^1 (1-p)^{N_{00}}p^{N_{01}}\,dp\int_0^1 q^{N_{10}}(1-q)^{N_{11}}\,dq, \tag{51}$$

where N_{ij} is defined in (4) for all $i,j \in \{0,1\}$. It determines the number of sampled tasks needed for finite-prior Bayes predictor to approximate uniform-prior Bayes predictor: smaller $\rho_{\mathrm{unif}}(s_{1:n})$ requires larger K. We also define the finite-prior marginal likelihood

$$\rho_K(s_{1:n}) := \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_k(s_{1:n}), \tag{52}$$

where $\mathbb{P}_k(s_{1:n})$ denotes the probability of observing $s_{1:n}$ under the k-th Markov chain. Note that $\rho_K(s_{1:n})$ is the empirical Monte Carlo approximation of $\rho_{\mathrm{unif}}(s_{1:n})$ obtained from the K sampled pretraining tasks. We next provide uniform lower bounds for both $\rho_{\mathrm{unif}}(s_{1:n})$ and $\rho_K(s_{1:n})$ as follows:

Lemma 4. *Consider the setup in Definition 1. The following statements hold:*
(i) It holds that

$$\inf_{s_{1:n}\in\{0,1\}^n}\rho_{\mathrm{unif}}(s_{1:n}) \ge \frac{4\min\{\mu_0,\mu_1\}}{(n+1)^2 2^{n-1}}. \tag{53}$$

(ii) If

$$K \ge \frac{2(n+1)^2 2^{n-1}}{\min\{\mu_0,\mu_1\}}\log\frac{2^n}{\delta}, \tag{54}$$

it holds with probability at least $1-\delta$ *that*

$$\inf_{s_{1:n}\in\{0,1\}^n}\rho_K(s_{1:n}) \ge \frac{2\min\{\mu_0,\mu_1\}}{(n+1)^2 2^{n-1}}. \tag{55}$$

Proof. Let $\mu_{\min} := \min\{\mu_0,\mu_1\}$ and fix any prefix $s_{1:n} \in \{0,1\}^n$. Under the Markov chain with transition matrix $\boldsymbol{P}(p,q)$, the probability of observing $s_{1:n}$ is

$$\mathbb{P}(s_{1:n} \mid p,q) = \mu_{s_1}(1-p)^{N_{00}}p^{N_{01}}q^{N_{10}}(1-q)^{N_{11}},$$

where $N_{ij} := N_{ij}(s_{1:n})$ is defined in (4).

(i) Using (51) and the identity

$$\int_0^1 t^a(1-t)^b\,dt = \frac{a!\,b!}{(a+b+1)!} = \frac{1}{(a+b+1)\binom{a+b}{a}}, \quad \forall a,b \in \mathbb{N}\cup\{0\},$$

we obtain

$$\rho_{\text{unif}}(s_{1:n}) = \mu_{s_1}\frac{1}{(N_0+1)\binom{N_0}{N_{01}}}\frac{1}{(N_1+1)\binom{N_1}{N_{10}}}.$$

Since

$$\binom{N_0}{N_{01}} \le 2^{N_0}, \quad \binom{N_1}{N_{10}} \le 2^{N_1},$$

and $\mu_{s_1} \ge \mu_{\min}$, we have

$$\rho_{\text{unif}}(s_{1:n}) \ge \frac{\mu_{\min}}{(N_0+1)(N_1+1)2^{N_0+N_1}}.$$

Moreover, $N_0 + N_1 = n-1$, and hence

$$(N_0+1)(N_1+1) \le \left(\frac{N_0+N_1+2}{2}\right)^2 = \left(\frac{n+1}{2}\right)^2.$$

Combining the last two displays gives

$$\rho_{\text{unif}}(s_{1:n}) \ge \frac{\mu_{\min}}{((n+1)/2)^2\,2^{n-1}} = \frac{4\mu_{\min}}{(n+1)^2 2^{n-1}}.$$

Since the prefix $s_{1:n}$ was arbitrary, we have (53).

(ii) Given $s_{1:n} \in \{0,1\}^n$, we define $X_k(s_{1:n}) := \mathbb{P}_k(s_{1:n})$ for all $k \in [K]$. Then $X_1,\dots,X_K$ are independent random variables in $[0,1]$. According to (52), we have $\mathbb{E}[X_k(s_{1:n})] = \rho_{\text{unif}}(s_{1:n})$. Using (53), we have

$$\rho_{\text{unif}}(s_{1:n}) \ge L_n := \frac{4\mu_{\min}}{(n+1)^2 2^{n-1}}.$$

By the multiplicative Chernoff bound for independent random variables bounded in $[0,1]$,

$$\mathbb{P}\left(\rho_K(s_{1:n}) \le \frac{1}{2}\rho_{\text{unif}}(s_{1:n})\right) \le \exp\left(-\frac{K\rho_{\text{unif}}(s_{1:n})}{8}\right).$$

Using $\rho_{\text{unif}}(s_{1:n}) \ge L_n$, we get

$$\mathbb{P}\left(\rho_K(s_{1:n}) \le \frac{1}{2}L_n\right) \le \exp\left(-\frac{KL_n}{8}\right).$$

Taking a union bound over all 2^n possible prefixes gives

$$\mathbb{P}\left(\exists s_{1:n} \in \{0,1\}^n : \rho_K(s_{1:n}) \le \frac{1}{2}L_n\right) \le 2^n \exp\left(-\frac{KL_n}{8}\right).$$

Substituting the definition of L_n yields

$$2^n \exp\left(-\frac{KL_n}{8}\right) = 2^n \exp\left(-\frac{K\mu_{\min}}{2(n+1)^2 2^{n-1}}\right).$$

Therefore, with probability at least

$$1 - 2^n \exp\left(-\frac{K\mu_{\min}}{2(n+1)^2 2^{n-1}}\right),$$

we have

$$\rho_K(s_{1:n}) \ge \frac{1}{2}L_n = \frac{2\mu_{\min}}{(n+1)^2 2^{n-1}}$$

simultaneously for all $s_{1:n} \in \{0,1\}^n$. The equivalent sample-size condition (54) follows by requiring

$$2^n \exp\left(-\frac{KL_n}{8}\right) \le \delta,$$

Under this condition, (55) holds with probability at least $1-\delta$. □

We next relate the uniform-prior Bayes predictor $f^*_{\text{unif}}(s_{1:n})$ in Lemma 3 to the empirical Markov predictor $\hat{\boldsymbol{P}}$ defined in (6). This step isolates the smoothing bias caused by the uniform prior.

Lemma 5. *Fix a prefix $s_{1:n} \in \{0,1\}^n$. Suppose that there exists a constant $c_1 \in (0,1)$ such that*

$$N_0 \geq c_1(n-1), \quad N_1 \geq c_1(n-1), \tag{56}$$

where N_0, N_1 are defined in (5). Then, we have

$$\left| f^*_{\text{unif}}(s_{1:n}) - \hat{\boldsymbol{P}}_{s_n,1} \right| \leq \frac{1}{c_1(n-1)+2}.$$

Proof. We consider two cases. First, suppose $s_n = 0$. It follows from Lemma 3 that

$$f^*_{\text{unif}}(s_{1:n}) = \frac{N_{01}+1}{N_{00}+N_{01}+2}, \quad \hat{\boldsymbol{P}}_{s_n,1} = \hat{\boldsymbol{P}}_{0,1} = \frac{N_{01}}{N_{00}+N_{01}},$$

where N_{ij} is defined in (4). It follows from (56) that $N_0 > 0$. Then, we compute

$$\left| f^*_{\text{unif}}(s_{1:n}) - \hat{\boldsymbol{P}}_{0,1} \right| = \left| \frac{N_{01}+1}{N_0+2} - \frac{N_{01}}{N_0} \right| = \left| \frac{N_0(N_{01}+1) - N_{01}(N_0+2)}{N_0(N_0+2)} \right| = \frac{|N_{00}-N_{01}|}{N_0(N_0+2)}.$$

Since $|N_{00} - N_{01}| \leq N_{00} + N_{01} = N_0$, we obtain

$$\left| f^*_{\text{unif}}(s_{1:n}) - \hat{\boldsymbol{P}}_{0,1} \right| \leq \frac{1}{N_0+2} \leq \frac{1}{c_1(n-1)+2},$$

where the last inequality follows from (56). The case $s_n = 1$ follows analogously, using

$$f^*_{\text{unif}}(s_{1:n}) = \frac{N_{11}+1}{N_1+2}, \quad \hat{\boldsymbol{P}}_{1,1} = \frac{N_{11}}{N_1}.$$

This completes the proof. □

Proof of Theorem 3. For simplicity, write $s_{1:n} = s^{\text{ICL}}_{1:n}$, $\rho := \rho_{\text{unif}}(s_{1:n})$, and $\mu_{\min} := \min\{\mu_0, \mu_1\}$. By Lemma 4, we have

$$\rho \geq \frac{4\mu_{\min}}{(n+1)^2 2^{n-1}}.$$

Therefore, the assumed lower bound in (14) on K implies $K \geq 32\log(4/\delta)/\rho$. This, together with Proposition 1, yields that, with probability at least $1-\delta$,

$$|f^*_K(s_{1:n}) - f^*_{\text{unif}}(s_{1:n})| \leq 8\sqrt{\frac{\log(4/\delta)}{K\rho}}. \tag{57}$$

It remains to compare the uniform-prior Bayes predictor with the empirical Markov predictor. Since $\hat{c}_0 \in [\gamma, 1-\gamma]$, we have

$$N_0 = \hat{c}_0(n-1) \geq \gamma(n-1), \quad N_1 = \hat{c}_1(n-1) = (1-\hat{c}_0)(n-1) \geq \gamma(n-1).$$

Thus, applying Lemma 5 with $c_1 = \gamma$ gives

$$\left| f^*_{\text{unif}}(s_{1:n}) - \hat{\boldsymbol{P}}_{s_n,1} \right| \leq \frac{1}{\gamma(n-1)+2} \leq \frac{1}{\gamma(n-1)}. \tag{58}$$

Finally, by the triangle inequality, on the same high-probability event,

$$\begin{aligned} \left| f^*_K(s_{1:n}) - \hat{\boldsymbol{P}}_{s_n,1} \right| &\leq |f^*_K(s_{1:n}) - f^*_{\text{unif}}(s_{1:n})| + \left| f^*_{\text{unif}}(s_{1:n}) - \hat{\boldsymbol{P}}_{s_n,1} \right| \\ &\leq 8\sqrt{\frac{\log(4/\delta)}{K\rho_{\text{unif}}(s_{1:n})}} + \frac{1}{\gamma(n-1)}. \end{aligned}$$

This completes the proof. □

Lemma 6. *For $3 \le n \le T-1$, let*

$$\mathcal{S}_n := \{s_{1:n} \in \{0,1\}^n : N_0 \ge 1,\ N_1 \ge 1\}, \tag{59}$$

where N_0 and N_1 are defined in (5). Then, it holds that

$$|\mathcal{S}_n| = 2^n - 4 \ge 2^{n-1}. \tag{60}$$

Moreover, for every $s_{1:n} \in \mathcal{S}_n$,

$$\hat{c}_0, \hat{c}_1 \in \left[\frac{1}{T+1}, 1 - \frac{1}{T+1}\right]. \tag{61}$$

Proof. Recall that N_0 and N_1 count the number of occurrences of states 0 and 1, respectively, among the first $n-1$ positions:

$$N_0 + N_1 = n - 1.$$

The condition $N_0, N_1 \ge 1$ fails if and only if the first $n-1$ symbols are all 0 or all 1. If $s_1 = \cdots = s_{n-1} = 0$, then s_n can be either 0 or 1, giving two bad prefixes. Similarly, if $s_1 = \cdots = s_{n-1} = 1$, this gives two more bad prefixes. Hence there are exactly 4 bad prefixes, and therefore

$$|\mathcal{S}_n| = 2^n - 4 \ge 2^{n-1},$$

where the last inequality uses $n \ge 3$.

For any $s_{1:n} \in \mathcal{S}_n$, we have $N_0, N_1 \ge 1$. Using (5), we have $\hat{c}_0, \hat{c}_1 \ge 1/(n-1)$. Moreover, since $N_0 + N_1 = n-1$ and both $N_0, N_1 \ge 1$,

$$\hat{c}_0, \hat{c}_1 \le 1 - \frac{1}{n-1}.$$

This, together with $n \le T-1$, implies (61). □

D.2. Proof of Theorem 4

Proof of Theorem 4. For ease of notation, we write $s_{1:n} := s_{1:n}^{\mathrm{ICL}}$ and $\mu_{\min} := \min\{\mu_0, \mu_1\}$. Let

$$\mathcal{S} := \bigcup_{m=\hat{T}}^{T-1} \{0,1\}^m, \quad M := |\mathcal{S}| \le 2^T, \quad \rho_T := \frac{4\mu_{\min}}{T^2 2^{T-2}}, \quad \underline{\gamma}_T := \frac{1}{T+1}.$$

For every non-degenerate prefix $s_{1:m} \in \mathcal{S}$, meaning $N_0, N_1 \ge 1$, by Lemma 6,

$$\hat{c}_0, \hat{c}_1 \in \left[\frac{1}{T+1}, 1 - \frac{1}{T+1}\right]. \tag{62}$$

(i) Uniform control of finite-prior Bayes approximation. By Lemma 4, for every $s_{1:m} \in \mathcal{S}$, since $m \le T-1$,

$$\rho_{\mathrm{unif}}(s_{1:m}) \ge \frac{4\mu_{\min}}{(m+1)^2 2^{m-1}} \ge \frac{4\mu_{\min}}{T^2 2^{T-2}} = \rho_T.$$

Suppose that

$$K \ge \frac{32 \log(4M/\delta_1)}{\rho_T}.$$

Applying Proposition 1 with failure probability δ_1/M, and then taking a union bound over all prefixes in $\mathcal{S}$, we obtain that, with probability at least $1-\delta_1$, for every $s_{1:m} \in \mathcal{S}$,

$$|f_K^*(s_{1:m}) - f_{\mathrm{unif}}^*(s_{1:m})| \le 8\sqrt{\frac{\log(4M/\delta_1)}{K\rho_T}} =: \alpha_K. \tag{63}$$

In the rest of the proof, we work on this high-probability event.

(ii) Controlling the trained predictor via finite-to-uniform excess loss. Next, we define the finite-to-uniform excess loss

$$\Delta_K := \mathcal{L}(f^*_{\text{unif}}) - \mathcal{L}(f^*_K). \tag{64}$$

For any predictor g, it follows from (2) that

$$\mathcal{L}(g) - \mathcal{L}(f^*_K) = \frac{1}{T-\hat{T}} \sum_{m=\hat{T}}^{T-1} \sum_{s_{1:m}} \rho_K(s_{1:m}) \text{KL}\left(\text{Bern}(f^*_K(s_{1:m})) \,\|\, \text{Bern}(g(s_{1:m}))\right). \tag{65}$$

Taking $g = f^*_{\text{unif}}$ and using (64), we have

$$\Delta_K = \frac{1}{T-\hat{T}} \sum_{m=\hat{T}}^{T-1} \sum_{s_{1:m}} \rho_K(s_{1:m}) \text{KL}\left(\text{Bern}(f^*_K(s_{1:m})) \,\|\, \text{Bern}(f^*_{\text{unif}}(s_{1:m}))\right). \tag{66}$$

We now upper bound this KL term. According to Lemma 3, we have

$$f^*_{\text{unif}}(s_{1:m}) \in \left[\frac{1}{T+1}, 1 - \frac{1}{T+1}\right]. \tag{67}$$

One can verify that

$$\text{KL}\left(\text{Bern}(a) \,\|\, \text{Bern}(b)\right) \le \frac{(a-b)^2}{b(1-b)}, \ \forall a \in [0,1], b \in (0,1).$$

This, together with $a = f^*_K(s_{1:m})$ and $b = f^*_{\text{unif}}(s_{1:m})$, yields for every $s_{1:m} \in \mathcal{S}$,

$$\text{KL}\left(\text{Bern}(f^*_K(s_{1:m})) \,\|\, \text{Bern}(f^*_{\text{unif}}(s_{1:m}))\right) \le \frac{(f^*_K(s_{1:m}) - f^*_{\text{unif}}(s_{1:m}))^2}{f^*_{\text{unif}}(s_{1:m})(1 - f^*_{\text{unif}}(s_{1:m}))} \le \frac{(T+1)^2 \alpha_K^2}{T},$$

where the last inequality follows from (63) and (67). This, together with (66), yields

$$\Delta_K \le \frac{1}{T-\hat{T}} \sum_{m=\hat{T}}^{T-1} \sum_{s_{1:m}} \frac{(T+1)^2 \alpha_K^2 \rho_K(s_{1:m})}{T} \le \frac{(T+1)^2 \alpha_K^2}{T},$$

where the last inequality uses $\sum_{s_{1:m}} \rho_K(s_{1:m}) = 1$ for each m. Substituting the definition of α_K in (63) yields

$$\Delta_K \le \frac{64(T+1)^2}{T\rho_T} \frac{\log(4M/\delta_1)}{K}. \tag{68}$$

We now compare the trained transformer with f^*_{unif}. Since $f^*_{\text{unif}} \in \mathcal{F}$ and $f_{\boldsymbol{\theta}^\star}$ globally minimizes $\mathcal{L}$ over $\mathcal{F}$, we have $\mathcal{L}(f_{\boldsymbol{\theta}^\star}) \le \mathcal{L}(f^*_{\text{unif}})$. Therefore,

$$\mathcal{L}(f_{\boldsymbol{\theta}^\star}) - \mathcal{L}(f^*_K) \le \mathcal{L}(f^*_{\text{unif}}) - \mathcal{L}(f^*_K) = \Delta_K. \tag{69}$$

Applying Pinsker's inequality (see Lemma 8) to (65) with $g = f_{\boldsymbol{\theta}^\star}$, and keeping only the length-n terms, gives

$$\Delta_K \ge \mathcal{L}(f_{\boldsymbol{\theta}^\star}) - \mathcal{L}(f^*_K) \ge \frac{2}{T-\hat{T}} \sum_{s_{1:n}} \rho_K(s_{1:n}) \left(f_{\boldsymbol{\theta}^\star}(s_{1:n}) - f^*_K(s_{1:n})\right)^2. \tag{70}$$

Let $\mathcal{S}_n$ be the set of non-degenerate prefixes of length n. We additionally work on the high-probability event from Lemma 4(ii). That is, suppose that

$$K \ge \frac{2(n+1)^2 2^{n-1}}{\mu_{\min}} \log\left(\frac{2^n}{\delta_2}\right),$$

with probability at least $1 - \delta_2$,

$$\rho_K(s_{1:n}) \ge \frac{2\mu_{\min}}{(n+1)^2 2^{n-1}}, \quad \forall s_{1:n} \in \{0,1\}^n.$$

Combining this event with the Bernstein finite-to-uniform event from part (i), it is sufficient to assume

$$K \geq \max\left\{\frac{32\log(4M/\delta_1)}{\rho_T},\ \frac{2(n+1)^2 2^{n-1}}{\mu_{\min}}\log\frac{2^n}{\delta_2}\right\}.$$

Since $n \leq T-1$, we have $(n+1)^2 \leq T^2$ and $2^{n-1} \leq 2^{T-2}$. Therefore,

$$\frac{2(n+1)^2 2^{n-1}}{\mu_{\min}} \leq \frac{2T^2 2^{T-2}}{\mu_{\min}} = \frac{8}{\rho_T}.$$

Thus, it is enough to impose the simpler sufficient condition

$$K \geq \max\left\{\frac{32\log(4M/\delta_1)}{\rho_T},\ \frac{8}{\rho_T}\log\frac{2^n}{\delta_2}\right\}. \tag{71}$$

Restricting the sum in (70) to $\mathcal{S}_n$ and using the lower bound on $\rho_K(s_{1:n})$, we obtain

$$\sum_{s_{1:n}\in\mathcal{S}_n}\left(f_{\boldsymbol{\theta}^\star}(s_{1:n}) - f_K^*(s_{1:n})\right)^2 \leq \frac{(n+1)^2 2^{n-1}(T-\hat{T})}{4\mu_{\min}}\Delta_K.$$

By Lemma 6, $|\mathcal{S}_n| \geq 2^{n-1}$. Hence,

$$\frac{1}{|\mathcal{S}_n|}\sum_{s_{1:n}\in\mathcal{S}_n}\left(f_{\boldsymbol{\theta}^\star}(s_{1:n}) - f_K^*(s_{1:n})\right)^2 \leq \frac{(n+1)^2(T-\hat{T})}{4\mu_{\min}}\Delta_K. \tag{72}$$

Similarly, applying the same KL decomposition and Pinsker's inequality with $g = f^*_{\text{unif}}$, and again keeping only the length-n terms, gives

$$\Delta_K = \mathcal{L}(f^*_{\text{unif}}) - \mathcal{L}(f_K^*) \geq \frac{2}{T-\hat{T}}\sum_{s_{1:n}}\rho_K(s_{1:n})\left(f^*_{\text{unif}}(s_{1:n}) - f_K^*(s_{1:n})\right)^2.$$

Repeating the same prefix-mass argument yields

$$\frac{1}{|\mathcal{S}_n|}\sum_{s_{1:n}\in\mathcal{S}_n}\left(f^*_{\text{unif}}(s_{1:n}) - f_K^*(s_{1:n})\right)^2 \leq \frac{(n+1)^2(T-\hat{T})}{4\mu_{\min}}\Delta_K. \tag{73}$$

(iii) From average error to prompt-wise task learning. For $s_{1:n} \in \mathcal{S}_n$, we define

$$Z(s_{1:n}) := \left|f_{\boldsymbol{\theta}^\star}(s_{1:n}) - f^*_{\text{unif}}(s_{1:n})\right|^2.$$

By the triangle inequality and $(a+b)^2 \leq 2a^2 + 2b^2$,

$$Z(s_{1:n}) \leq 2\left(f_{\boldsymbol{\theta}^\star}(s_{1:n}) - f_K^*(s_{1:n})\right)^2 + 2\left(f_K^*(s_{1:n}) - f^*_{\text{unif}}(s_{1:n})\right)^2.$$

Averaging over $s_{1:n} \in \mathcal{S}_n$, and using (72) and (73), gives

$$\frac{1}{|\mathcal{S}_n|}\sum_{s_{1:n}\in\mathcal{S}_n} Z(s_{1:n}) \leq \frac{(n+1)^2(T-\hat{T})}{\mu_{\min}}\Delta_K.$$

Therefore, if $s_{1:n}^{\text{ICL}}$ is sampled uniformly from $\mathcal{S}_n$, Markov's inequality gives

$$\mathbb{P}\left(Z(s_{1:n}^{\text{ICL}}) > \frac{(n+1)^2(T-\hat{T})}{\alpha\mu_{\min}}\Delta_K\right) \leq \alpha.$$

Equivalently, with probability at least $1-\alpha$ over the sampled prompt,

$$\left|f_{\boldsymbol{\theta}^\star}(s_{1:n}^{\text{ICL}}) - f^*_{\text{unif}}(s_{1:n}^{\text{ICL}})\right| \leq \sqrt{\frac{(n+1)^2(T-\hat{T})}{\alpha\mu_{\min}}\Delta_K}. \tag{74}$$

It remains to compare f^*_{unif} with the empirical Markov predictor. Since $s^{\text{ICL}}_{1:n} \in \mathcal{S}_n$ is non-degenerate and satisfies $\hat{c}_0 \in [\gamma, 1-\gamma]$, Lemma 5 gives

$$\left| f^*_{\text{unif}}(s^{\text{ICL}}_{1:n}) - \hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1} \right| \le \frac{1}{\gamma(n-1)+2}. \tag{75}$$

Combining (74) and (75) by the triangle inequality yields

$$\left| f_{\boldsymbol{\theta}^\star}(s^{\text{ICL}}_{1:n}) - \hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1} \right| \le \sqrt{\frac{(n+1)^2(T-\hat{T})}{\alpha\mu_{\min}}\Delta_K} + \frac{1}{\gamma(n-1)+2}.$$

Finally, substituting (68) gives, with probability at least $1-\delta_1-\delta_2-\alpha$,

$$\left| f_{\boldsymbol{\theta}^\star}(s^{\text{ICL}}_{1:n}) - \hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1} \right| \le \sqrt{\frac{64(n+1)^2(T-\hat{T})(T+1)^2}{\alpha\mu_{\min}T\rho_T}\frac{\log(4M/\delta_1)}{K}} + \frac{1}{\gamma(n-1)+2}.$$

Using $M \le 2^T$ and $\rho_T = 4\mu_{\min}/(T^2 2^{T-2})$, this can be summarized as

$$\left| f_{\boldsymbol{\theta}^\star}(s^{\text{ICL}}_{1:n}) - \hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1} \right| = O_{\mathbb{P}}\left(2^{T/2}T^{3/2}(n+1)\sqrt{\frac{(T-\hat{T})\log(4M/\delta_1)}{\alpha K}} + \frac{1}{n} \right),$$

where the hidden constant may depend on γ and $\mu_{\min}$.

In particular, choosing

$$\alpha = \frac{1}{\log K}, \quad \delta_1 = \delta_2 = \frac{1}{2K},$$

gives probability at least $1 - 1/\log K - 1/K$ and

$$\left| f_{\boldsymbol{\theta}^\star}(s^{\text{ICL}}_{1:n}) - \hat{\boldsymbol{P}}_{s^{\text{ICL}}_n,1} \right| = O\left(2^{T/2}T^{3/2}(n+1)\sqrt{\frac{(T-\hat{T})\log K\,(T+\log K)}{K}} + \frac{1}{n} \right).$$

For fixed T, this simplifies to (18). □

E. Proofs in Section 4

Proof of Theorem 5. To simplify notation, we write $s^{\text{ICL}}_{1:n}$ as $s_{1:n}$. For each $t = m, \dots, n-1$, denote the length-m context by

$$u_t := s_{t-m+1:t}.$$

(i) For the k-th m-th order Markov chain, the probability of observing the prefix $s_{1:n}$ is

$$\mathbb{P}_k(s_{1:n}) = \mu_{s_{1:m}} \prod_{t=m}^{n-1} \boldsymbol{P}^{(k)}_{u_t, s_{t+1}},$$

where $\mu_{s_{1:m}}$ is the probability of the initial m tokens. Using the transition counts $N_{u,j}$, this can be written as

$$\mathbb{P}_k(s_{1:n}) = \mu_{s_{1:m}} \prod_{u\in[S]^m} \prod_{j\in[S]} \left(\boldsymbol{P}^{(k)}_{u,j}\right)^{N_{u,j}}.$$

Taking logarithms, we obtain

$$\begin{aligned}\log \prod_{u\in[S]^m} \prod_{j\in[S]} \left(\boldsymbol{P}^{(k)}_{u,j}\right)^{N_{u,j}} &= \sum_{u\in[S]^m}\sum_{j\in[S]} N_{u,j}\log \boldsymbol{P}^{(k)}_{u,j} = (n-m)\sum_{u:\,N_u>0}\hat{c}_u \sum_{j\in[S]} \hat{\boldsymbol{P}}_{u,j}\log \boldsymbol{P}^{(k)}_{u,j} \\ &= -(n-m)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}),\end{aligned}$$

where $\hat{\boldsymbol{P}}$ is defined in (20) and $D(\cdot,\cdot)$ is defined in (21). Therefore,

$$\mathbb{P}_k(s_{1:n}) = \mu_{s_{1:m}} \exp\left(-(n-m)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\right).$$

Since the initial distribution $\boldsymbol{\mu}$ is shared across all tasks, the factor $\mu_{s_{1:m}}$ cancels in the posterior weights. Hence,

$$\alpha_k(s_{1:n}) = \frac{\exp\left(-(n-m)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\right)}{\sum_{r=1}^{K} \exp\left(-(n-m)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(r)})\right)}.$$

Let $w_k := \exp\left(-(n-m)D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\right)$. Then the Bayes predictor satisfies, for every $j \in [S]$,

$$f_j^*(s_{1:n}) = \frac{\sum_{k=1}^{K} w_k \boldsymbol{P}^{(k)}_{u_n,j}}{\sum_{k=1}^{K} w_k}, \quad \text{where } u_n := s_{n-m+1:n}.$$

By the definitions of k^* and Δ, for every $k \neq k^*$,

$$D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}) - D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k^*)}) \geq \Delta.$$

Therefore, $w_k \leq w_{k^*} \exp\bigl(-(n-m)\Delta\bigr)$ for all $k \neq k^*$. Using the representation of $f^*_{K,j}(s_{1:n})$, we have

$$\begin{aligned}\left|f_j^*(s_{1:n}) - \boldsymbol{P}^{(k^*)}_{u_n,j}\right| &= \left|\frac{\sum_{k\neq k^*} w_k \left(\boldsymbol{P}^{(k)}_{u_n,j} - \boldsymbol{P}^{(k^*)}_{u_n,j}\right)}{w_{k^*} + \sum_{k\neq k^*} w_k}\right| \leq \frac{\sum_{k\neq k^*} w_k}{w_{k^*} + \sum_{k\neq k^*} w_k} \leq \sum_{k\neq k^*} \frac{w_k}{w_{k^*}} \\ &\leq (K-1)\exp\bigl(-(n-m)\Delta\bigr),\end{aligned}$$

where we used $\boldsymbol{P}^{(k)}_{u_n,j} \in [0,1]$. This proves part (i).

(ii) For part (ii), we follow the same decomposition as in the proof of Theorem 3, with the binary empirical transition matrix replaced by the empirical transition tensor. In the multi-state m-th order setting, the empirical Markov predictor at the current context $u_n = s_{n-m+1:n}$ is the probability vector $\hat{\boldsymbol{P}}_{u_n,\cdot} \in \Delta^{S-1}$, whose j-th coordinate $\hat{\boldsymbol{P}}_{u_n,j}$ is the empirical probability that token j follows the context u_n.

The finite-to-uniform step follows by applying the Bernstein argument in Proposition 1 to the S coordinates of the next-token distribution and taking a union bound over these S numerators and the marginal-likelihood denominator. Thus, under the stated lower bound on K, with probability at least $1-\delta$,

$$\left|f^*_{K,j}(s_{1:n}) - f^*_{\text{unif},j}(s_{1:n})\right| \leq 8\sqrt{\frac{\log(2(S+1)/\delta)}{K\rho_{\text{unif}}(s_{1:n})}}, \quad \forall j \in [S].$$

It remains to compare the uniform-prior Bayes predictor with the empirical Markov predictor. Under the uniform Dirichlet prior on each transition vector,

$$f^*_{\text{unif},j}(s_{1:n}) = \frac{N_{u_n,j} + 1}{N_{u_n} + S}.$$

Hence, using $N_{u_n} \geq \gamma(n-m)$,

$$\left|f^*_{\text{unif},j}(s_{1:n}) - \hat{\boldsymbol{P}}_{u_n,j}\right| \leq \frac{S-1}{N_{u_n} + S} \leq \frac{S-1}{\gamma(n-m) + S}.$$

Combining the two displays by the triangle inequality yields the desired bound.

□

F. Auxiliary Results

Lemma 7 (Bernstein inequality (Vershynin, 2018))**.** *Let $Z_1, \ldots, Z_K$ be independent mean-zero random variables satisfying $|Z_k| \le M$ for all $k \in [K]$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{k=1}^{K} Z_k\right| \ge t\right) \le 2\exp\left(-\frac{t^2/2}{\sigma^2 + Mt/3}\right),$$

where $\sigma^2 := \sum_{k=1}^{K} \mathbb{E}[Z_k^2]$.

Based on the above lemma, we have the following corollary:

Corollary 2. *Let $X_1, \ldots, X_K$ be independent random variables satisfying $0 \le X_k \le 1$, $\mathbb{E}[X_k] = m$, and $\mathrm{Var}(X_k) \le \sigma^2$ for all $k \in [K]$. Define*

$$\widehat{m}_K := \frac{1}{K}\sum_{k=1}^{K} X_k.$$

Then, for any $u > 0$,

$$\mathbb{P}\left(|\widehat{m}_K - m| \ge \sqrt{\frac{2\sigma^2 u}{K}} + \frac{2u}{3K}\right) \le 2e^{-u}.$$

Proof. Let $Z_k := X_k - m$. Then $\mathbb{E}[Z_k] = 0$. Since $0 \le X_k \le 1$, we have $m \in [0, 1]$, and hence $-m \le Z_k \le 1 - m$, which implies $|Z_k| \le 1$. Moreover,

$$\sum_{k=1}^{K} \mathbb{E}[Z_k^2] = \sum_{k=1}^{K} \mathrm{Var}(X_k) \le K\sigma^2.$$

Applying Lemma 7 with $t = K\epsilon$ gives

$$\mathbb{P}\left(|\widehat{m}_K - m| \ge \epsilon\right) \le 2\exp\left(-\frac{K^2\epsilon^2/2}{K\sigma^2 + K\epsilon/3}\right).$$

Equivalently, the Bernstein tail bound implies that, with probability at least $1 - 2e^{-u}$,

$$|\widehat{m}_K - m| \le \sqrt{\frac{2\sigma^2 u}{K}} + \frac{2u}{3K}.$$

This proves the result. □

Lemma 8 (Pinsker's inequality for Bernoulli distributions)**.** *For any $x, y \in (0, 1)$, we have*

$$\mathrm{KL}(\mathrm{Bern}(x) \,\|\, \mathrm{Bern}(y)) \ge 2(x - y)^2, \tag{76}$$

where $\mathrm{Bern}(a)$ denotes the Bernoulli distribution with mean a and $\mathrm{KL}(\cdot, \cdot)$ denotes the Kullback–Leibler divergence.

Proof. We compute

$$\mathrm{KL}(\mathrm{Bern}(x) \,\|\, \mathrm{Bern}(y)) = x \log\frac{x}{y} + (1 - x)\log\frac{1 - x}{1 - y} \tag{77}$$

Fix $x \in (0, 1)$ and define, for $y \in (0, 1)$,

$$f(y) := x \log\frac{x}{y} + (1 - x)\log\frac{1 - x}{1 - y} - 2(x - y)^2.$$

It suffices to show that $f(y) \ge 0$ for all $y \in (0, 1)$. Differentiating with respect to y, we obtain

$$f'(y) = -\frac{x}{y} + \frac{1 - x}{1 - y} + 4(x - y).$$

Since

$$-\frac{x}{y} + \frac{1-x}{1-y} = \frac{y-x}{y(1-y)},$$

we have

$$f'(y) = (y-x)\left(\frac{1}{y(1-y)} - 4\right).$$

Because $y(1-y) \leq 1/4$ for all $y \in (0,1)$, it follows that

$$\frac{1}{y(1-y)} - 4 \geq 0.$$

Hence,

$$f'(y) \leq 0 \quad \text{if } y < x, \qquad f'(y) \geq 0 \quad \text{if } y > x.$$

Therefore, f attains its global minimum at $y = x$. Since

$$f(x) = x \log 1 + (1-x) \log 1 - 2(x-x)^2 = 0,$$

we conclude that $f(y) \geq 0$ for all $y \in (0,1)$. This, together with (77), yields (76). □