Optimal Prediction of Markov Chains With and Without Spectral Gap

Yanjun Han, *Member, IEEE*, Soham Jana, and Yihong Wu

***Abstract*— We study the following learning problem with dependent data: Observing a trajectory of length n from a stationary Markov chain with k states, the goal is to predict the next state. For $3 \le k \le O(\sqrt{n})$, using techniques from universal compression, the optimal prediction risk in Kullback-Leibler divergence is shown to be $\Theta(\frac{k^2}{n}\log\frac{n}{k^2})$, in contrast to the optimal rate of $\Theta(\frac{\log\log n}{n})$ for $k = 2$ previously shown in Falahatgar et al. (2016). These rates, slower than the parametric rate of $O(\frac{k^2}{n})$, can be attributed to the memory in the data, as the spectral gap of the Markov chain can be arbitrarily small. To quantify the memory effect, we study irreducible reversible chains with a prescribed spectral gap. In addition to characterizing the optimal prediction risk for two states, we show that, as long as the spectral gap is not excessively small, the prediction risk in the Markov model is $O(\frac{k^2}{n})$, which coincides with that of an iid model with the same number of parameters. Extensions to higher-order Markov chains are also obtained.**

***Index Terms*— Markov chains, prediction, redundancy, spectral gap, mixing time, Kullback Leibler risk, higher-order Markov chains.**

I. Introduction

LEARNING distributions from samples is a central question in statistics and machine learning. While significant progress has been achieved in property testing and estimation based on independent and identically distributed (iid) data, for many applications, most notably natural language processing, two new challenges arise: (a) Modeling data as independent observations fails to capture their temporal dependency; (b) Distributions are commonly supported on a large domain whose cardinality is comparable to or even exceeds the sample size. Continuing the progress made in [1] and [2], in this paper we study the following prediction problem with dependent data modeled as Markov chains.

Manuscript received 6 May 2022; revised 10 October 2022; accepted 8 January 2023. Date of publication 27 January 2023; date of current version 19 May 2023. The work of Yihong Wu was supported in part by NSF under Grant CCF-1900507, in part by NSF CAREER under Award CCF-1651588, and in part by the Alfred Sloan Fellowship. An earlier version of this paper was presented in part at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). *(Corresponding author: Soham Jana.)*

Yanjun Han is with the Institute of Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: yjhan@mit.edu).

Soham Jana is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: soham.jana@princeton.edu).

Yihong Wu is with the Department of Statistics and Data Science, Yale University, New Haven, CT 06520 USA (e-mail: yihong.wu@yale.edu).

Communicated by V. Tan, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2023.3239508

Suppose $X_1, X_2, \ldots$ is a stationary first-order Markov chain on state space $[k] \triangleq \{1, \ldots, k\}$ with unknown statistics. Observing a trajectory $X^n \triangleq (X_1, \ldots, X_n)$, the goal is to predict the next state X_{n+1} by estimating its distribution conditioned on the present data. We use the Kullback-Leibler (KL) divergence as the loss function: For distributions $P = [p_1, \ldots, p_k], Q = [q_1, \ldots, q_k]$, $D(P\|Q) = \sum_{i=1}^k p_i \log \frac{p_i}{q_i}$ if $p_i = 0$ whenever $q_i = 0$ and $D(P\|Q) = \infty$ otherwise. The minimax prediction risk is given by

$$\begin{aligned}\mathsf{Risk}_{k,n} &\triangleq \inf_{\widehat{M}} \sup_{\pi, M} \mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))] \\ &= \inf_{\widehat{M}} \sup_{\pi, M} \sum_{i=1}^k \mathbb{E}[D(M(\cdot|i)\|\widehat{M}(\cdot|i))\mathbf{1}_{\{X_n=i\}}] \quad (1)\end{aligned}$$

where the supremum is taken over all stationary distributions π and transition matrices M (row-stochastic) such that $\pi M = \pi$, the infimum is taken over all estimators $\widehat{M} = \widehat{M}(X_1, \ldots, X_n)$ that are proper Markov kernels (i.e. rows sum to 1), and $M(\cdot|i)$ denotes the ith row of M. Our main objective is to characterize this minimax risk within universal constant factors as a function of n and k.

The prediction problem (1) is distinct from the parameter estimation problem such as estimating the transition matrix [3], [4], [5], [6] or its properties [7], [8], [9], [10] in that the quantity to be estimated (conditional distribution of the next state) depends on the sample path itself. This is precisely what renders the prediction problem closely relevant to natural applications such as autocomplete and text generation. In addition, this formulation allows more flexibility with far less assumptions compared to the estimation framework. For example, if certain state has very small probability under the stationary distribution, consistent estimation of the transition matrix with respect to usual loss function, e.g. squared risk, may not be possible, whereas the prediction problem is unencumbered by such rare states.

In the special case of iid data, the prediction problem reduces to estimating the distribution in KL divergence. In this setting the optimal risk is well understood, which is known to be $\frac{k-1}{2n}(1+o(1))$ when k is fixed and $n \to \infty$ [11] and $\Theta(\frac{k}{n})$ for $k = O(n)$ [12], [13].[1] Typical in parametric models, this rate $\frac{k}{n}$ is commonly referred to the "parametric rate", which leads to a sample complexity that scales proportionally to the

[1]Here and below $\asymp, \lesssim, \gtrsim$ or $\Theta(\cdot), O(\cdot), \Omega(\cdot)$ denote equality and inequalities up to universal multiplicative constants.

number of parameters and inverse proportionally to the desired accuracy.

In the setting of Markov chains, however, the prediction problem is much less understood especially for large state space. Recently the seminal work [1] showed the surprising result that for stationary Markov chains on two states, the optimal prediction risk satisfies

$$\mathsf{Risk}_{2,n} = \Theta\left(\frac{\log\log n}{n}\right), \tag{2}$$

which has a nonparametric rate even when the problem has only two parameters. The follow-up work [2] studied general k-state chains and showed a lower bound of $\Omega(\frac{k\log\log n}{n})$ for uniform (not necessarily stationary) initial distribution; however, the upper bound $O(\frac{k^2\log\log n}{n})$ in [2] relies on implicit assumptions on mixing time such as spectral gap conditions: the proof of the upper bound for prediction (Lemma 7 in the supplement) and for estimation (Lemma 17 of the supplement) is based on Berstein-type concentration results of the empirical transition counts, which depend on spectral gap. The following theorem resolves the optimal risk for k-state Markov chains:

Theorem 1 (Optimal rates without spectral gap): There exists a universal constant $C > 0$ such that for all $3 \le k \le \sqrt{n}/C$,

$$\frac{k^2}{Cn}\log\left(\frac{n}{k^2}\right) \le \mathsf{Risk}_{k,n} \le \frac{Ck^2}{n}\log\left(\frac{n}{k^2}\right). \tag{3}$$

Furthermore, the lower bound continues to hold even if the Markov chain is restricted to be irreducible and reversible.

Remark 1: The optimal prediction risk of $O(\frac{k^2}{n}\log\frac{n}{k^2})$ can be achieved by an average version of the *add-one estimator* (i.e. Laplace's rule of succession).

Given a trajectory $x^n = (x_1, \ldots, x_n)$ of length n, denote the transition counts (with the convention $N_i \equiv N_{ij} \equiv 0$ if $n = 0, 1$)

$$N_i = \sum_{\ell=1}^{n-1} \mathbf{1}_{\{x_\ell = i\}}, \quad N_{ij} = \sum_{\ell=1}^{n-1} \mathbf{1}_{\{x_\ell = i, x_{\ell+1} = j\}}. \tag{4}$$

The add-one estimator for the transition probability $M(j|i)$ is given by

$$\widehat{M}^{+1}_{x^n}(j|i) \triangleq \frac{N_{ij}+1}{N_i + k}, \tag{5}$$

which is an additively smoothed version of the empirical frequency. Finally, the optimal rate in (3) can be achieved by the following estimator $\widehat{M}$ defined as an average of add-one estimators over different sample sizes:

$$\widehat{M}_{x^n}(x_{n+1}|x_n) \triangleq \frac{1}{n}\sum_{t=1}^{n} \widehat{M}^{+1}_{x_{n-t+1}^n}(x_{n+1}|x_n). \tag{6}$$

In other words, we apply the add-one estimator to the most recent t observations $(X_{n-t+1}, \ldots, X_n)$ to predict the next X_{n+1}, then average over $t = 1, \ldots, n$. Such Cesàro-mean-type estimators have been introduced before in the density estimation literature (see, e.g., [14]). It remains open whether the usual add-one estimator (namely, the last term in (6) which uses all the data) or any add-c estimator for constant c achieves the optimal rate. In contrast, for two-state chains the optimal risk (2) is attained by a hybrid strategy [1], applying add-c estimator for $c = \frac{1}{\log n}$ for trajectories with at most one transition and $c = 1$ otherwise. Also note that the estimator in (6) can be computed in $O(nk)$ time. To derive this first note that given any $j \in [k]$ calculating $\widehat{M}^{+1}_{x_1^{n-1}}(j|x_{n-1})$ takes $O(n)$ time and given any $M^{+1}_{x_{n-t+1}^{n-1}}(j|x_{n-1})$ we need $O(1)$ time to calculate $\widehat{M}^{+1}_{x_{n-t+2}^{n-1}}(j|x_{n-1})$. Summing over all j we get the algorithmic complexity upper bound.

Theorem 1 shows that the departure from the parametric rate of $\frac{k^2}{n}$, first discovered in [1] and [2] for binary chains, is even more pronounced for larger state space. As will become clear in the proof, there is some fundamental difference between two-state and three-state chains, resulting in $\mathsf{Risk}_{3,n} = \Theta(\frac{\log n}{n}) \gg \mathsf{Risk}_{2,n} = \Theta(\frac{\log\log n}{n})$. It is instructive to compare the sample complexity for prediction in the iid and Markov model. Denote by d the number of parameters, which is $k-1$ for the iid case and $k(k-1)$ for Markov chains. Define the sample complexity $n^*(d, \epsilon)$ as the smallest sample size n in order to achieve a prescribed prediction risk ϵ. For $\epsilon = O(1)$, we have

$$n^*(d,\epsilon) \asymp \begin{cases} \frac{d}{\epsilon} & \text{iid} \\ \frac{d}{\epsilon}\log\log\frac{1}{\epsilon} & \text{Markov with 2 states} \\ \frac{d}{\epsilon}\log\frac{1}{\epsilon} & \text{Markov with } k \ge 3 \text{ states.} \end{cases} \tag{7}$$

At a high level, the nonparametric rates in the Markov model can be attributed to the memory in the data. On the one hand, Theorem 1 as well as (2) affirm that one can obtain meaningful prediction without imposing any mixing conditions;[2] such decoupling between learning and mixing has also been observed in other problems such as learning linear dynamics [15], [16]. On the other hand, the dependency in the data does lead to a strictly higher sample complexity than that of the iid case; in fact, the lower bound in Theorem 1 is proved by constructing chains with spectral gap as small as $O(\frac{1}{n})$ (see Section III). Thus, it is conceivable that with sufficiently favorable mixing conditions, the prediction risk improves over that of the worst case and, at some point, reaches the parametric rate. To make this precise, we focus on Markov chains with a prescribed spectral gap.

It is well-known that for an irreducible and reversible chain, the transition matrix M has k real eigenvalues satisfying $1 = \lambda_1 \ge \lambda_2 \ge \ldots \lambda_k \ge -1$.

The *absolute spectral gap* of M, defined as

$$\gamma_* \triangleq 1 - \max\{|\lambda_i| : i \ne 1\}, \tag{8}$$

quantifies the memory of the Markov chain. For example, the mixing time is determined by $1/\gamma^*$ (relaxation time) up to logarithmic factors. As extreme cases, the chain which does not move (M is identity) and which is iid (M is rank-one) have spectral gap equal to 0 and 1, respectively. We refer the reader to [17] for more background. Note that the definition of absolute spectral gap requires irreducibility and reversibility, thus we restrict ourselves to this class of Markov chains (it is

[2]To see this, it is helpful to consider the extreme case where the chain does not move at all or is periodic, in which case predicting the next state is in fact easy.

possible to use more general notions such as pseudo spectral gap to quantify the memory of the process, which is beyond the scope of the current paper). Given $\gamma_0 \in (0,1)$, define $\mathcal{M}_k(\gamma_0)$ as the set of transition matrices corresponding to irreducible and reversible chains whose absolute spectral gap exceeds γ_0. Restricting (1) to this subcollection and noticing the stationary distribution here is uniquely determined by M, we define the corresponding minimax risk:

$$\mathsf{Risk}_{k,n}(\gamma_0) \triangleq \inf_{\widehat{M}} \sup_{M \in \mathcal{M}_k(\gamma_0)} \mathbb{E}\left[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))\right] \tag{9}$$

Extending the result (2) of [1], the following theorem characterizes the optimal prediction risk for two-state chains with prescribed spectral gaps (the case $\gamma_0 = 0$ corresponds to the minimax rate in [1] over all binary Markov chains):

Theorem 2 (Spectral gap dependent rates for binary chain): For any $\gamma_0 \in (0,1)$

$$\mathsf{Risk}_{2,n}(\gamma_0) \asymp \frac{1}{n} \max\left\{1, \log\log\left(\min\left\{n, \frac{1}{\gamma_0}\right\}\right)\right\}.$$

Theorem 2 shows that for binary chains, parametric rate $O(\frac{1}{n})$ is achievable if and only if the spectral gap is nonvanishing. While this holds for bounded state space (see Corollary 4 below), for large state space, it turns out that much weaker conditions on the absolute spectral gap suffice to guarantee the parametric rate $O(\frac{k^2}{n})$, achieved by the add-one estimator applied to the entire trajectory. In other words, as long as the spectral gap is not excessively small, the prediction risk in the Markov model behaves in the same way as that of an iid model with equal number of parameters. A similar conclusion has been established previously for the sample complexity of estimating the entropy rate of Markov chains in [9, Theorem 1].

Theorem 3: The add-one estimator in (5) achieves the following risk bound.

(i) For any $k \geq 2$, $\mathsf{Risk}_{k,n}(\gamma_0) \lesssim \frac{k^2}{n}$ provided that $\gamma_0 \gtrsim (\frac{\log k}{k})^{1/4}$.

(ii) In addition, for $k \gtrsim (\log n)^6$, $\mathsf{Risk}_{k,n}(\gamma_0) \lesssim \frac{k^2}{n}$ provided that $\gamma_0 \gtrsim \frac{(\log(n+k))^2}{k}$.

Corollary 4: For any fixed $k \geq 2$, $\mathsf{Risk}_{k,n}(\gamma_0) = O(\frac{1}{n})$ if and only if $\gamma_0 = \Omega(1)$.

Remark 2 (The role of reversibility): The main reason why we require reversibility in Theorems 2 and 3 is to make use of the absolute spectral gap, which might not be well-defined without reversibility. For non-reversible chains, a related notion is the *pseudo spectral gap* (used later in Lemma 31), which is technically more involved. The result of Theorem 3(ii) is still true under the pseudo spectral gap, using the concentration inequality for non-reversible chains [18, Theorem 3.4]. As for other results in Theorems 2 and 3(i), currently we do not know whether they can be extended to the pseudo spectral gap.

Next, we address the optimal prediction risk for higher-order Markov chains:

Theorem 5: There is a constant C_m depending on m such that for any $2 \leq k \leq n^{\frac{1}{m+1}}/C_m$ and constant $m \geq 2$ the minimax prediction rate for m^{th}-order Markov chains with stationary initialization is $\Theta_m\left(\frac{k^{m+1}}{n} \log \frac{n}{k^{m+1}}\right)$.

Notably, for binary states, it turns out that the optimal rate $\Theta\left(\frac{\log\log n}{n}\right)$ for first-order Markov chains determined by [1] is something very special, as we show that for second-order chains the optimal rate is $\Theta\left(\frac{\log n}{n}\right)$.

Finally, we discuss some basic results for the prediction problem with stationary reversible chains when other f-divergences are considered. For the spectral gap independent results, our general proof techniques for the KL-based prediction strongly depend on the reduction to redundancy, which is usually unavailable for other general divergences. As a result, the corresponding risk bounds do not directly follow, and one needs to resort to other techniques. Nevertheless, we show that for the specific case of the squared total variation loss our results can establish a $\Theta_k\left(\frac{\log n}{n}\right)$ when $k \geq 3$ (cf. (136)). On the other hand, we investigate the prediction risk for the stronger loss function given by the Chi-square divergence for the spectral gap dependent results. We show that whenever the absolute spectral gap is non-vanishing in n, k, we can achieve the parametric error rate (cf. Theorem 33).

A. Proof Techniques

The proof of Theorem 1 deviates from existing approaches based on concentration inequalities for Markov chains. For instance, the standard program for analyzing the add-one estimator (5) involves proving concentration of the empirical counts on their population version, namely, $N_i \approx n\pi_i$ and $N_{ij} \approx n\pi_i M(j|i)$, and bounding the risk in the atypical case by concentration inequalities, such as the Chernoff-type bounds in [19] and [18], which have been widely used in recent work on statistical inference with Markov chains [2], [6], [8], [9], [10]. However, these concentration inequalities inevitably depends on the spectral gap of the Markov chain, leading to results which deteriorate as the spectral gap becomes smaller. For two-state chains, results free of the spectral gap are obtained in [1] using explicit joint distribution of the transition counts; this refined analysis, however, is difficult to extend to larger state space as the probability mass function of (N_{ij}) is given by Whittle's formula [20] which takes an unwieldy determinantal form.

Eschewing concentration-based arguments, the crux of our proof of Theorem 1, for both the upper and lower bound, revolves around the following quantity known as *redundancy*:

$$\begin{aligned}\mathsf{Red}_{k,n} &\triangleq \inf_{Q_{X^n}} \sup_{P_{X^n}} D(P_{X^n}\|Q_{X^n}) \\ &= \inf_{Q_{X^n}} \sup_{P_{X^n}} \sum_{x^n} P_{X^n}(x^n) \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}. \end{aligned} \tag{10}$$

Here the supremum is taken over all joint distributions of stationary Markov chains X^n on k states, and the infimum is over all joint distributions Q_{X^n}. A central quantity which measures the minimax regret in universal compression, the redundancy (10) corresponds to minimax cumulative risk (namely, the total prediction risk when the sample size ranges from 1 to n), while (1) is the individual minimax risk at sample

size n – see Section II for a detailed discussion. We prove the following reduction between prediction risk and redundancy:

$$\frac{1}{n}\mathsf{Red}^{\mathsf{sym}}_{k-1,n} - \frac{\log k}{n} \lesssim \mathsf{Risk}_{k,n} \le \frac{1}{n-1}\mathsf{Red}_{k,n} \tag{11}$$

where $\mathsf{Red}^{\mathsf{sym}}$ denotes the redundancy for *symmetric* Markov chains. The upper bound is standard: thanks to the convexity of the loss function and stationarity of the Markov chain, the risk of the Cesàro-mean estimator (6) can be upper bounded using the cumulative risk and, in turn, the redundancy. The proof of the lower bound is more involved. Given a $(k-1)$-state chain, we embed it into a larger state space by introducing a new state, such that with constant probability, the chain starts from and gets stuck at this state for a period time that is approximately uniform in $[n]$, then enters the original chain. Effectively, this scenario is equivalent to a prediction problem on $k-1$ states with a *random* (approximately uniform) sample size, whose prediction risk can then be related to the cumulative risk and redundancy. This intuition can be made precise by considering a Bayesian setting, in which the $(k-1)$-state chain is randomized according to the least favorable prior for (10), and representing the Bayes risk as conditional mutual information and applying the chain rule.

Given the above reduction in (11), it suffices to show both redundancies therein are on the order of $\frac{k^2}{n}\log\frac{n}{k^2}$. The redundancy is upper bounded by *pointwise redundancy*, which replaces the average in (10) by the maximum over all trajectories. Following [21] and [22], we consider an explicit probability assignment defined by add-one smoothing and using combinatorial arguments to bound the pointwise redundancy, shown optimal by information-theoretic arguments.

The optimal spectral gap-dependent rate in Theorem 2 relies on the key observation in [1] that, for binary chains, the dominating contribution to the prediction risk comes from trajectories with a single transition, for which we may apply an add-c estimator with c depending appropriately on the spectral gap. The lower bound is shown using a Bayesian argument similar to that of [2, Theorem 1]. The proof of Theorem 3 relies on more delicate concentration arguments as the spectral gap is allowed to be vanishingly small. Notably, for small k, direct application of existing Bernstein inequalities for Markov chains in [19] and [18] falls short of establishing the parametric rate of $O(\frac{k^2}{n})$ (see Remark 5 in Section IV-B for details); instead, we use a fourth moment bound which turns out to be well suited for analyzing concentration of empirical counts conditional on the terminal state.

For large k, we further improve the spectral gap condition using a simulation argument for Markov chains using independent samples [5], [9]. A key step is a new concentration inequality for $D(P\|\widehat{P}^{+1}_{n,k})$, where $\widehat{P}^{+1}_{n,k}$ is the add-one estimator based on n iid observations of P supported on $[k]$:

$$\mathbb{P}\left(D(P\|\widehat{P}^{+1}_{n,k}) \ge c\cdot\frac{k}{n} + \frac{\mathsf{polylog}(n)\cdot\sqrt{k}}{n}\right) \le \frac{1}{\mathsf{poly}(n)}, \tag{12}$$

for some absolute constant $c>0$. Note that an application of the classical concentration inequality of McDiarmid would result in the second term being $\mathsf{polylog}(n)/\sqrt{n}$, and (12) crucially improves this to $\mathsf{polylog}(n)\cdot\sqrt{k}/n$. Such an improvement has been recently observed by [23], [24], and [25] in studying the similar quantity $D(\widehat{P}_n\|P)$ for the (unsmoothed) empirical distribution $\widehat{P}_n$; however, these results, based on either the method of types or an explicit upper bound of the moment generating function, are not directly applicable to (12) in which the true distribution P appears as the first argument in the KL divergence.

The nonasymptotic spectral gap-independent analysis of the prediction rate for higher-order chains with large alphabets is based on a similar redundancy-based reduction as the first-order chain. However, the nonasymptotic redundancy lower bound for higher-order chains is more challenging. The main technical difficulty is that even if $\{X_t\}_{t=1}^n$ is a reversible m-th order chain, the first-order chain $\{(X_t,\ldots,X_{t+m-1})\}_{t=1}^{n-m+1}$ is usually not reversible. Consequently, as also noted in [26], existing analysis in [27, Sec III] based on simple mixing conditions from [28] leads to suboptimal results on large alphabets. To bypass this issue, we show the pseudo spectral gap [18] of the transition matrix of the first-order chain $\{(X_t,\ldots,X_{t+m-1})\}_{t=1}^{n-m+1}$ is at least a constant. This is accomplished by a careful construction of a prior on m^{th}-order transition matrices with $\Theta\left(k^{m+1}\right)$ degrees of freedom. The high-level idea is to add *laziness* to the first-order chain $\{(X_t,\ldots,X_{t+m-1})\}_{t=1}^{n-m+1}$ so that its pseudo spectral gap is of the same order of the Poincaré's constant [29, Corollary 1.15], a property which requires reversibility without laziness.

B. Related Work

While the exact prediction problem studied in this paper has recently been in focus since [1] and [2], there exists a large body of literature on relate works. As mentioned before some of our proof strategies draws inspiration and results from the study of redundancy in universal compression, its connection to mutual information, as well as the perspective of sequential probability assignment as prediction, dating back to [21], [30], [31], [32], and [33]. Asymptotic characterization of the minimax redundancy for Markov sources, both average and pointwise, were obtained in [27], [34], and [35], in the regime of fixed alphabet size k and large sample size n. Nonasymptotic characterization was obtained in [27] for $n \gg k^2\log k$ and recently extended to $n \asymp k^2$ in [26], which further showed that the behavior of the redundancy remains unchanged even if the Markov chain is very close to being iid in terms of spectral gap $\gamma^* = 1-o(1)$.

The current paper adds to a growing body of literature devoted to statistical learning with dependent data, in particular those dealing with Markov chains. Estimation of the transition matrix [3], [4], [5], [36] and testing the order of Markov chains [7] have been well studied in the large-sample regime. More recently attention has been shifted towards large state space and nonasymptotics. For example, [6] studied the estimation of transition matrix in $\ell_\infty \to \ell_\infty$ induced norm for Markov chains with prescribed pseudo spectral gap and minimum probability mass of the stationary distribution,

and determined sample complexity bounds up to logarithmic factors. Similar results have been obtained for estimating properties of Markov chains, including mixing time and spectral gap [10], entropy rate [8], [9], [37], graph statistics based on random walk [38], as well as identity testing [39], [40], [41], [42]. Most of these results rely on assumptions on the Markov chains such as lower bounds on the spectral gap and the stationary distribution, which afford concentration for sample statistics of Markov chains. In contrast, one of the main contributions in this paper, in particular Theorem 1, is that optimal prediction can be achieved without these assumptions, thereby providing a novel way of tackling these seemingly unavoidable issues. This is ultimately accomplished by information-theoretic and combinatorial techniques from universal compression.

C. Notations and Preliminaries

For $n \in \mathbb{N}$, let $[n] \triangleq \{1, \ldots, n\}$. Denote $x^n = (x_1, \ldots, x_n)$ and $x_t^n = (x_t, \ldots, x_n)$. The distribution of a random variable X is denoted by P_X. In a Bayesian setting, the distribution of a parameter θ is referred to as a prior, denoted by P_θ. We recall the following definitions from information theory [43], [44]. The conditional KL divergence is defined as as an average of KL divergence between conditional distributions:

$$\begin{aligned} D(P_{A|B}\|Q_{A|B}|P_B) &\triangleq \mathbb{E}_{B\sim P_B}[D(P_{A|B}\|Q_{A|B})] \\ &= \int P_B(db) D(P_{A|B=b}\|Q_{A|B=b}). \end{aligned} \tag{13}$$

The mutual information between random variables A and B with joint distribution P_{AB} is $I(A;B) \triangleq D(P_{B|A}\|P_B|P_A)$; similarly, the conditional mutual information is defined as

$$I(A;B|C) \triangleq D(P_{B|A,C}\|P_{B|C}|P_{A,C}).$$

The following variational representation of (conditional) mutual information is well-known

$$\begin{aligned} I(A;B) &= \min_{Q_B} D(P_{B|A}\|Q_B|P_A), \\ I(A;B|C) &= \min_{Q_{B|C}} D(P_{B|A,C}\|Q_{B|C}|P_{AC}). \end{aligned} \tag{14}$$

The entropy of a discrete random variables X is $H(X) \triangleq \sum_x P_X(x) \log \frac{1}{P_X(x)}$.

D. Organization

The rest of the paper is organized as follows. In Section II we describe the general paradigm of minimax redundancy and prediction risk and their dual representation in terms of mutual information. We give a general redundancy-based bound on the prediction risk, which, combined with redundancy bounds for Markov chains, leads to the upper bound in Theorem 1. Section III presents the lower bound construction, starting from three states and then extending to k states. Spectral-gap dependent risk bounds in Theorems 2 and 3 are given in Section IV. Section V presents the results and proofs for m^{th}-order Markov chains. In Section VI we discuss prediction risks assessed by other loss functions than the KL divergence. Section VII discusses the assumptions and implications of our results and related open problems.

II. Two General Paradigms

A. Redundancy, Prediction Risk, and Mutual Information Representation

For $n \in \mathbb{N}$, let $\mathcal{P} = \{P_{X^{n+1}|\theta} : \theta \in \Theta\}$ be a collection of joint distributions parameterized by θ.

1) "Compression": Consider a sample $X^n \triangleq (X_1, \ldots, X_n)$ of size n drawn from $P_{X^n|\theta}$ for some unknown $\theta \in \Theta$. The *redundancy* of a probability assignment (joint distribution) Q_{X^n} is defined as the worst-case KL risk of fitting the joint distribution of X^n, namely

$$\mathsf{Red}(Q_{X^n}) \triangleq \sup_{\theta\in\Theta} D(P_{X^n|\theta}\|Q_{X^n}). \tag{15}$$

Optimizing over Q_{X^n}, the minimax redundancy is defined as

$$\mathsf{Red}_n \triangleq \inf_{Q_{X^n}} \mathsf{Red}_n(Q_{X^n}), \tag{16}$$

where the infimum is over all joint distribution Q_{X^n}. This quantity can be operationalized as the redundancy (i.e. regret) in the setting of universal data compression, that is, the excess number of bits compared to the optimal compressor of X^n that knows θ [44, Chapter 13].

The capacity-redundancy theorem (see [45] for a very general result) provides the following mutual information characterization of (16):

$$\mathsf{Red}_n = \sup_{P_\theta} I(\theta; X^n), \tag{17}$$

where the supremum is over all distributions (priors) P_θ on Θ. In view of the variational representation (14), this result can be interpreted as a minimax theorem:

$$\begin{aligned} \mathsf{Red}_n &= \inf_{Q_{X^n}} \sup_{P_\theta} D(P_{X^n|\theta}\|Q_{X^n}|P_\theta) \\ &= \sup_{P_\theta} \inf_{Q_{X^n}} D(P_{X^n|\theta}\|Q_{X^n}|P_\theta). \end{aligned}$$

Typically, for fixed model size and $n \to \infty$, one expects that $\mathsf{Red}_n = \frac{d}{2}\log n(1+o(1))$, where d is the number of parameters; see [31] for a general theory of this type. Indeed, on a fixed alphabet of size k, we have $\mathsf{Red}_n = \frac{k-1}{2}\log n(1+o(1))$ for iid model [30] and $\mathsf{Red}_n = \frac{k^m(k-1)}{2}\log n(1+o(1))$ for m^{th}-order Markov models [46], with more refined asymptotics shown in [47] and [48]. For large alphabets, nonasymptotic results have also been obtained. For example, for first-order Markov model, $\mathsf{Red}_n \asymp k^2 \log \frac{n}{k^2}$ provided that $n \gtrsim k^2$ [26].

2) "Prediction": Consider the problem of predicting the next unseen data point X_{n+1} based on the observations $X_1, \ldots, X_n$, where $(X_1, \ldots, X_{n+1})$ are jointly distributed as $P_{X^{n+1}|\theta}$ for some unknown $\theta \in \Theta$. Here, an estimator is a distribution (for X_{n+1}) as a function of X^n, which, in turn, can be written as a conditional distribution $Q_{X_{n+1}|X^n}$. As such, its worst-case average risk is

$$\mathsf{Risk}(Q_{X_{n+1}|X^n}) \triangleq \sup_{\theta\in\Theta} D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n|\theta}), \tag{18}$$

where the conditional KL divergence is defined in (13). The minimax prediction risk is then defined as

$$\mathsf{Risk}_n \triangleq \inf_{Q_{X_{n+1}|X^n}} \mathsf{Risk}_n(Q_{X_{n+1}|X^n}), \tag{19}$$

While (16) does not directly correspond to a statistical estimation problem, (19) is exactly the familiar setting of "density estimation", where $Q_{X_{n+1}|X^n}$ is understood as an estimator for the distribution of the unseen X_{n+1} based on the available data $X_1, \ldots, X_n$.

In the Bayesian setting where θ is drawn from a prior P_θ, the Bayes prediction risk coincides with the conditional mutual information as a consequence of the variational representation (14):

$$\inf_{Q_{X_{n+1}|X^n}} \mathbb{E}_\theta[D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n|\theta})] = I(\theta; X_{n+1}|X^n). \tag{20}$$

Furthermore, the Bayes estimator that achieves this infimum takes the following form:

$$Q^{\mathsf{Bayes}}_{X_{n+1}|X^n} = P_{X_{n+1}|X^n} = \frac{\int_\Theta P_{X^{n+1}|\theta}\, dP_\theta}{\int_\Theta P_{X^n|\theta}\, dP_\theta}, \tag{21}$$

known as the Bayes predictive density [30], [49]. These representations play a crucial role in the lower bound proof of Theorem 1. Under appropriate conditions which hold for Markov models (see Lemma 34 in Appendix A), the minimax prediction risk (19) also admits a dual representation analogous to (17):

$$\mathsf{Risk}_n = \sup_{\theta\sim\pi} I(\theta; X_{n+1}|X^n), \tag{22}$$

which, in view of (20), show that the principle of "minimax=worst-case Bayes" continues to hold for prediction problem in Markov models.

The following result relates the redundancy and the prediction risk.

Lemma 6: For any model $\mathcal{P}$,

$$\mathsf{Red}_n \le \sum_{t=0}^{n-1} \mathsf{Risk}_t. \tag{23}$$

In addition, suppose that each $P_{X^n|\theta} \in \mathcal{P}$ is stationary and m^{th}-order Markov. Then for all $n \ge m+1$,

$$\mathsf{Risk}_n \le \mathsf{Risk}_{n-1} \le \frac{\mathsf{Red}_n}{n-m}. \tag{24}$$

Furthermore, for any joint distribution Q_{X^n} factorizing as $Q_{X^n} = \prod_{t=1}^n Q_{X_t|X^{t-1}}$, the prediction risk of the estimator

$$\widetilde{Q}_{X_n|X^{n-1}}(x_n|x^{n-1}) \triangleq \frac{1}{n-m}\sum_{t=m+1}^{n} Q_{X_t|X^{t-1}}(x_n|x_{n-t+1}^{n-1}) \tag{25}$$

is bounded by the redundancy of Q_{X^n} as

$$\mathsf{Risk}(\widetilde{Q}_{X_n|X^{n-1}}) \le \frac{1}{n-m}\mathsf{Red}(Q_{X^n}). \tag{26}$$

Remark 3: Note that the upper bound (23) on redundancy, known as the "estimation-compression inequality" [1], [13], holds without conditions, while the lower bound (24) relies on stationarity and Markovity. For iid data, the estimation-compression inequality is almost an equality; however, this is not the case for Markov chains, as both sides of (23) differ by an unbounded factor of $\Theta(\log\log n)$ for $k=2$ and $\Theta(\log n)$ for fixed $k \ge 3$ – see (2) and Theorem 1. On the other hand, Markov chains with at least three states offers a rare instance where (24) is tight, namely, $\mathsf{Risk}_n \asymp \frac{\mathsf{Red}_n}{n}$ (cf. Lemma 7).

Proof: The upper bound on the redundancy follows from the chain rule of KL divergence:

$$D(P_{X^n|\theta}\|Q_{X^n}) = \sum_{t=1}^n D(P_{X_t|X^{t-1},\theta}\|Q_{X_t|X^{t-1}}|P_{X^{t-1}}). \tag{27}$$

Thus

$$\sup_{\theta\in\Theta} D(P_{X^n|\theta}\|Q_{X^n}) \le \sum_{t=1}^n \sup_{\theta\in\Theta} D(P_{X_t|X^{t-1},\theta}\|Q_{X_t|X^{t-1}}|P_{X^{t-1}}).$$

Minimizing both sides over Q_{X^n} (or equivalently, $Q_{X_t|X^{t-1}}$ for $t=1,\ldots,n$) yields (23).

To upper bound the prediction risk using redundancy, fix any Q_{X^n}, which gives rise to $Q_{X_t|X^{t-1}}$ for $t=1,\ldots,n$. For clarity, let use denote the t^{th} estimator as $\widehat{P}_t(\cdot|x^{t-1}) = Q_{X_t|X^{t-1}=x^{t-1}}$. Consider the estimator $\widetilde{Q}_{X_n|X^{n-1}}$ defined in (25), namely,

$$\widetilde{Q}_{X_n|X^{n-1}=x^{n-1}} \triangleq \frac{1}{n-m}\sum_{t=m+1}^{n} \widehat{P}_t(\cdot|x_{n-t+1},\ldots,x_{n-1}). \tag{28}$$

That is, we apply $\widehat{P}_t$ to the most recent $t-1$ symbols prior to X_n for predicting its distribution, then average over t. We may bound the prediction risk of this estimator by redundancy as follows: Fix $\theta \in \Theta$. To simplify notation, we suppress the dependency of θ and write $P_{X^n|\theta} \equiv P_{X^n}$. Then

$$\begin{aligned}
&D(P_{X_n|X^{n-1}}\|\widetilde{Q}_{X_n|X^{n-1}}|P_{X^{n-1}}) \\
&\overset{(a)}{=} \mathbb{E}\left[D\left(P_{X_n|X_{n-m}^{n-1}}\Big\|\frac{1}{n}\sum_{t=1}^{n}\widehat{P}_t(\cdot|X_{n-t+1}^{n-1})\right)\right] \\
&\overset{(b)}{\le} \frac{1}{n-m}\sum_{t=m+1}^{n}\mathbb{E}\left[D(P_{X_n|X_{n-m}^{n-1}}\|\widehat{P}_t(\cdot|X_{n-t+1}^{n-1}))\right] \\
&\overset{(c)}{=} \frac{1}{n-m}\sum_{t=m+1}^{n}\mathbb{E}\left[D(P_{X_t|X_{t-m}^{t-1}}\|\widehat{P}_t(\cdot|X^{t-1}))\right] \\
&\overset{(d)}{=} \frac{1}{n-m}\sum_{t=m+1}^{n} D(P_{X_t|X^{t-1}}\|Q_{X^t|X^{t-1}}|P_{X^{t-1}}) \\
&\le \frac{1}{n-m}\sum_{t=1}^{n} D(P_{X_t|X^{t-1}}\|Q_{X^t|X^{t-1}}|P_{X^{t-1}}) \\
&\overset{(e)}{=} \frac{1}{n-m} D(P_{X^n}\|Q_{X^n}),
\end{aligned}$$

where (a) uses the m^{th}-order Markovian assumption; (b) is due to the convexity of the KL divergence; (c) uses the crucial fact that for all $t = 1,\ldots,n-1$, $(X_{n-t},\ldots,X_{n-1}) \overset{\text{law}}{=} (X_1,\ldots,X_t)$, thanks to stationarity; (d) follows from substituting $\widehat{P}_t(\cdot|x^{t-1}) = Q_{X_t|X^{t-1}=x^{t-1}}$,

the Markovian assumption $P_{X_t|X_{t-m}^{t-1}} = P_{X_t|X^{t-1}}$, and rewriting the expectation as conditional KL divergence; (e) is by the chain rule (27) of KL divergence. Since the above holds for any $\theta \in \Theta$, the desired (26) follows which implies that $\mathsf{Risk}_{n-1} \le \frac{\mathsf{Red}_n}{n-m}$. Finally, $\mathsf{Risk}_{n-1} \ge \mathsf{Risk}_n$ follows from $\mathbb{E}[D(P_{X_{n+1}|X_n}\|\widehat{P}_n(X_2^n))] = \mathbb{E}[D(P_{X_n|X_{n-1}}\|\widehat{P}_n(X_1^{n-1}))]$, since $(X_2,\dots,X_n)$ and $(X_1,\dots,X_{n-1})$ are equal in law. □

Remark 4: Alternatively, Lemma 6 also follows from the mutual information representation (17) and (22). Indeed, by the chain rule for mutual information,

$$I(\theta; X^n) = \sum_{t=1}^{n} I(\theta; X_t|X^{t-1}), \tag{29}$$

taking the supremum over π (the distribution of θ) on both sides yields (17). For (22), it suffices to show that $I(\theta; X_t|X^{t-1})$ is decreasing in t: for any $\theta \sim \pi$,

$$\begin{aligned} I(\theta; X_{n+1}|X^n) &= \mathbb{E}\log\frac{P_{X_{n+1}|X^n,\theta}}{P_{X_{n+1}|X^n}} \\ &= \mathbb{E}\log\frac{P_{X_{n+1}|X^n,\theta}}{P_{X_{n+1}|X_2^n}} + \underbrace{\mathbb{E}\log\frac{P_{X_{n+1}|X_2^n}}{P_{X_{n+1}|X^n}}}_{-I(X_1;X_{n+1}|X_2^n)}, \end{aligned}$$

and the first term is

$$\begin{aligned} \mathbb{E}\log\frac{P_{X_{n+1}|X^n,\theta}}{P_{X_{n+1}|X_2^n}} &= \mathbb{E}\log\frac{P_{X_{n+1}|X_{n-m+1}^n,\theta}}{P_{X_{n+1}|X_2^n}} \\ = \mathbb{E}\log\frac{P_{X_n|X_{n-m}^{n-1},\theta}}{P_{X_n|X^{n-1}}} &= I(\theta; X_n|X^{n-1}) \end{aligned}$$

where the first and second equalities follow from the m^{th}-order Markovity and stationarity, respectively. Taking supremum over π yields $\mathsf{Risk}_n \le \mathsf{Risk}_{n-1}$. Finally, by the chain rule (29), we have

$$I(\theta; X^n) \ge (n-m) I(\theta; X_n|X^{n-1}),$$

yielding $\mathsf{Risk}_{n-1} \le \frac{\mathsf{Red}_n}{n-m}$.

B. Proof of the Upper Bound Part of Theorem 1

Specializing to first-order stationary Markov chains with k states, we denote the redundancy and prediction risk in (16) and (19) by $\mathsf{Red}_{k,n}$ and $\mathsf{Risk}_{k,n}$, the latter of which is precisely the quantity previously defined in (1). Applying Lemma 6 yields $\mathsf{Risk}_{k,n} \le \frac{1}{n-1}\mathsf{Red}_{k,n}$. To upper bound $\mathsf{Red}_{k,n}$, consider the following probability assignment:

$$Q(x_1,\cdots,x_n) = \frac{1}{k}\prod_{t=1}^{n-1} \widehat{M}_{x^t}^{+1}(x_{t+1}|x_t) \tag{30}$$

where $\widehat{M}^{+1}$ is the add-one estimator defined in (5). This Q factorizes as $Q(x_1) = \frac{1}{k}$ and $Q(x_{t+1}|x^t) = \widehat{M}_{x^t}^{+1}(x_{t+1}|x_t)$. The following lemma bounds the redundancy of Q:

Lemma 7:

$$\mathsf{Red}(Q) \le k(k-1)\left[\log\left(1+\frac{n-1}{k(k-1)}\right)+1\right] + \log k.$$

Combined with Lemma 6, Lemma 7 shows that $\mathsf{Risk}_{k,n} \le C\frac{k^2}{n}\log\frac{n}{k^2}$ for all $k \le \sqrt{n/C}$ and some universal constant C, achieved by the estimator (6), which is obtained by applying the rule (25) to (30).

It remains to show Lemma 7. To do so, we in fact bound the pointwise redundancy of the add-one probability assignment (30) over all (not necessarily stationary) Markov chains on k states. The proof is similar to those of [22, Theorems 6.3 and 6.5], which, in turn, follow the arguments of [21, Sec. III-B].

Proof: We show that for every Markov chain with transition matrix M and initial distribution π, and every trajectory $(x_1,\cdots,x_n)$, it holds that

$$\begin{aligned} &\log\frac{\pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)}{Q(x_1,\cdots,x_n)} \\ &\le k(k-1)\left[\log\left(1+\frac{n}{k(k-1)}\right)+1\right] + \log k \end{aligned} \tag{31}$$

where we abbreviate the add-one estimator $M_{x^t}^{+1}(x_{t+1}|x_t)$ defined in (5) as $\widehat{M}^{+1}(x_{t+1}|x_t)$.

To establish (31), note that $Q(x_1,\cdots,x_n)$ could be equivalently expressed using the empirical counts N_i and N_{ij} in (4) as

$$Q(x_1,\cdots,x_n) = \frac{1}{k}\prod_{i=1}^{k}\frac{\prod_{j=1}^{k} N_{ij}!}{k\cdot(k+1)\cdot\cdots\cdot(N_i+k-1)}.$$

Note that

$$\prod_{t=1}^{n-1} M(x_{t+1}|x_t) = \prod_{i=1}^{k}\prod_{j=1}^{k} M(j|i)^{N_{ij}} \le \prod_{i=1}^{k}\prod_{j=1}^{k}(N_{ij}/N_i)^{N_{ij}},$$

where the inequality follows from $\sum_j \frac{N_{ij}}{N_i}\log\frac{N_{ij}/N_i}{M(j|i)} \ge 0$ for each i, by the nonnegativity of the KL divergence. Therefore, we have

$$\begin{aligned} &\frac{\pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)}{Q(x_1,\cdots,x_n)} \\ &\le k\cdot\prod_{i=1}^{k}\frac{k\cdot(k+1)\cdot\cdots\cdot(N_i+k-1)}{N_i^{N_i}}\prod_{j=1}^{k}\frac{N_{ij}^{N_{ij}}}{N_{ij}!}. \end{aligned} \tag{32}$$

We claim that: for $n_1,\cdots,n_k \in \mathbb{Z}_+$ and $n = \sum_{i=1}^k n_i \in \mathbb{N}$, it holds that

$$\prod_{i=1}^{k}\left(\frac{n_i}{n}\right)^{n_i} \le \frac{\prod_{i=1}^{k} n_i!}{n!}, \tag{33}$$

with the understanding that $(\frac{0}{n})^0 = 0! = 1$. Applying this claim to (32) gives

$$\begin{aligned} &\log\frac{\pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)}{Q(x_1,\cdots,x_n)} \\ &\le \log k + \sum_{i=1}^{k}\log\frac{k\cdot(k+1)\cdot\cdots\cdot(N_i+k-1)}{N_i!} \\ &= \log k + \sum_{i=1}^{k}\sum_{\ell=1}^{N_i}\log\left(1+\frac{k-1}{\ell}\right) \end{aligned}$$

$$\le \log k + \sum_{i=1}^{k} \int_0^{N_i} \log\left(1 + \frac{k-1}{x}\right) dx$$
$$= \log k + \sum_{i=1}^{k} \left((k-1)\log\left(1 + \frac{N_i}{k-1}\right) + N_i \log\left(1 + \frac{k-1}{N_i}\right)\right)$$
$$\overset{(a)}{\le} k(k-1)\log\left(1 + \frac{n-1}{k(k-1)}\right) + k(k-1) + \log k,$$

where (a) follows from the concavity of $x \mapsto \log x$, $\sum_{i=1}^{k} N_i = n-1$, and $\log(1+x) \le x$.

It remains to justify (33), which has a simple information-theoretic proof: Let T denote the collection of sequences x^n in $[k]^n$ whose *type* is given by $(n_1, \ldots, n_k)$. Namely, for each $x^n \in T$, i appears exactly n_i times for each $i \in [k]$. Let $(X_1, \ldots, X_n)$ be drawn uniformly at random from the set T. Then

$$\log \frac{n!}{\prod_{i=1}^{k} n_i!} = H(X_1, \ldots, X_n) \overset{(a)}{\le} \sum_{j=1}^{n} H(X_j) \overset{(b)}{=} n \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n}{n_i},$$

where (a) follows from the fact that the joint entropy is at most the sum of marginal entropies; (b) is because each X_j is distributed as $(\frac{n_1}{n}, \ldots, \frac{n_k}{n})$. □

III. Optimal Rates Without Spectral Gap

In this section, we prove the lower bound part of Theorem 1, which shows the optimality of the average version of the add-one estimator (25). We first describe the lower bound construction for three-state chains, which is subsequently extended to k states.

A. Warmup: An $\Omega(\frac{\log n}{n})$ Lower Bound for Three-State Chains

Theorem 8: $\mathsf{Risk}_{3,n} = \Omega\left(\frac{\log n}{n}\right)$.

To show Theorem 8, consider the following one-parameter family of transition matrices:

$$\mathcal{M} = \left\{ M_p = \begin{bmatrix} 1-\frac{2}{n} & \frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & 1-\frac{1}{n}-p & p \\ \frac{1}{n} & p & 1-\frac{1}{n}-p \end{bmatrix} : 0 \le p \le 1 - \frac{1}{n} \right\}. \tag{34}$$

Note that each transition matrix in $\mathcal{M}$ is symmetric (hence doubly stochastic), whose corresponding chain is reversible with a uniform stationary distribution and spectral gap $\Theta(\frac{1}{n})$; see Fig. 1.

The main idea is as follows. Notice that by design, with constant probability, the trajectory is of the following form: The chain starts and stays at state 1 for t steps, and then transitions into state 2 or 3 and never returns to state 1, where $t = 1, \ldots, n-1$. Since p is the single unknown parameter, the only useful observations are visits to state 2 and 3 and each visit entails one observation about p by flipping a coin with bias roughly p. Thus the effective sample size for estimating p is $n-t-1$ and we expect the best estimation error is of the order of $\frac{1}{n-t}$. However, t is not fixed. In fact, conditioned on the trajectory is of this form, t is roughly uniformly distributed between 1 and $n-1$. As such, we anticipate the estimation error of p is approximately

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-t} = \Theta\left(\frac{\log n}{n}\right).$$

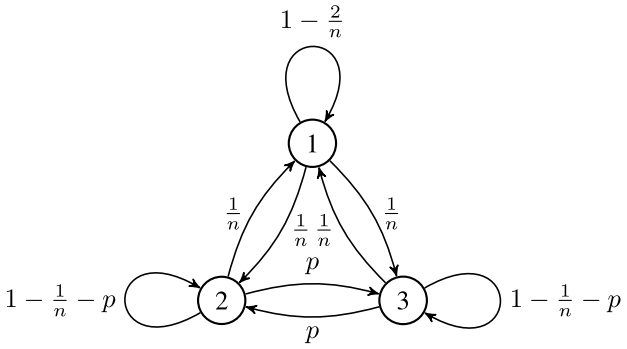

Fig. 1. Lower bound construction for three-state chains.

Intuitively speaking, the construction in Fig. 1 "embeds" a symmetric two-state chain (with states 2 and 3) with unknown parameter p into a space of three states, by adding a "nuisance" state 1, which effectively slows down the exploration of the useful part of the state space, so that in a trajectory of length n, the effective number of observations we get to make about p is roughly uniformly distributed between 1 and n. This explains the extra log factor in Theorem 8, which actually stems from the harmonic sum in $\mathbb{E}[\frac{1}{\mathrm{Uniform}([n])}]$. We will fully explore this embedding idea in Section III-B to deal with larger state space.

Next we make the above intuition rigorous using a Bayesian argument. Let us start by recalling the following well-known lemma.

Lemma 9: Let $q \sim \mathrm{Uniform}(0,1)$. Conditioned on q, let $N \sim \mathrm{Binom}(m, q)$. Then the Bayes estimator of q given N is the "add-one" estimator:

$$\mathbb{E}[q|N] = \frac{N+1}{m+2}$$

and the Bayes risk is given by

$$\mathbb{E}[(q - \mathbb{E}[q|N])^2] = \frac{1}{6(m+2)}.$$

Proof of Theorem 8: Consider the following Bayesian setting: First, we draw p uniformly at random from $[0, 1-\frac{1}{n}]$. Then, we generate the sample path $X^n = (X_1, \ldots, X_n)$ of a stationary (uniform) Markov chain with transition matrix M_p as defined in (34). Define for $t = 1, \ldots, n-1$

$$\mathcal{X}_t = \{x^n : x_1 = \ldots = x_t = 1, x_i \ne 1, i = t+1, \ldots, n\},$$
$$\mathcal{X} = \cup_{t=1}^{n-1} \mathcal{X}_t. \tag{35}$$

Let $\mu(x^n|p) = \mathbb{P}[X = x^n]$. Then for $x^n \in \mathcal{X}_t$ we have

$$\mu(x^n|p) = \frac{1}{3}\left(1-\frac{2}{n}\right)^{t-1}\frac{2}{n}p^{N(x^n)}\left(1-\frac{1}{n}-p\right)^{n-t-1-N(x^n)}, \tag{36}$$

where $N(x^n)$ denotes the number of transitions from state 2 to 3 or from 3 to 2. Then

$$\begin{aligned}
&\mathbb{P}[X^n \in \mathcal{X}_t] \\
&= \frac{1}{3}\left(1-\frac{2}{n}\right)^{t-1} \cdot \\
&\quad \frac{2}{n}\sum_{k=0}^{n-t-1}\binom{n-t-1}{k}p^k\left(1-\frac{1}{n}-p\right)^{n-t-1-k} \\
&= \frac{1}{3}\left(1-\frac{2}{n}\right)^{t-1}\frac{2}{n}\left(1-\frac{1}{n}\right)^{n-t-1} \\
&= \frac{2}{3n}\left(1-\frac{1}{n}\right)^{n-2}\left(1-\frac{1}{n-1}\right)^{t-1}
\end{aligned} \tag{37}$$

and hence

$$\begin{aligned}
&\mathbb{P}[X^n \in \mathcal{X}] \\
&= \sum_{t=1}^{n-1}\mathbb{P}[X^n \in \mathcal{X}_t] \\
&= \frac{2(n-1)}{3n}\left(1-\frac{1}{n}\right)^{n-2}\left(1-\left(1-\frac{1}{n-1}\right)^{n-1}\right) \\
&= \frac{2(1-1/e)}{3e} + o_n(1).
\end{aligned} \tag{38}$$

Consider the Bayes estimator (for estimating p under the mean-squared error)

$$\widehat{p}(x^n) = \mathbb{E}[p|x^n] = \frac{\mathbb{E}[p \cdot \mu(x^n|p)]}{\mathbb{E}[\mu(x^n|p)]}.$$

For $x^n \in \mathcal{X}_t$, using (36) we have (with notation $p \sim \text{Uniform}\left(0, \frac{n-1}{n}\right)$ and $U \sim \text{Uniform}(0,1)$)

$$\begin{aligned}
\widehat{p}(x^n) &= \frac{\mathbb{E}\left[p^{N(x^n)+1}\left(1-\frac{1}{n}-p\right)^{n-t-1-N(x^n)}\right]}{\mathbb{E}\left[p^{N(x^n)}\left(1-\frac{1}{n}-p\right)^{n-t-1-N(x^n)}\right]} \\
&= \frac{n-1}{n}\frac{\mathbb{E}\left[U^{N(x^n)+1}(1-U)^{n-t-1-N(x^n)}\right]}{\mathbb{E}\left[U^{N(x^n)}(1-U)^{n-t-1-N(x^n)}\right]} \\
&= \frac{n-1}{n}\frac{N(x^n)+1}{n-t+1},
\end{aligned}$$

where the last step follows from Lemma 9. From (36), we conclude that conditioned on $X^n \in \mathcal{X}_t$ and on p, $N(X^n) \sim \text{Binom}(n-t-1, q)$, where $q = \frac{p}{1-\frac{1}{n}} \sim \text{Uniform}(0,1)$. Applying Lemma 9 (with $m = n-t-1$ and $N = N(X^n)$), we get

$$\begin{aligned}
&\mathbb{E}[(p-\widehat{p}(X^n))^2|X^n \in \mathcal{X}_t] \\
&= \left(\frac{n-1}{n}\right)^2\mathbb{E}\left[\left(q-\frac{N(x^n)+1}{n-t+1}\right)^2\right] \\
&= \left(\frac{n-1}{n}\right)^2\frac{1}{6(n-t+1)}.
\end{aligned}$$

Finally, note that conditioned on $X^n \in \mathcal{X}$, the probability of $X^n \in \mathcal{X}_t$ is close to uniform. Indeed, from (37) and (38) we get for $t = 1, \ldots, n-1$

$$\begin{aligned}
\mathbb{P}[X^n \in \mathcal{X}_t|\mathcal{X}] &= \frac{1}{n-1}\frac{\left(1-\frac{1}{n-1}\right)^{t-1}}{1-\left(1-\frac{1}{n-1}\right)^{n-1}} \\
&\geq \frac{1}{n-1}\left(\frac{1}{e-1}+o_n(1)\right).
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}[(p-\widehat{p}(X^n))^2\mathbf{1}_{\{X^n\in\mathcal{X}\}}] \\
&= \mathbb{P}[X^n \in \mathcal{X}]\sum_{t=1}^{n-1}\mathbb{E}[(p-\widehat{p}(X^n))^2|X^n \in \mathcal{X}_t]\mathbb{P}[X^n \in \mathcal{X}_t|\mathcal{X}] \\
&\gtrsim \frac{1}{n-1}\sum_{t=1}^{n-1}\frac{1}{n-t+1} = \Theta\left(\frac{\log n}{n}\right).
\end{aligned} \tag{39}$$

Finally, we relate (39) formally to the minimax prediction risk under the KL divergence. Consider any predictor $\widehat{M}(\cdot|i)$ (as a function of the sample path X) for the ith row of M, $i = 1, 2, 3$. By Pinsker inequality, we conclude that

$$D(M(\cdot|2)\|\widehat{M}(\cdot|2)) \geq \frac{1}{2}\|M(\cdot|2)-\widehat{M}(\cdot|2)\|_{\ell_1}^2 \geq \frac{1}{2}(p-\widehat{M}(3|2))^2 \tag{40}$$

and similarly, $D(M(\cdot|3)\|\widehat{M}(\cdot|3)) \geq \frac{1}{2}(p-\widehat{M}(2|3))^2$. Abbreviate $\widehat{M}(3|2) \equiv \widehat{p}_2$ and $\widehat{M}(2|3) \equiv \widehat{p}_3$, both functions of X. Taking expectations over both p and X, the Bayes prediction risk can be bounded as follows

$$\begin{aligned}
&\sum_{i=1}^{3}\mathbb{E}[D(M(\cdot|i)\|\widehat{M}(\cdot|i))\mathbf{1}_{\{X_n=i\}}] \\
&\geq \frac{1}{2}\mathbb{E}[(p-\widehat{p}_2)^2\mathbf{1}_{\{X_n=2\}}+(p-\widehat{p}_3)^2\mathbf{1}_{\{X_n=3\}}] \\
&\geq \frac{1}{2}\sum_{x\in\mathcal{X}}\mu(x^n)\left(\mathbb{E}[(p-\widehat{p}_2)^2|X=x^n]\mathbf{1}_{\{x_n=2\}}\right. \\
&\quad \left.+\mathbb{E}[(p-\widehat{p}_3)^2|X=x^n]\mathbf{1}_{\{x_n=3\}}\right) \\
&\geq \frac{1}{2}\sum_{x^n\in\mathcal{X}}\mu(x^n)\mathbb{E}[(p-\widehat{p}(x^n))^2|X=x^n](\mathbf{1}_{\{x_n=2\}}+\mathbf{1}_{\{x_n=3\}}) \\
&= \frac{1}{2}\sum_{x^n\in\mathcal{X}}\mu(x^n)\mathbb{E}[(p-\widehat{p}(x^n))^2|X=x^n] \\
&= \frac{1}{2}\mathbb{E}[(p-\widehat{p}(X))^2\mathbf{1}_{\{X\in\mathcal{X}\}}] \overset{(39)}{=} \Theta\left(\frac{\log n}{n}\right).
\end{aligned}$$

□

B. k-State Chains

The lower bound construction for 3-state chains in Section III-A can be generalized to k-state chains.

The high-level argument is again to augment a $(k-1)$-state chain into a k-state chain. Specifically, we partition the state space $[k]$ into two sets $\mathcal{S}_1 = \{1\}$ and $\mathcal{S}_2 = \{2, 3, \cdots, k\}$. Consider a k-state Markov chain such that the transition probabilities from $\mathcal{S}_1$ to $\mathcal{S}_2$, and from $\mathcal{S}_2$ to $\mathcal{S}_1$, are both very small (on the order of $\Theta(1/n)$). At state 1, the chain either

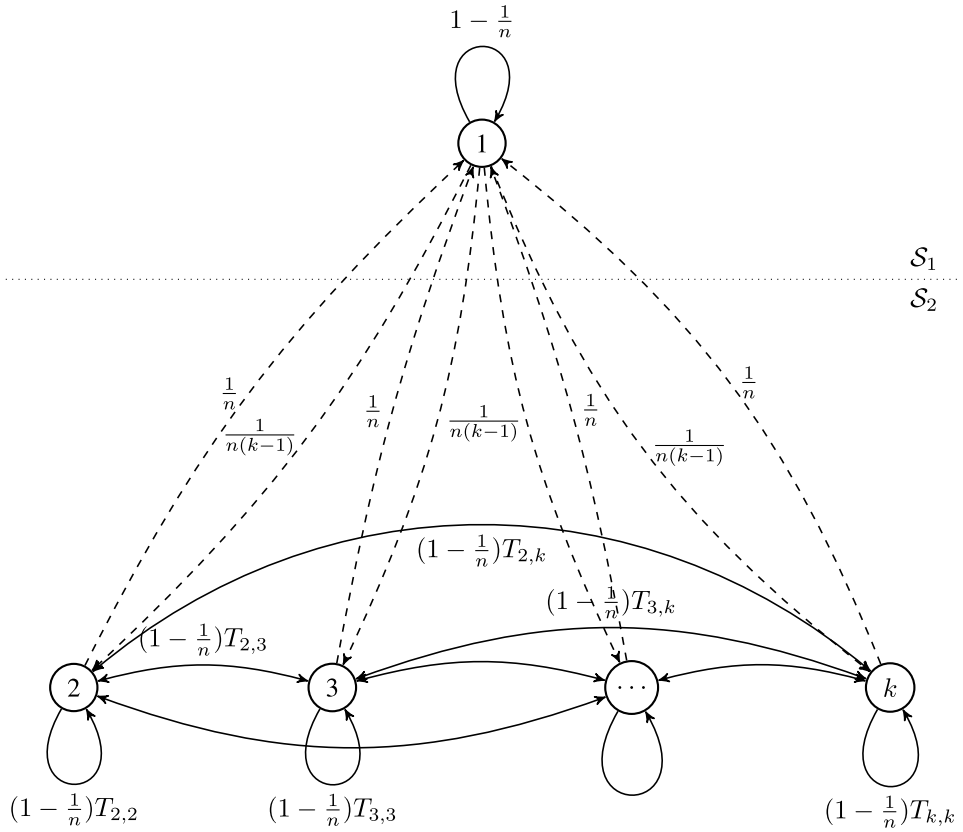

Fig. 2. Lower bound construction for k-state chains. Solid arrows represent transitions within $\mathcal{S}_1$ and $\mathcal{S}_2$, and dashed arrows represent transitions between $\mathcal{S}_1$ and $\mathcal{S}_2$. The double-headed arrows denote transitions in both directions with equal probabilities.

stays at 1 with probability $1-1/n$ or moves to one of the states in $\mathcal{S}_2$ with equal probability $\frac{1}{n(k-1)}$; at each state in $\mathcal{S}_2$, the chain moves to 1 with probability $\frac{1}{n}$; otherwise, within the state subspace $\mathcal{S}_2$, the chain evolves according to some symmetric transition matrix T. (See Fig. 2 in Section III-B.1 for the precise transition diagram.)

The key feature of such a chain is as follows. Let $\mathcal{X}_t$ be the event that $X_1, X_2, \cdots, X_t \in \mathcal{S}_1$ and $X_{t+1}, \cdots, X_n \in \mathcal{S}_2$. For each $t \in [n-1]$, one can show that $\mathbb{P}(\mathcal{X}_t) \geq c/n$ for some absolute constant $c > 0$. Moreover, conditioned on the event $\mathcal{X}_t$, $(X_{t+1}, \ldots, X_n)$ is equal in law to a stationary Markov chain $(Y_1, \cdots, Y_{n-t})$ on state space $\mathcal{S}_2$ with symmetric transition matrix T. It is not hard to show that estimating M and T are nearly equivalent. Consider the Bayesian setting where T is drawn from some prior. We have

$$\begin{aligned}
&\inf_{\widehat{M}} \mathbb{E}_T \left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))|\mathcal{X}_t]\right] \\
&\approx \inf_{\widehat{T}} \mathbb{E}_T \left[\mathbb{E}[D(T(\cdot|Y_{n-t})\|\widehat{T}(\cdot|Y_{n-t}))]\right] \\
&= I(T; Y_{n-t+1}|Y^{n-t}),
\end{aligned}$$

where the last equality follows from the representation (20) of Bayes prediction risk as conditional mutual information. Lower bounding the minimax risk by the Bayes risk, we have

$$\begin{aligned}
&\mathsf{Risk}_{k,n} \\
&\geq \inf_{\widehat{M}} \mathbb{E}_T \left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))]\right] \\
&\geq \inf_{\widehat{M}} \sum_{t=1}^{n-1} \mathbb{E}_M \left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))|\mathcal{X}_t] \cdot \mathbb{P}(\mathcal{X}_t)\right] \\
&\geq \frac{c}{n} \cdot \sum_{t=1}^{n-1} \inf_{\widehat{M}} \mathbb{E}_M \left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))|\mathcal{X}_t]\right] \\
&\approx \frac{c}{n} \cdot \sum_{t=1}^{n-1} I(T; Y_{n-t+1}|Y^{n-t}) \\
&= \frac{c}{n} \cdot (I(T; Y^n) - I(T; Y_1)). \qquad (41)
\end{aligned}$$

Note that $I(T; Y_1) \leq H(Y_1) \leq \log(k-1)$ since Y_1 takes values in $\mathcal{S}_2$. Maximizing the right hand side over the prior P_T and recalling the dual representation for redundancy in (17), the above inequality (41) leads to a risk lower bound of $\mathsf{Risk}_{k,n} \gtrsim \frac{1}{n}(\mathsf{Red}^{\mathsf{sym}}_{k-1,n} - \log k)$, where $\mathsf{Red}^{\mathsf{sym}}_{k-1,n} = \sup I(T; Y_1)$ is the redundancy for *symmetric* Markov chains with $k-1$ states and sample size n. Since symmetric transition matrices still have $\Theta(k^2)$ degrees of freedom, it is expected that $\mathsf{Red}^{\mathsf{sym}}_{k,n} \asymp k^2 \log \frac{n}{k^2}$ for $n \gtrsim k^2$, so that (41) yields the desired lower bound $\mathsf{Risk}_{k,n} = \Omega(\frac{k^2}{n} \log \frac{n}{k^2})$ in Theorem 1.

Next we rigorously carry out the lower bound proof sketched above: In Section III-B.1, we explicitly construct the k-state chain which satisfies the desired properties in Section III-B. In Section III-B.2, we make the steps in (41) precise and bound the Bayes risk from below by an appropriate mutual information. In Section III-B.3, we choose a prior distribution on the transition probabilities and prove a lower bound on the resulting mutual information, thereby completing

the proof of Theorem 1, with the added bonus that the construction is restricted to irreducible and reversible chains.

1) Construction of the k-State Chain: We construct a k-state chain with the following transition probability matrix:

$$M = \begin{bmatrix} 1-\frac{1}{n} & \frac{1}{n(k-1)} & \frac{1}{n(k-1)} & \cdots & \frac{1}{n(k-1)} \\ 1/n & & & & \\ 1/n & & \left(1-\frac{1}{n}\right) T & & \\ \vdots & & & & \\ 1/n & & & & \end{bmatrix}, \quad (42)$$

where $T \in \mathbb{R}^{\mathcal{S}_2 \times \mathcal{S}_2}$ is a symmetric stochastic matrix to be chosen later. The transition diagram of M is shown in Figure 2. One can also verify that the spectral gap of M is $\Theta(\frac{1}{n})$. Let $(X_1, \ldots, X_n)$ be the trajectory of a stationary Markov chain with transition matrix M. We observe the following properties:

(P1) This Markov chain is irreducible and reversible, with stationary distribution $(\frac{1}{2}, \frac{1}{2(k-1)}, \cdots, \frac{1}{2(k-1)})$;

(P2) For $t \in [n-1]$, let $\mathcal{X}_t$ denote the collections of trajectories x^n such that $x_1, x_2, \cdots, x_t \in \mathcal{S}_1$ and $x_{t+1}, \cdots, x_n \in \mathcal{S}_2$. Then

$$\begin{aligned} &\mathbb{P}(X^n \in \mathcal{X}_t) \\ &= \mathbb{P}(X_1 = \cdots = X_t = 1) \cdot \mathbb{P}(X_{t+1} \neq 1 | X_t = 1) \cdot \\ &\quad \prod_{s=t+1}^{n-1} \mathbb{P}(X_{s+1} \neq 1 | X_s \neq 1) \\ &= \frac{1}{2} \cdot \left(1-\frac{1}{n}\right)^{t-1} \cdot \frac{1}{n} \cdot \left(1-\frac{1}{n}\right)^{n-1-t} \\ &\geq \frac{1}{2en}. \end{aligned} \quad (43)$$

Moreover, this probability does not depend of the choice of T;

(P3) Conditioned on the event that $X^n \in \mathcal{X}_t$, the trajectory $(X_{t+1}, \cdots, X_n)$ has the same distribution as a length-$(n-t)$ trajectory of a stationary Markov chain with state space $\mathcal{S}_2 = \{2, 3, \cdots, k\}$ and transition probability T, and the uniform initial distribution. Indeed,

$$\begin{aligned} &\mathbb{P}\left[X_{t+1} = x_{t+1}, \ldots, X_n = x_n | X^n \in \mathcal{X}_t\right] \\ &= \frac{\frac{1}{2} \cdot \left(1-\frac{1}{n}\right)^{t-1} \cdot \frac{1}{n(k-1)} \prod_{s=t+1}^{n-1} M(x_{s+1}|x_s)}{\frac{1}{2} \cdot \left(1-\frac{1}{n}\right)^{t-1} \cdot \frac{1}{n} \cdot \left(1-\frac{1}{n}\right)^{n-1-t}} \\ &= \frac{1}{k-1} \prod_{s=t+1}^{n-1} T(x_{s+1}|x_s). \end{aligned}$$

2) Reducing the Bayes Prediction Risk to Redundancy: Let $\mathcal{M}_{k-1}^{\mathsf{sym}}$ be the collection of all symmetric transition matrices on state space $\mathcal{S}_2 = \{2, \ldots, k\}$. Consider a Bayesian setting where the transition matrix M is constructed in (42) and the submatrix T is drawn from an arbitrary prior on $\mathcal{M}_{k-1}^{\mathsf{sym}}$. The following lemma lower bounds the Bayes prediction risk.

Lemma 10: Conditioned on T, let $Y^n = (Y_1, \ldots, Y_n)$ denote a stationary Markov chain on state space $\mathcal{S}_2$ with transition matrix T and uniform initial distribution. Then

$$\begin{aligned} &\inf_{\widehat{M}} \mathbb{E}_T \left[\mathbb{E}[D(M(\cdot|X_n) \| \widehat{M}(\cdot|X_n))]\right] \\ &\geq \frac{n-1}{2en^2} \left(I(T; Y^n) - \log(k-1)\right). \end{aligned}$$

Lemma 10 is the formal statement of the inequality (41) presented in the proof sketch. Maximizing the lower bound over the prior on T and in view of the mutual information representation (17), we obtain the following corollary.

Corollary 11: Let $\mathsf{Risk}_{k,n}^{\mathsf{rev}}$ denote the minimax prediction risk for stationary irreducible and reversible Markov chains on k states and $\mathsf{Red}_{k,n}^{\mathsf{sym}}$ the redundancy for stationary symmetric Markov chains on k states. Then

$$\mathsf{Risk}_{k,n}^{\mathsf{rev}} \geq \frac{n-1}{2en^2} (\mathsf{Red}_{k-1,n}^{\mathsf{sym}} - \log(k-1)).$$

We make use of the properties (P1)–(P3) in Section III-B.1 to prove Lemma 10.

Proof of Lemma 10: Recall that in the Bayesian setting, we first draw T from some prior on $\mathcal{M}_{k-1}^{\mathsf{sym}}$, then generate the stationary Markov chain $X^n = (X_1, \ldots, X_n)$ with state space $[k]$ and transition matrix M in (42), and $(Y_1, \ldots, Y_n)$ with state space $\mathcal{S}_2 = \{2, \ldots, k\}$ and transition matrix T.

We first relate the Bayes estimator of M and T (given the X and Y chain respectively). For clarity, we spell out the explicit dependence of the estimators on the input trajectory. For each $t \in [n]$, denote by $\widehat{M}_t = \widehat{M}_t(\cdot|x^t)$ the Bayes estimator of $M(\cdot|x_t)$ give $X^t = x^t$, and $\widehat{T}_t(\cdot|y^t)$ the Bayes estimator of $T(\cdot|y_t)$ give $Y^t = y^t$. For each $t = 1, \ldots, n-1$ and for each trajectory $x^n = (1, \ldots, 1, x_{t+1}, \ldots, x_n) \in \mathcal{X}_t$, recalling the form (21) of the Bayes estimator, we have, for each $j \in \mathcal{S}_2$,

$$\begin{aligned} &\widehat{M}_n(j|x^n) \\ &= \frac{\mathbb{P}\left[X^{n+1} = (x^n, j)\right]}{\mathbb{P}\left[X^n = x^n\right]} \\ &= \frac{\mathbb{E}\left[\frac{1}{2} M(1|1)^{t-1} M(x_{t+1}|1) M(x_{t+2}|x_{t+1}) \ldots M(x_n|x_{n-1}) M(j|x_n)\right]}{\mathbb{E}\left[\frac{1}{2} M(1|1)^{t-1} M(x_{t+1}|1) M(x_{t+2}|x_{t+1}) \ldots M(x_n|x_{n-1})\right]} \\ &= \left(1-\frac{1}{n}\right) \frac{\mathbb{E}[T(x_{t+2}|x_{t+1}) \ldots T(x_n|x_{n-1}) T(j|x_n)]}{\mathbb{E}[T(x_{t+2}|x_{t+1}) \ldots T(x_n|x_{n-1})]} \\ &= \left(1-\frac{1}{n}\right) \widehat{T}_{n-t}(j|x_{t+1}^n), \end{aligned}$$

where we used the stationary distribution of X in (P1) and the uniformity of the stationary distribution of Y, neither of which depends on T. Furthermore, by construction in (42), $\widehat{M}_n(1|x^n) = \frac{1}{n}$ is deterministic. In all, we have

$$\widehat{M}_n(\cdot|x^n) = \frac{1}{n}\delta_1 + \left(1-\frac{1}{n}\right) \widehat{T}_{n-t}(\cdot|x_{t+1}^n), \quad x^n \in \mathcal{X}_t. \quad (44)$$

with δ_1 denoting the point mass at state 1, which parallels the fact that

$$M(\cdot|x) = \frac{1}{n}\delta_1 + \left(1-\frac{1}{n}\right) T(\cdot|x), \quad x \in \mathcal{S}_2. \quad (45)$$

By (P2), each event $\{X^n \in \mathcal{X}_t\}$ occurs with probability at least $1/(2en)$, and is independent of T. Therefore,

$$\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X^n))]\right] \geq \frac{1}{2en}\sum_{t=1}^{n-1}\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X^n))|X^n\in\mathcal{X}_t]\right]. \tag{46}$$

By (P3), the conditional joint law of $(T, X_{t+1}, \ldots, X_n)$ on the event $\{X^n \in \mathcal{X}_t\}$ is the same as the joint law of $(T, Y_1, \ldots, Y_{n-t})$. Thus, we may express the Bayes prediction risk in the X chain as

$$\begin{aligned}&\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X^n))|X^n\in\mathcal{X}_t]\right]\\ &\overset{(a)}{=}\left(1-\frac{1}{n}\right)\cdot\mathbb{E}_T\left[\mathbb{E}[D(T(\cdot|Y_{n-t})\|\widehat{T}(\cdot|Y^{n-t}))]\right]\\ &\overset{(b)}{=}\left(1-\frac{1}{n}\right)\cdot I(T;Y_{n-t+1}|Y^{n-t}),\end{aligned} \tag{47}$$

where (a) follows from (44), (45), and the fact that for distributions P, Q supported on $\mathcal{S}_2$, $D(\epsilon\delta_1+(1-\epsilon)P\|\epsilon\delta_1+(1-\epsilon)Q) = (1-\epsilon)D(P\|Q)$; (b) is the mutual information representation (20) of the Bayes prediction risk. Finally, the lemma follows from (46), (47), and the chain rule

$$\begin{aligned}&\sum_{t=1}^{n-1} I(T;Y_{n-t+1}|Y^{n-t})\\ &= I(T;Y^n) - I(T;Y_1) \geq I(T;Y^n) - \log(k-1),\end{aligned}$$

as $I(T;Y_1) \leq H(Y_1) \leq \log(k-1)$. □

3) Prior Construction and Lower Bounding the Mutual Information: In view of Lemma 10, it remains to find a prior on $\mathcal{M}_{k-1}^{\mathsf{sym}}$ for T, such that the mutual information $I(T;Y^n)$ is large. We make use of the connection identified in [21], [27], and [31] between estimation error and mutual information (see also [22, Theorem 7.1] for a self-contained exposition). To lower the mutual information, a key step is to find a good estimator $\widehat{T}(Y^n)$ of T. This is carried out in the following lemma.

Lemma 12: In the setting of Lemma 10, suppose that $T \in \mathcal{M}_k^{\mathsf{sym}}$ with $T_{ij} \in [\frac{1}{2k}, \frac{3}{2k}]$ for all $i, j \in [k]$. Then there is an estimator $\widehat{T}$ based on Y^n such that

$$\mathbb{E}[\|\widehat{T}-T\|_{\mathsf{F}}^2] \leq \frac{16k^2}{n-1},$$

where $\|\widehat{T}-T\|_{\mathsf{F}} = \sqrt{\sum_{ij}(\widehat{T}_{ij}-T_{ij})^2}$ denotes the Frobenius norm.

We show how Lemma 12 leads to the desired lower bound on the mutual information $I(T;Y^n)$. Since $k \geq 3$, we may assume that $k-1 = 2k_0$ is an even integer. Consider the following prior distribution π on T: let $u = (u_{i,j})_{i,j\in[k_0],i\leq j}$ be iid and uniformly distributed in $[1/(4k_0), 3/(4k_0)]$, and $u_{i,j} = u_{j,i}$ for $i > j$. Let the transition matrix T be given by

$$\begin{aligned}T_{2i-1,2j-1} = T_{2i,2j} &= u_{i,j},\\ T_{2i-1,2j} = T_{2i,2j-1} &= \frac{1}{k_0} - u_{i,j},\\ &\forall i,j\in[k].\end{aligned} \tag{48}$$

It is easy to verify that T is symmetric and a stochastic matrix, and each entry of T is supported in the interval $[1/(4k_0), 3/(4k_0)]$. Since $2k_0 = k-1$, the condition of Lemma 12 is fulfilled, so there exist estimators $\widehat{T}(Y^n)$ and $\widehat{u}(Y^n)$ such that

$$\mathbb{E}[\|\widehat{u}(Y^n)-u\|_2^2] \leq \mathbb{E}[\|\widehat{T}(Y^n)-T\|_{\mathsf{F}}^2] \leq \frac{64k_0^2}{n-1}. \tag{49}$$

Here and below, we identify u and $\widehat{u}$ as $\frac{k_0(k_0+1)}{2}$-dimensional vectors.

Let $h(X) = \int -f_X(x)\log f_X(x)dx$ denote the differential entropy of a continuous random vector X with density f_X w.r.t the Lebesgue measure and $h(X|Y) = \int -f_{XY}(xy)\log f_{X|Y}(x|y)dxdy$ the conditional differential entropy (cf. e.g. [44]). Then

$$h(u) = \sum_{i,j\in[k_0],i\leq j} h(u_{i,j}) = -\frac{k_0(k_0+1)}{2}\log(2k_0). \tag{50}$$

Then

$$\begin{aligned}&I(T;Y^n)\\ &\overset{(a)}{=} I(u;Y^n)\\ &\overset{(b)}{\geq} I(u;\widehat{u}(Y^n)) = h(u) - h(u|\widehat{u}(Y^n))\\ &\overset{(c)}{\geq} h(u) - h(u-\widehat{u}(Y^n))\\ &\overset{(d)}{\geq} \frac{k_0(k_0+1)}{4}\log\left(\frac{n-1}{1024\pi e k_0^2}\right) \geq \frac{k^2}{16}\log\left(\frac{n-1}{256\pi e k^2}\right).\end{aligned}$$

where (a) is because u and T are in one-to-one correspondence by (48); (b) follows from the data processing inequality; (c) is because $h(\cdot)$ is translation invariant and concave; (d) follows from the maximum entropy principle [44]: $h(u-\widehat{u}(Y^n)) \leq \frac{k_0(k_0+1)}{4}\log\left(\frac{2\pi e}{k_0(k_0+1)/2}\cdot\mathbb{E}[\|\widehat{u}(Y^n)-u\|_2^2]\right)$, which in turn is bounded by (49). Plugging this lower bound into Lemma 10 completes the lower bound proof of Theorem 1.

Proof of Lemma 12: Since T is symmetric, the stationary distribution is uniform, and there is a one-to-one correspondence between the joint distribution of (Y_1, Y_2) and the transition probabilities. Motivated by this observation, consider the following estimator $\widehat{T}$: for $i, j \in [k]$, let

$$\widehat{T}_{ij} = k\cdot\frac{\sum_{t=1}^n \mathbf{1}_{\{Y_t=i,Y_{t+1}=j\}}}{n-1}.$$

Clearly $\mathbb{E}[\widehat{T}_{ij}] = k\cdot\mathbb{P}(Y_1=i,Y_2=j) = T_{ij}$. The following variance bound is shown in [26, Lemma 7, Lemma 8] using the concentration inequality of [18]:

$$\mathrm{Var}(\widehat{T}_{ij}) \leq k^2\cdot\frac{8T_{ij}k^{-1}}{\gamma_*(T)(n-1)},$$

where $\gamma_*(T)$ is the absolute spectral gap of T defined in (8). Note that $T = k^{-1}\mathbf{J}+\Delta$, where $\mathbf{J}$ is the all-one matrix and each entry of Δ lying in $[-1/(2k), 1/(2k)]$. Thus the spectral radius of Δ is at most $1/2$ and thus $\gamma_*(T) \geq 1/2$. Consequently, we have

$$\mathbb{E}[\|\widehat{T}-T\|_{\mathsf{F}}^2] = \sum_{i,j\in[k]}\mathrm{Var}(\widehat{T}_{ij}) \leq \sum_{i,j\in[k]}\frac{16kT_{ij}}{n-1} = \frac{16k^2}{n-1},$$

completing the proof. □

IV. Spectral Gap-Dependent Risk Bounds

A. Two States

The proof for the spectral gap dependent result in the specific case of two states chain follow closely along the techniques presented in [1]. More specifically, let us consider the sequences in $\{1,2\}^n$ which have exactly one transition, either from 1 to 2 or from 2 to 1

$$\mathcal{S} = \left\{2^{n-\ell}1^{\ell}, 1^{n-\ell}2^{\ell} : 1 \le \ell \le n-1\right\}. \tag{51}$$

Then [1, Lemma 6,7] and its supplemental results establish that the add half estimator $\widehat{M}^{+\frac{1}{2}}(j|i) = \frac{N_{ij}+\frac{1}{2}}{N_i+1}$ achieves $O\left(\frac{1}{n}\right)$ risk upper bound over the trajectory set $\mathcal{S}$. This implies that analyzing the risk bound for $\mathcal{S}$ is sufficient for detecting any non-parametric rate.

For establishing a minimax upper bound on parameter space $\mathcal{M}_2(\gamma_0)$ we consider the following estimator: if $x^n = 2^{n-\ell}1^{\ell}$, we use the estimator

$$\widehat{M}_\ell(2|1) = 1/(\ell \log(1/\gamma_0)), \quad \widehat{M}_\ell(1|1) = 1 - \widehat{M}_\ell(2|1).$$

and symmetrically construct the estimator for $1^{n-\ell}2^{\ell}$. Note that this is similar to the estimator used in [1, Equation 5], which we modified specifically to serve our purpose. For this estimator, we show that the desired minimax rate is achieved for a strictly larger parameter space that consists of binary Markov chains with the spectral gap being at least γ_0, instead of the absolute spectral gap. We analyze the risk individually for every trajectory in $\mathcal{S}$ and add them up to achieve the result.

For establishing the minimax lower bound we consider a Bayesian strategy. We provide here a short description of the analysis when $\log\log(1/\gamma_0) > 0$, the rest of the analysis has been provided later. We use the prior distribution that is uniformly distributed over the following class binary Markov chains for $\alpha = \log(1/\gamma_0)$

$$\mathcal{M} = \left\{M : M(1|2) = \frac{1}{n}, M(2|1) = \frac{1}{\alpha^m}, \right.$$
$$\left. m \in \mathbb{N} \cap \left(\left\lceil \frac{\alpha}{5\log\alpha}\right\rceil, 5\left\lceil \frac{\alpha}{5\log\alpha}\right\rceil\right)\right\}.$$

This prior is similar to the prior used in [2, Section 3], and so is the proof strategy, which we modified specifically to fit our setup. We analyze the Bayes risk for trajectories over the set $\mathcal{S}$. For each $M \in \mathcal{M}$ with $M(2|1) = \frac{1}{\alpha^m}$ we consider trajectories of the form $2^{n-\ell}1^{\ell}$ such that $\left\lceil \frac{\alpha^m}{\log\alpha}\right\rceil \le \ell \le \lfloor \alpha^m \log\alpha \rfloor$. We show that for each such trajectory the expected contribution to the risk is significantly big which sum up to the desired risk lower bound.

We now present the entire proof in detail.

Proof of Theorem 2: To show Theorem 2, let us prove a refined version. In addition to the absolute spectral gap defined in (8), define the spectral gap

$$\gamma \triangleq 1 - \lambda_2 \tag{52}$$

and $\mathcal{M}'_k(\gamma_0)$ the collection of transition matrices whose spectral gap exceeds γ_0. Paralleling $\mathsf{Risk}_{k,n}(\gamma_0)$ defined in (9), define $\mathsf{Risk}'_{k,n}(\gamma_0)$ as the minimax prediction risk restricted to $M \in \mathcal{M}'_k(\gamma_0)$ Since $\gamma \ge \gamma^*$, we have $\mathcal{M}_k(\gamma_0) \subseteq \mathcal{M}'_k(\gamma_0)$ and hence $\mathsf{Risk}'_{k,n}(\gamma_0) \ge \mathsf{Risk}_{k,n}(\gamma_0)$. Nevertheless, the next result shows that for $k = 2$ they have the same rate:

Theorem 13 (Spectral gap dependent rates for binary chain): For any $\gamma_0 \in (0,1)$

$$\mathsf{Risk}_{2,n}(\gamma_0) \asymp \mathsf{Risk}'_{2,n}(\gamma_0)$$
$$\asymp \frac{1}{n}\max\left\{1, \log\log\left(\min\left\{n, \frac{1}{\gamma_0}\right\}\right)\right\}.$$

We first prove the upper bound on $\mathsf{Risk}'_{2,n}$. Note that it is enough to show

$$\mathsf{Risk}'_{2,n}(\gamma_0) \lesssim \frac{\log\log(1/\gamma_0)}{n}, \quad \text{if } n^{-0.9} \le \gamma_0 \le e^{-e^5}. \tag{53}$$

Indeed, for any $\gamma_0 \le n^{-0.9}$, the upper bound $O(\log\log n/n)$ proven in [1], which does not depend on the spectral gap, suffices; for any $\gamma_0 > e^{-e^5}$, by monotonicity we can use the upper bound $\mathsf{Risk}'_{2,n}(e^{-e^5})$.

We now define an estimator that achieves (53). Following [1], consider trajectories with a single transition, namely, $\left\{2^{n-\ell}1^{\ell}, 1^{n-\ell}2^{\ell} : 1 \le \ell \le n-1\right\}$, where $2^{n-\ell}1^{\ell}$ denotes the trajectory $(x_1, \cdots, x_n)$ with $x_1 = \cdots = x_{n-\ell} = 2$ and $x_{n-\ell+1} = \cdots = x_n = 1$. We refer to this type of x^n as *step sequences*. For all non-step sequences x^n, we apply the add-$\frac{1}{2}$ estimator similar to (5), namely

$$\widehat{M}_{x^n}(j|i) = \frac{N_{ij}+\frac{1}{2}}{N_i+1}, \qquad i,j \in \{1,2\},$$

where the empirical counts N_i and N_{ij} are defined in (4); for step sequences of the form $2^{n-\ell}1^{\ell}$, we estimate by

$$\widehat{M}_\ell(2|1) = 1/(\ell \log(1/\gamma_0)), \quad \widehat{M}_\ell(1|1) = 1 - \widehat{M}_\ell(2|1). \tag{54}$$

The other type of step sequences $1^{n-\ell}2^{\ell}$ are dealt with by symmetry.

Due to symmetry it suffices to analyze the risk for sequences ending in 1. The risk of add-$\frac{1}{2}$ estimator for the non-step sequence 1^n is bounded as

$$\mathbb{E}\left[\mathbf{1}_{\{X^n=1^n\}} D(M(\cdot|1)\|\widehat{M}_{1^n}(\cdot|1))\right]$$
$$= P_{X^n}(1^n)\left\{M(2|1)\log\left(\frac{M(2|1)}{1/(2n)}\right)\right.$$
$$\left. + M(1|1)\log\left(\frac{M(1|1)}{(n-\frac{1}{2})/n}\right)\right\}$$
$$\le (1-M(2|1))^{n-1}\left\{2M(2|1)^2 n + \log\left(\frac{n}{n-\frac{1}{2}}\right)\right\}$$
$$\lesssim \frac{1}{n}.$$

where the last step followed by using $(1-x)^{n-1}x^2 \le n^{-2}$ with $x = M(2|1)$ and $\log x \le x-1$. From [1, Lemma 7,8] we have that the total risk of other non-step sequences is bounded from above by $O\left(\frac{1}{n}\right)$ and hence it is enough to analyze the risk for step sequences, and further by symmetry, those in $\left\{2^{n-\ell}1^{\ell} : 1 \le \ell \le n-1\right\}$. The desired upper bound (53) then follows from Lemma 14 next.

Lemma 14: For any $n^{-0.9} \le \gamma_0 \le e^{-e^5}$, $\widehat{M}_\ell(\cdot|1)$ in (54) satisfies

$$\sup_{M\in\mathcal{M}_2'(\gamma_0)} \sum_{\ell=1}^{n-1} \mathbb{E}\left[\mathbf{1}_{\{X^n=2^{n-\ell}1^\ell\}} D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1))\right] \lesssim \frac{\log\log(1/\gamma_0)}{n}.$$

Proof: For each ℓ using $\log\left(\frac{1}{1-x}\right) \le 2x, x \le \frac{1}{2}$ with $x = \frac{1}{\ell\log(1/\gamma_0)}$,

$$\begin{aligned} &D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1)) \\ &= M(1|1)\log\left(\frac{M(1|1)}{1-\frac{1}{\ell\log(1/\gamma_0)}}\right) + M(2|1)\log\left(M(2|1)\ell\log(1/\gamma_0)\right) \\ &\lesssim \frac{1}{\ell\log(1/\gamma_0)} + M(2|1)\log(M(2|1)\ell) + M(2|1)\log\log(1/\gamma_0) \\ &\le \frac{1}{\ell\log(1/\gamma_0)} + M(2|1)\log_+(M(2|1)\ell) + M(2|1)\log\log(1/\gamma_0), \end{aligned} \tag{55}$$

where we define $\log_+(x) = \max\{1, \log x\}$. Recall the following Chebyshev's sum inequality: for $a_1 \le a_2 \le \cdots \le a_n$ and $b_1 \ge b_2 \ge \cdots \ge b_n$, it holds that

$$\sum_{i=1}^n a_i b_i \le \frac{1}{n}\left(\sum_{i=1}^n a_i\right)\left(\sum_{i=1}^n b_i\right).$$

The following inequalities are thus direct corollaries: for $x, y \in [0,1]$,

$$\begin{aligned} &\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1} y(1-y)^{\ell-1} \\ &\le \frac{1}{n-1}\left(\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}\right)\left(\sum_{\ell=1}^{n-1} y(1-y)^{\ell-1}\right) \\ &\le \frac{1}{n-1}, \end{aligned} \tag{56}$$

$$\begin{aligned} &\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1} y(1-y)^{\ell-1}\log_+(\ell y) \\ &\le \frac{1}{n-1}\left(\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}\right) \\ &\quad \left(\sum_{\ell=1}^{n-1} y(1-y)^{\ell-1}\log_+(\ell y)\right) \\ &\le \frac{1}{n-1}\sum_{\ell=1}^{n-1} y(1-y)^{\ell-1}(1+\ell y) \le \frac{2}{n-1}, \end{aligned} \tag{57}$$

where in (57) we need to verify that $\ell \mapsto y(1-y)^{\ell-1}\log_+(\ell y)$ is non-increasing. To verify it, w.l.o.g. we may assume that $(\ell+1)y \ge e$, and therefore

$$\begin{aligned} \frac{y(1-y)^\ell \log_+((\ell+1)y)}{y(1-y)^{\ell-1}\log_+(\ell y)} &= \frac{(1-y)\log((\ell+1)y)}{\log_+(\ell y)} \\ &\le \left(1-\frac{e}{\ell+1}\right)\left(1+\frac{\log(1+1/\ell)}{\log_+(\ell y)}\right) \\ &\le \left(1-\frac{e}{\ell+1}\right)\left(1+\frac{1}{\ell}\right) < 1+\frac{1}{\ell}-\frac{e}{\ell+1} < 1. \end{aligned}$$

Therefore,

$$\begin{aligned} &\sum_{\ell=1}^{n-1} \mathbb{E}\left[\mathbf{1}_{\{X^n=2^{n-\ell}1^\ell\}} D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1))\right] \\ &\le \sum_{\ell=1}^{n-1} M(2|2)^{n-\ell-1} M(1|2) M(1|1)^{\ell-1} \\ &\quad D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1)) \\ &\overset{(a)}{\lesssim} \sum_{\ell=1}^{n-1} M(2|2)^{n-\ell-1} M(1|2) M(1|1)^{\ell-1} \\ &\quad \left(\begin{array}{c} \frac{1}{\ell\log(1/\gamma_0)} + M(2|1)\log_+(M(2|1)\ell) \\ +M(2|1)\log\log(1/\gamma_0) \end{array}\right) \\ &\overset{(b)}{\le} \sum_{\ell=1}^{n-1} \frac{M(2|2)^{n-\ell-1} M(1|2) M(1|1)^{\ell-1}}{\ell\log(1/\gamma_0)} \\ &\quad + \frac{2+\log\log(1/\gamma_0)}{n-1}, \end{aligned} \tag{58}$$

where (a) is due to (55), (b) follows from (56) and (57) applied to $x = M(1|2), y = M(2|1)$. To deal with the remaining sum, we distinguish into two cases. Sticking to the above definitions of x and y, if $y > \gamma_0/2$, then

$$\begin{aligned} &\sum_{\ell=1}^{n-1} \frac{x(1-x)^{n-\ell-1}(1-y)^{\ell-1}}{\ell} \\ &\le \frac{1}{n-1}\left(\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}\right)\left(\sum_{\ell=1}^{n-1}\frac{(1-y)^{\ell-1}}{\ell}\right) \\ &\le \frac{\log(2/\gamma_0)}{n-1}, \end{aligned}$$

where the last step has used that $\sum_{\ell=1}^\infty t^{\ell-1}/\ell = \log(1/(1-t))$ for $|t| < 1$. If $y \le \gamma_0/2$, notice that for two-state chain the spectral gap is given explicitly by $\gamma = M(1|2) + M(2|1) = x + y$, so that the assumption $\gamma \ge \gamma_0$ implies that $x \ge \gamma_0/2$. In this case,

$$\begin{aligned} &\sum_{\ell=1}^{n-1} \frac{x(1-x)^{n-\ell-1}(1-y)^{\ell-1}}{\ell} \\ &\le \sum_{\ell<n/2} (1-x)^{n/2-1} + \sum_{\ell\ge n/2} \frac{x(1-x)^{n-\ell-1}}{n/2} \\ &\le \frac{n}{2} e^{-(n/2-1)\gamma_0} + \frac{2}{n} \lesssim \frac{1}{n}, \end{aligned}$$

thanks to the assumption $\gamma_0 \ge n^{-0.9}$. Therefore, in both cases, the first term in (58) is $O(1/n)$, as desired. □

Next we prove the lower bound on $\mathsf{Risk}_{2,n}$. It is enough to show that $\mathsf{Risk}_{2,n}(\gamma_0) \gtrsim \frac{1}{n}\log\log(1/\gamma_0)$ for $n^{-1} \le \gamma_0 \le e^{-e^5}$. Indeed, for $\gamma_0 \ge e^{-e^5}$, we can apply the result in the iid setting (see, e.g., [11]), in which the absolute spectral gap is 1, to obtain the usual parametric-rate lower bound $\Omega\left(\frac{1}{n}\right)$; for $\gamma_0 < n^{-1}$, we simply bound $\mathsf{Risk}_{2,n}(\gamma_0)$ from below by $\mathsf{Risk}_{2,n}(n^{-1})$. Define

$$\alpha = \log(1/\gamma_0), \quad \beta = \left\lceil \frac{\alpha}{5\log\alpha} \right\rceil, \tag{59}$$

and consider the prior distribution

$$\mathscr{M} = \text{Uniform}(\mathcal{M}),$$
$$\mathcal{M} = \left\{ M : M(1|2) = \frac{1}{n}, M(2|1) = \frac{1}{\alpha^m} : m \in \mathbb{N} \cap (\beta, 5\beta) \right\}. \tag{60}$$

Then the lower bound part of Theorem 2 follows from the next lemma.

Lemma 15: Assume that $n^{-0.9} \le \gamma_0 \le e^{-e^5}$. Then

(i) $\gamma_* > \gamma_0$ for each $M \in \mathcal{M}$;

(ii) the Bayes risk with respect to the prior $\mathscr{M}$ is at least $\Omega\left(\frac{\log\log(1/\gamma_0)}{n}\right)$.

Proof: Part (i) follows by noting that absolute spectral gap for any two states matrix M is $1-|1-M(2|1)-M(1|2)|$ and for any $M \in \mathcal{M}$, $M(2|1) \in \left(\alpha^{-5\beta}, \alpha^{-\beta}\right) \subseteq (\gamma_0, \gamma_0^{1/5}) \subseteq (\gamma_0, 1/2)$ which guarantees $\gamma_* = M(1|2) + M(2|1) > \gamma_0$.

To show part (ii) we lower bound the Bayes risk when the observed trajectory X^n is a step sequence in $\left\{2^{n-\ell}1^\ell : 1 \le \ell \le n-1\right\}$. Our argument closely follows that of [2, Theorem 1]. Since $\gamma_0 \ge n^{-1}$, for each $M \in \mathcal{M}$, the corresponding stationary distribution π satisfies

$$\pi_2 = \frac{M(2|1)}{M(2|1)+M(1|2)} \ge \frac{1}{2}.$$

Denote by $\mathsf{Risk}(\mathscr{M})$ the Bayes risk with respect to the prior $\mathscr{M}$ and by $\widehat{M}_\ell^{\mathsf{B}}(\cdot|1)$ the Bayes estimator for prior $\mathscr{M}$ given $X^n = 2^{n-\ell}1^\ell$. Note that

$$\begin{aligned}&\mathbb{P}\left[X^n = 2^{n-\ell}1^\ell\right]\\ &= \pi_2 \left(1-\frac{1}{n}\right)^{n-\ell-1} \frac{1}{n} M(1|1)^{\ell-1} \ge \frac{1}{2en} M(1|1)^{\ell-1}.\end{aligned} \tag{61}$$

Then

$$\begin{aligned}&\mathsf{Risk}(\mathscr{M})\\ &\ge \mathbb{E}_{M\sim\mathscr{M}}\left[\sum_{\ell=1}^{n-1} \mathbb{E}\left[\mathbf{1}_{\{X^n = 2^{n-\ell}1^\ell\}} D(M(\cdot|1)\|\widehat{M}_\ell^{\mathsf{B}}(\cdot|1))\right]\right]\\ &\ge \mathbb{E}_{M\sim\mathscr{M}}\left[\sum_{\ell=1}^{n-1} \frac{M(1|1)^{\ell-1}}{2en} D(M(\cdot|1)\|\widehat{M}_\ell^{\mathsf{B}}(\cdot|1))\right]\\ &= \frac{1}{2en}\sum_{\ell=1}^{n-1} \mathbb{E}_{M\sim\mathscr{M}}\left[M(1|1)^{\ell-1} D(M(\cdot|1)\|\widehat{M}_\ell^{\mathsf{B}}(\cdot|1))\right].\end{aligned} \tag{62}$$

Recalling the general form of the Bayes estimator in (21) and in view of (61), we get

$$\begin{aligned}\widehat{M}_\ell^{\mathsf{B}}(2|1) &= \frac{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]},\\ \widehat{M}_\ell^{\mathsf{B}}(1|1) &= 1 - \widehat{M}_\ell^{\mathsf{B}}(2|1).\end{aligned} \tag{63}$$

Plugging (63) into (62), and using

$$\begin{aligned}&D((x,1-x)\|(y,1-y))\\ &= x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y} \ge x\max\left\{0, \log\frac{x}{y} - 1\right\},\end{aligned}$$

we arrive at the following lower bound for the Bayes risk:

$$\begin{aligned}&\mathsf{Risk}(\mathscr{M})\\ &\ge \frac{1}{2en}\sum_{\ell=1}^{n-1} \mathbb{E}_{M\sim\mathscr{M}}\left[M(1|1)^{\ell-1}M(2|1)\right.\\ &\quad\left.\cdot\max\left\{0, \log\left(\frac{M(2|1)\cdot\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]}\right) - 1\right\}\right].\end{aligned} \tag{64}$$

Under the prior $\mathscr{M}$, $M(2|1) = 1 - M(1|1) = \alpha^{-m}$ with $\beta \le m \le 5\beta$.

We further lower bound (64) by summing over an appropriate range of ℓ. For any $m \in [\beta, 3\beta]$, define

$$\ell_1(m) = \left\lceil \frac{\alpha^m}{\log\alpha} \right\rceil, \qquad \ell_2(m) = \lfloor \alpha^m \log\alpha \rfloor.$$

Since $\gamma_0 \le e^{-e^5}$, our choice of α ensures that the intervals $\{[\ell_1(m), \ell_2(m)]\}_{\beta\le m\le 3\beta}$ are disjoint. We will establish the following claim: for all $m \in [\beta, 3\beta]$ and $\ell \in [\ell_1(m), \ell_2(m)]$, it holds that

$$\frac{\alpha^{-m}\cdot\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]} \gtrsim \frac{\log(1/\gamma_0)}{\log\log(1/\gamma_0)}. \tag{65}$$

We first complete the proof of the Bayes risk bound assuming (65). Using (64) and (65), we have

$$\begin{aligned}&\mathsf{Risk}(\mathscr{M})\\ &\gtrsim \frac{1}{n}\cdot\frac{1}{4\beta}\sum_{m=\beta}^{3\beta}\sum_{\ell=\ell_1(m)}^{\ell_2(m)} \alpha^{-m}(1-\alpha^{-m})^{\ell-1}\cdot\log\log(1/\gamma_0)\\ &= \frac{\log\log(1/\gamma_0)}{4n\beta}\sum_{m=\beta}^{3\beta}\left\{(1-\alpha^{-m})^{\ell_1(m)-1} - (1-\alpha^{-m})^{\ell_2(m)}\right\}\\ &\overset{(a)}{\gtrsim} \frac{\log\log(1/\gamma_0)}{4n\beta}\sum_{m=\beta}^{3\beta}\left(\left(\frac{1}{4}\right)^{\frac{1}{\log\alpha}} - \left(\frac{1}{e}\right)^{-1+\log\alpha}\right)\\ &\gtrsim \frac{\log\log(1/\gamma_0)}{n},\end{aligned}$$

with (a) following from $\frac{1}{4} \le (1-x)^{\frac{1}{x}} \le \frac{1}{e}$ if $x \le \frac{1}{2}$, and $\alpha^{-m} \le \alpha^{-\beta} \le \gamma_0^{1/5} \le \frac{1}{2}$.

Next we prove the claim (65). Expanding the expectation in (60), we write the LHS of (65) as

$$\frac{\alpha^{-m}\cdot\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]} = \frac{X_\ell + A_\ell + B_\ell}{X_\ell + C_\ell + D_\ell},$$

where

$$X_\ell = \left(1-\alpha^{-m}\right)^\ell, \quad A_\ell = \sum_{j=\beta}^{m-1}\left(1-\alpha^{-j}\right)^\ell,$$
$$B_\ell = \sum_{j=m+1}^{5\beta}\left(1-\alpha^{-j}\right)^\ell,$$
$$C_\ell = \sum_{j=\beta}^{m-1}\left(1-\alpha^{-j}\right)^\ell\alpha^{m-j}, \quad D_\ell = \sum_{j=m+1}^{5\beta}\left(1-\alpha^{-j}\right)^\ell\alpha^{m-j}.$$

We bound each of the terms individually. Clearly, $X_\ell \in (0,1)$ and $A_\ell \geq 0$. Thus it suffices to show that $B_\ell \gtrsim \beta$ and $C_\ell, D_\ell \lesssim 1$, for $m \in [\beta, 3\beta]$ and $\ell_1(m) \leq \ell \leq \ell_2(m)$. Indeed,

- For $j \geq m+1$, we have
$$\left(1-\alpha^{-j}\right)^{\ell} \geq \left(1-\alpha^{-j}\right)^{\ell_2(m)} \overset{(a)}{\geq} (1/4)^{\frac{\ell_2(m)}{\alpha^j}} \geq (1/4)^{\frac{\log\alpha}{\alpha}} \geq 1/4,$$
where in (a) we use the inequality $(1-x)^{1/x} \geq 1/4$ for $x \leq 1/2$. Consequently, $B_\ell \geq \beta/2$;
- For $j \leq m-1$, we have
$$\left(1-\alpha^{-j}\right)^{\ell} \leq \left(1-\alpha^{-j}\right)^{\ell_1(m)} \overset{(b)}{\leq} e^{-\frac{\alpha^{m-j}}{\log\alpha}} = \gamma_0^{\frac{\alpha^{m-j-1}}{\log\alpha}},$$
where (b) follows from $(1-x)^{1/x} \leq 1/e$ and the definition of $\ell_1(m)$. Consequently,
$$C_\ell \leq \gamma_0^{\frac{\alpha}{\log\alpha}} \sum_{j=\beta}^{m-2} \alpha^{m-j} + \alpha\gamma_0^{\frac{1}{\log\alpha}} \leq e^{-\frac{\alpha^2}{\log\alpha}+(2\beta+1)\log\alpha} + e^{\log\alpha-\frac{\alpha}{\log\alpha}} \leq 2,$$
where the last step uses the definition of β in (59);
- $D_\ell \leq \sum_{j=m+1}^{5\beta} \alpha^{m-j} \leq 1$, since $\alpha = \log\frac{1}{\gamma_0} \geq e^5$.

Combining the above bounds completes the proof of (65). □

B. k States

We first present a high-level proof strategy for the add-one estimator to achieve the spectral gap dependent risk bounds when k is large. Given any realization of transition matrix M and add-one estimator $\widehat{M}^{+1}$ the expected risk is

$$\mathbb{E}\left[\sum_{i=1}^{k} \mathbf{1}_{\{X_n=i\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right], \quad \widehat{M}^{+1}(j|i) = \frac{N_{ij}+1}{N_i+k}. \tag{66}$$

As $\widehat{M}^{+1} \geq \frac{1}{n+k}$ we get

$$D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)) \leq \log(n+k), i \in [k]. \tag{67}$$

Let γ be the spectral gap of M. For absolute constants $a_0, a_1, \ldots$ to be chosen later we define

$$\epsilon(m) = \frac{a_0 k}{m} + \frac{a_1(\log n)^3\sqrt{k}}{m}, \quad c_n = a_2\frac{\log n}{n\gamma},$$
$$n_i^{\pm} = n\pi_i \pm a_3 \max\left\{\frac{\log n}{n\gamma}, \sqrt{\frac{\pi_i \log n}{n\gamma}}\right\}, \tag{68}$$

where $i = 1, \ldots, k$. For each $i \in [k]$ we bound $\mathbb{E}\left[\mathbf{1}_{\{X_n=i\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$ based on the following cases.

- $\pi_i \leq c_n$: We bound the expected loss by $\log(n+k)c_n$.
- $\pi_i > c_n$: We further divide this case as
  - $N_i < n_i^-$ or $N_i > n_i^+$: The probability of the event can be made $O(n^{-4})$ by properly choosing a_2, a_3
  - $n_i^- \leq N_i \leq n_i^+$ and $D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)) > \epsilon(N_i)$: The probability of the event can be made $O(\frac{1}{n^3})$ by properly choosing a_0, a_1, a_2, a_3
  - $n_i^- \leq N_i \leq n_i^+$ and $D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)) \leq \epsilon(N_i)$: We bound the expected loss by $\pi_i \max_{n_i^- \leq N_i \leq n_i^+} \epsilon(N_i)$.

Summing up the bounds from the above cases and further summing over $i \in [k]$ we get

$$\mathbb{E}\left[\sum_{i=1}^{k} \mathbf{1}_{\{X_n=i\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right] \lesssim k\log(n+k)\left(c_n + \frac{k}{n^2} + n^{-4}\right) + \sum_{i=1}^{k}\pi_i\epsilon(n_i^-) \lesssim \frac{k^2}{n} + \frac{(\log n)^3 k^{3/2}}{n} + \frac{k(\log(n+k))^2}{n\gamma}. \tag{69}$$

As $k \gtrsim (\log n)^6$ we get the desired result.

As evident from the above, we need a different analysis when $k \leq (\log n)^6$. To this end, we work with the spectral structure and deduce a more detailed concentration guarantees for both N_i and $N_{i,j}$. The concentration results use the following moment bounds which uses the absolute spectral gap. The proofs are deferred to Appendix B.

Lemma 16: Finite reversible and irreducible chains observe the following moment bounds:

(i) $\mathbb{E}\left[(N_{ij} - N_i M(j|i))^2 | X_n = i\right] \lesssim n\pi_i M(j|i)(1 - M(j|i)) + \frac{\sqrt{M(j|i)}}{\gamma_*} + \frac{M(j|i)}{\gamma_*^2}$

(ii) $\mathbb{E}\left[(N_{ij} - N_i M(j|i))^4 | X_n = i\right] \lesssim (n\pi_i M(j|i)(1 - M(j|i)))^2 + \frac{\sqrt{M(j|i)}}{\gamma_*} + \frac{M(j|i)^2}{\gamma_*^4}$

(iii) $\mathbb{E}\left[(N_i - (n-1)\pi_i)^4 | X_n = i\right] \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{1}{\gamma_*^4}$.

When γ_* is high this shows that the moments behave as if for each $i \in [k]$, N_1 is approximately Binomial$(n-1, \pi_i)$ and N_{ij} is approximately Binomial$(N_i, M(j|i))$, which happens in case of iid sampling.

The above results imply that N_i is highly concentrated around $(n-1)\pi_i$ and $N_{ij} - N_i M(j|i)$ is highly concentrated around 0. Using this, we bound $\mathbb{E}\left[\mathbf{1}_{\{X_n=i\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$ for each $i \in [k]$ based on the following cases.

- $N_i \leq \frac{(n-1)\pi_i}{2}$: We bound the expected loss by $\pi_i \log(n\pi_i + k)\mathbb{P}\left[N_i \leq \frac{(n-1)\pi_i}{2} \middle| X_i = i\right]$.
- $N_i > \frac{(n-1)\pi_i}{2}$: We further divide expected loss for each i as $\sum_{j=1}^{k}\Delta_j$ where Δ_j is defined as
$$\Delta_j = \log\left(\frac{M(j|i)(N_i+k)}{N_{ij}+1}\right) + \frac{N_{ij}+1}{N_i+k} - M(j|i). \tag{70}$$
We bound $\mathbb{E}\left[\mathbf{1}_{\left\{X_n=i, N_i > \frac{(n-1)\pi_i}{2}\right\}}\Delta_j\right]$ for each $j \in [k]$.
  - When either of the expected transition counts $n\pi_i n, n\pi_i M(j|i)$ are small we use
$$\Delta_i \leq \frac{(M(i|1)N_1 - N_{1i} + M(i|1)k - 1)^2}{(N_1+k)(N_{1i}+1)}. \tag{71}$$

show that

$$\mathbb{E}\left[\mathbf{1}_{\left\{X_n=i,N_i>\frac{(n-1)\pi_i}{2}\right\}}\Delta_i\right] \lesssim \mathbb{E}\left[\mathbf{1}_{\{X_n=i\}}\frac{(M(j|i)N_i-N_{ij})^2}{n\pi+k}\right]. \tag{72}$$

Then we use Lemma 16 to bound $\mathbb{E}\left[\mathbf{1}_{\{X_n=i\}}(M(j|i)N_i-N_{ij})^2\right]$.

– When both $n\pi_i, n\pi_i M(j|i)$ are big we consider two sub-cases. If $N_{ij} > \frac{(n-1)\pi_i M(j|i)}{4}$ we use (71) and second moment bound to control the expected risk. Otherwise, if $N_{ij} \le \frac{(n-1)\pi_i M(j|i)}{4}$, we get $|N_i M(j|i) - N_{ij}| > \frac{(n-1)\pi_i M(j|i)}{4}$. The probability pertaining to this event can be controlled using a concentration result based on the fourth moment bound in Lemma 16(ii). Furthermore, controlling Δ_i by $\log\left(M(j|i)(N_i+k)\right)$ and using the above probabilistic results we bound the related expected risk.

Combining the above bounds we show that whenever the absolute spectral gap of the underlying chain satisfies $\gamma_* \ge \gamma_0$, we have

$$\mathbb{E}\left[\mathbf{1}_{\{X_n=i\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right] \lesssim \frac{k}{n}\left(1+\sqrt{\frac{\log k}{k\gamma_0^4}}\right). \tag{73}$$

Summing over $i \in [k]$ we get the desired $\mathsf{Risk}_{k,n}(\gamma_0) \lesssim \frac{k^2}{n}\left(1+\sqrt{\frac{\log k}{k\gamma_0^4}}\right)$.

The technical details of the proof of Theorem 3 are given below.

Proof of Theorem 3(i):

Proof: The analysis for $\mathbb{E}\left[\mathbf{1}_{\{X_n=i\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$ is identical for each $i \in [k]$, so we only consider the case $i=1$ and show the bound in (73). We split the risk based on N_1 as

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{X_n=1\}}D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]\\ &=\mathbb{E}\left[\mathbf{1}_{\{A^{\le}\}}D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]\\ &\quad+\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]\end{aligned}$$

where the typical set $A^{>}$ and atypical set $A^{\le}$ are defined as

$$\begin{aligned}A^{\le} &\triangleq \{X_n=1, N_1 \le (n-1)\pi_1/2\},\\ A^{>} &\triangleq \{X_n=1, N_1 > (n-1)\pi_1/2\}.\end{aligned} \tag{74}$$

Analysis for the atypical set $A^{\le}$: For the atypical case, note the following deterministic property of the add-one estimator. Let $\widehat{Q}$ be an add-one estimator with sample size n and alphabet size k of the form $\widehat{Q}_i = \frac{n_i+1}{n+k}$, where $\sum n_i = n$. Since $\widehat{Q}$ is bounded below by $\frac{1}{n+k}$ everywhere, for any distribution P, we have

$$D(P\|\widehat{Q}) \le \log(n+k). \tag{75}$$

Applying this bound on the event $A^{\le}$, we have

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{A^{\le}\}}D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]\\ &\le \log(n\pi_1+k)\,\mathbb{P}\left[X_n=1, N_1 \le (n-1)\pi_1/2\right]\\ &\overset{(a)}{\lesssim} \mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\pi_1\log(n\pi_1+k)\\ &\quad+\mathbf{1}_{\{n\pi_1\gamma_*>10\}}\pi_1\log(n\pi_1+k)\frac{\mathbb{E}\left[(N_1-(n-1)\pi_1)^4\,|X_n=1\right]}{n^4\pi_1^4}\end{aligned} \tag{76}$$

$$\begin{aligned}&\overset{(b)}{\le} \mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\frac{10}{n\gamma_*}\log\left(\frac{10}{\gamma_*}+k\right)\\ &\quad+\mathbf{1}_{\{n\pi_1\gamma_*>10\}}\log(n\pi_1+k)\left(\frac{1}{n^2\pi_1\gamma_*^2}+\frac{1}{n^4\pi_1^3\gamma_*^4}\right)\\ &\overset{(c)}{\lesssim} \frac{1}{n}\left\{\mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\frac{\log(1/\gamma_*)+\log k}{\gamma_*}\right.\\ &\qquad\left.+\mathbf{1}_{\{n\pi_1\gamma_*>10\}}(n\pi_1+\log k)\left(\frac{1}{n\pi_1\gamma_*^2}+\frac{1}{n^3\pi_1^3\gamma_*^4}\right)\right\}\\ &\lesssim \frac{1}{n}\left\{\mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\left(\frac{1}{\gamma_*^2}+\frac{\log k}{\gamma_*}\right)\right.\\ &\qquad\left.+\mathbf{1}_{\{n\pi_1\gamma_*>10\}}\left(\frac{1}{\gamma_*^2}+\frac{\log k}{\gamma_*}\right)\right\} \lesssim \frac{1}{n\gamma_0^2}+\frac{\log k}{n\gamma_0}.\end{aligned} \tag{77}$$

where we got (a) from Markov inequality, (b) from Lemma 16(iii) and (c) using $x+y \le xy, x, y \ge 2$.

Analysis for the typical set $A^{>}$: Next we bound $\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]$. Define

$$\Delta_i = M(i|1)\log\left(\frac{M(i|1)}{\widehat{M}^{+1}(i|1)}\right) - M(i|1) + \widehat{M}^{+1}(i|1).$$

As $D(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)) = \sum_{i=1}^k \Delta_i$ it suffices to bound $\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}\Delta_i\right]$ for each i. For some $r \ge 1$ to be optimized later consider the following cases separately.

- $n\pi_1 \le r$ or $n\pi_1 M(i|1) \le 10$:

Using the fact $y\log(y) - y + 1 \le (y-1)^2$ with $y = \frac{M(i|1)}{\widehat{M}^{+1}(i|1)} = \frac{M(i|1)(N_1+k)}{N_{1i}+1}$ we get

$$\Delta_i \le \frac{(M(i|1)N_1 - N_{1i} + M(i|1)k - 1)^2}{(N_1+k)(N_{1i}+1)}. \tag{78}$$

This implies

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}\Delta_i\right]\\ &\le \mathbb{E}\left[\frac{\mathbf{1}_{\{A^{>}\}}(M(i|1)N_1 - N_{1i} + M(i|1)k - 1)^2}{(N_1+k)(N_{1i}+1)}\right]\\ &\overset{(a)}{\lesssim} \frac{\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}(M(i|1)N_1 - N_{1i})^2\right] + k^2\pi_1 M(i|1)^2 + \pi_1}{n\pi_1+k}\\ &\overset{(b)}{\lesssim} \frac{\pi_1\mathbb{E}\left[(M(i|1)N_1 - N_{1i})^2\,\middle|\,X_n=1\right]}{n\pi_1+k} + \frac{1+rkM(i|1)}{n}\end{aligned} \tag{79}$$

where (a) follows from $N_1 > \frac{(n-1)\pi_1}{2}$ in $A^{>}$ and the fact that $(x+y+z)^2 \le 3(x^2+y^2+z^2)$; (b) uses the assumption that either $n\pi_1 \le r$ or $n\pi_1 M(i|1) \le 10$.

Applying Lemma 16(i) and the fact that $x + x^2 \le 2(1 + x^2)$, continuing the last display we get

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\Delta_i\right]\\ &\lesssim \frac{n\pi_1 M(i|1) + \left(1 + \frac{M(i|1)}{\gamma_*^2}\right)}{n} + \frac{1 + rkM(i|1)}{n}\\ &\lesssim \frac{1 + rkM(i|1)}{n} + \frac{M(i|1)}{n\gamma_0^2}.\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{A^>\}} D(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right]\\ &= \sum_{i=1}^{k}\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\Delta_i\right] \lesssim \frac{rk}{n} + \frac{1}{\gamma_0^2}.\end{aligned} \tag{80}$$

- $\boldsymbol{n\pi_1 > r}$ and $\boldsymbol{n\pi_1 M(i|1) > 10}$:

We decompose $A^>$ based on count of N_{1i} into atypical part $B^\le$ and typical part $B^>$

$$\begin{aligned}B^\le &\triangleq \{X_n = 1, N_1 > (n-1)\pi_1/2,\\ &\qquad N_{1i} \le (n-1)\pi_1 M(i|1)/4\}\\ B^> &\triangleq \{X_n = 1, N_1 > (n-1)\pi_1/2,\\ &\qquad N_{1i} > (n-1)\pi_1 M(i|1)/4\}\end{aligned}$$

and bound each of $\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right]$ and $\mathbb{E}\left[\mathbf{1}_{\{B^>\}}\Delta_i\right]$ separately.

I. **Bound on** $\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\boldsymbol{\Delta_i}\right]$:

Using $\widehat{M}^{+1}(i|1) \ge \frac{1}{N_1+k}$ and $N_{1i} < N_1 M(i|1)/2$ in $B^\le$ we get

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right]\\ &= \mathbb{E}\left[\mathbf{1}_{\{B^\le\}} M(i|1)\log\left(\frac{M(i|1)(N_1+k)}{N_{1i}+1}\right)\right]\\ &\quad + \mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\left(\frac{N_{1i}+1}{N_1+k} - M(i|1)\right)\right]\\ &\le \mathbb{E}\left[\mathbf{1}_{\{B^\le\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right]\\ &\quad + \mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\left(\frac{N_{1i}}{N_1} - M(i|1)\right)\right] + \mathbb{E}\left[\frac{\mathbf{1}_{\{B^\le\}}}{N_1}\right]\\ &\lesssim \mathbb{E}\left[\mathbf{1}_{\{B^\le\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right] + \frac{1}{n}\end{aligned} \tag{81}$$

where the last inequality followed as $\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}/N_1\right] \lesssim \mathbb{P}[X_n = 1]/n\pi_1 = \frac{1}{n}$. Note that for any event B and any function g,

$$\begin{aligned}&\mathbb{E}\left[g(N_1)\mathbf{1}_{\{N_1 \ge t_0, B\}}\right]\\ &= g(t_0)\mathbb{P}[N_1 \ge t_0, B] + \sum_{t=t_0+1}^{n}\left(g(t) - g(t-1)\right)\mathbb{P}[N_1 \ge t, B].\end{aligned}$$

Applying this identity with $t_0 = \lceil (n-1)\pi_1/2\rceil$, we can bound the expectation term in (81) as

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^\le\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right]\\ &= M(i|1)\log\left(M(i|1)(t_0+k)\right)\\ &\quad \mathbb{P}\left[N_1 \ge t_0, N_{1i} \le \frac{n\pi_1 M(i|1)}{4}, X_n = 1\right]\\ &\quad + M(i|1)\sum_{t=t_0+1}^{n-1}\log\left(1 + \frac{1}{t-1+k}\right)\\ &\quad \mathbb{P}\left[N_1 \ge t+1, N_{1i} \le \frac{n\pi_1 M(i|1)}{4}, X_n = 1\right]\\ &\le \pi_1 M(i|1)\log\left(M(i|1)(t_0+k)\right)\\ &\quad \mathbb{P}\left[M(i|1)N_1 - N_{1i} \ge \frac{M(i|1)t_0}{4}\middle| X_n = 1\right]\\ &\quad + \frac{M(i|1)}{n}\sum_{t=t_0+1}^{n-1}\mathbb{P}\left[M(i|1)N_1 - N_{1i} \ge \frac{M(i|1)t}{4}\middle| X_n = 1\right]\end{aligned} \tag{82}$$

where last inequality uses $\log\left(1 + \frac{1}{t-1+k}\right) \le \frac{1}{t} \lesssim \frac{1}{n\pi_1}$ for all $t \ge t_0$. Using Markov inequality $\mathbb{P}[Z > c] \le c^{-4}\mathbb{E}[Z^4]$ for $c > 0$, Lemma 16(ii) and $x + x^4 \le 2(1 + x^4)$ with $x = \sqrt{M(i|1)}/\gamma_*$

$$\begin{aligned}&\mathbb{P}\left[M(i|1)N_1 - N_{1i} \ge \frac{M(i|1)t}{4}\middle| X_n = 1\right]\\ &\lesssim \frac{(n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}}{(tM(i|1))^4}.\end{aligned}$$

In view of above continuing (82) we get

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^\le\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right]\\ &\lesssim \left((n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}\right)\\ &\quad \left(\frac{\pi_1 M(i|1)\log(M(i|1)(n\pi_1+k))}{(n\pi_1 M(i|1))^4}\right.\\ &\quad \left. + \frac{1}{n(M(i|1))^3}\sum_{t=t_0+1}^{n}\frac{1}{t^4}\right)\\ &\lesssim \left(\frac{(n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}}{n}\right)\\ &\quad \left(\frac{\log(n\pi_1 M(i|1) + kM(i|1))}{(n\pi_1 M(i|1))^3} + \frac{1}{(n\pi_1 M(i|1))^3}\right)\\ &\lesssim \frac{1}{n}\left((n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}\right)\\ &\quad \left(\frac{\log(n\pi_1 M(i|1) + kM(i|1))}{(n\pi_1 M(i|1))^3}\right)\\ &\lesssim \frac{1}{n}\left(\frac{\log(n\pi_1 M(i|1) + kM(i|1))}{n\pi_1 M(i|1)}\right.\\ &\quad \left. + \frac{M(i|1)\log(n\pi_1 M(i|1) + k)}{n\pi_1\gamma_*^4(n\pi_1 M(i|1))^2}\right)\\ &\overset{(a)}{\lesssim} \frac{1}{n}\left\{\frac{n\pi_1 M(i|1) + kM(i|1)}{n\pi_1 M(i|1)}\right.\\ &\quad \left. + \frac{M(i|1)\log(n\pi_1 M(i|1))}{n\pi_1\gamma_*^4(n\pi_1 M(i|1))^2}\right.\end{aligned}$$

$$+\frac{M(i|1)\log k}{n\pi_1\gamma_*^4(n\pi_1 M(i|1))^2}\Bigg\}$$
$$\overset{(b)}{\lesssim}\frac{1}{n}\left(1+kM(i|1)+\frac{M(i|1)\log k}{r\gamma_0^4}\right)$$

where (a) followed using $x+y\le xy$ for $x,y\ge 2$ and (b) followed as $n\pi_1\ge r, n\pi_1 M(i|1)\ge 10$ and $\log(n\pi_1 M(i|1))\le n\pi_1 M(i|1)$. In view of (81) this implies

$$\sum_{i=1}^k \mathbb{E}\left[\mathbf{1}_{\{B^{\le}\}}\Delta_i\right]\lesssim\sum_{i=1}^k\frac{1}{n}\left(1+kM(i|1)\left(1+\frac{\log k}{rk\gamma_0^4}\right)\right)$$
$$\lesssim\frac{k}{n}\left(1+\frac{\log k}{rk\gamma_0^4}\right). \tag{83}$$

II. **Bound on $\mathbb{E}\left[\mathbf{1}_{\{B^{>}\}}\boldsymbol{\Delta_i}\right]$:**
Using the inequality (78)

$$\mathbb{E}\left[\mathbf{1}_{\{B^{>}\}}\Delta_i\right]$$
$$\le\mathbb{E}\left[\frac{\mathbf{1}_{\{B^{>}\}}\left(M(i|1)N_1-N_{1i}+M(i|1)k-1\right)^2}{(N_1+k)(N_{1i}+1)}\right]$$
$$\overset{(a)}{\lesssim}\frac{\mathbb{E}\left[\mathbf{1}_{\{B^{>}\}}\left\{(M(i|1)N_1-N_{1i})^2\right\}\right]+\left\{\substack{k^2\pi_1 M(i|1)^2\\+\pi_1}\right\}}{(n\pi_1+k)(n\pi_1 M(i|1)+1)}$$
$$\lesssim\frac{\pi_1\mathbb{E}\left[(M(i|1)N_1-N_{1i})^2\middle|X_n=1\right]}{(n\pi_1+k)(n\pi_1 M(i|1)+1)}+\frac{kM(i|1)}{n}$$

where (a) follows using properties of the set $B^{>}$ along with $(x+y+z)^2\le 3(x^2+y^2+z^2)$. Using Lemma 16(i) we get

$$\mathbb{E}\left[\mathbf{1}_{\{B^{>}\}}\Delta_i\right]\lesssim\frac{n\pi_1 M(i|1)+\left(1+\frac{M(i|1)}{\gamma_*^2}\right)}{n(n\pi_1 M(i|1)+1)}$$
$$+\frac{kM(i|1)}{n}\lesssim\frac{1+kM(i|1)}{n}+\frac{M(i|1)}{n\gamma_0^2}.$$

Summing up the last bound over $i\in[k]$ and using (83) we get for $n\pi_1>r, n\pi_1 M(i|1)>10$

$$\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}D(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right]$$
$$=\sum_{i=1}^k\left[\mathbb{E}\left[\mathbf{1}_{\{B^{\le}\}}\Delta_i\right]+\mathbb{E}\left[\mathbf{1}_{\{B^{>}\}}\Delta_i\right]\right]$$
$$\lesssim\frac{k}{n}\left(1+\frac{1}{k\gamma_0^2}+\frac{\log k}{rk\gamma_0^4}\right).$$

Combining this with (80) we obtain

$$\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}D(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right]$$
$$\lesssim\frac{k}{n}\left(\frac{1}{k\gamma_0^2}+r+\frac{\log k}{rk\gamma_0^4}\right)\lesssim\frac{k}{n}\left(1+\sqrt{\frac{\log k}{k\gamma_0^4}}\right)$$

where we chose $r=10+\sqrt{\frac{\log k}{k\gamma_0^4}}$ for the last inequality. In view of (77) this implies the required bound. □

Remark 5: We explain the subtlety of the concentration bound in Lemma 16 based on fourth moment and why existing Chernoff bound or Chebyshev inequality falls short. For example, the risk bound in (77) relies on bounding the probability that N_1 is atypically small. To this end, one may use the classical Chernoff-type inequality for reversible chains (see [19, Theorem 1.1] or [18, Proposition 3.10 and Theorem 3.3])

$$\mathbb{P}\left[N_1\le\frac{(n-1)\pi_1}{2}|X_1=1\right]\lesssim\frac{e^{-\Theta(n\pi_1\gamma_*)}}{\sqrt{\pi_1}}; \tag{84}$$

in contrast, the fourth moment bound in (76) yields $\mathbb{P}\left[N_1\le(n-1)\pi_1/2|X_1=1\right]=O(\frac{1}{(n\pi_1\gamma_*)^2})$. Although the exponential tail in (84) is much better, the pre-factor $\frac{1}{\sqrt{\pi_1}}$, due to conditioning on the initial state, can lead to a suboptimal result when π_1 is small. (As a concrete example, consider two states with $M(2|1)=\Theta(\frac{1}{n})$ and $M(1|2)=\Theta(1)$. Then $\pi_1=\Theta(\frac{1}{n}), \gamma=\gamma_*\approx\Theta(1)$, and (84) leads to $\mathbb{P}\left[N_1\le(n-1)\pi_1/2, X_n=1\right]=O(\frac{1}{\sqrt{n}})$ as opposed to the desired $O(\frac{1}{n})$.)

In the same context it is also insufficient to use second moment based bound (Chebyshev), which leads to $\mathbb{P}\left[N_1\le(n-1)\pi_1/2|X_1=1\right]=O(\frac{1}{n\pi_1\gamma_*})$. This bound is too loose, which, upon substitution into (76), results in an extra $\log n$ factor in the final risk bound when π_1 and γ_* are large.

Proof of Theorem 3(ii):

Proof: Let $k\ge(\log n)^6$ and $\gamma_0\ge\frac{(\log(n+k))^2}{k}$. We prove a stronger result using the spectral gap as opposed to the absolute spectral gap. Fix M such that $\gamma\ge\gamma_0$. Denote its stationary distribution by π. For absolute constants $\tau>0$ to be chosen later and c_0 as in Lemma 17 below, define, with $i=1,\dots,k$,

$$\epsilon(m)=\frac{2k}{m}+\frac{c_0(\log n)^3\sqrt{k}}{m},\quad c_n=100\tau^2\frac{\log n}{n\gamma},$$
$$n_i^{\pm}=n\pi_i\pm\tau\max\left\{\frac{\log n}{n\gamma},\sqrt{\frac{\pi_i\log n}{n\gamma}}\right\}, \tag{85}$$

Let N_i be the number of visits to state i as in (4). We bound the risk by accounting for the contributions from different ranges of N_i and π_i separately:

$$\mathbb{E}\left[\sum_{i=1}^k\mathbf{1}_{\{X_n=i\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$$
$$=\sum_{i:\pi_i\ge c_n}\mathbb{E}\left[\mathbf{1}_{\{X_n=i,n_i^-\le N_i\le n_i^+\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$$
$$+\sum_{i:\pi_i\ge c_n}\mathbb{E}\left[\mathbf{1}_{\{X_n=i,N_i>n_i^+\text{ or }N_i<n_i^-\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$$
$$+\sum_{i:\pi_i<c_n}\mathbb{E}\left[\mathbf{1}_{\{X_n=i\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$$
$$\le\log(n+k)\sum_{i:\pi_i\ge c_n}\mathbb{P}\left[\substack{D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i))\\>\epsilon(N_i), n_i^-\le N_i\le n_i^+}\right]$$
$$+\sum_{i:\pi_i\ge c_n}\mathbb{E}\left[\mathbf{1}_{\{X_n=i,n_i^-\le N_i\le n_i^+\}}\epsilon(N_i)\right]$$
$$+\log(n+k)\sum_{i:\pi_i\ge c_n}\left[\mathbb{P}\left[N_i\ge n_i^+\right]+\mathbb{P}\left[N_i\le n_i^-\right]\right]$$
$$+\sum_{i:\pi_i\le c_n}\pi_i\log(n+k)$$

$$\lesssim \log(n+k)\sum_{i:\pi_i\ge c_n}\mathbb{P}\left[\begin{array}{c}D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i))>\epsilon(N_i),\\ n_i^-\le N_i\le n_i^+\end{array}\right]$$
$$+\sum_{i:\pi_i\ge c_n}\pi_i\max_{n_i^-\le m\le n_i^+}\epsilon(m)$$
$$+\log(n+k)\sum_{i:\pi_i\ge c_n}\left(\mathbb{P}\left[N_i>n_i^+\right]+\mathbb{P}\left[N_i<n_i^-\right]\right)$$
$$+\frac{k\left(\log(n+k)\right)^2}{n\gamma}. \tag{86}$$

where the first inequality uses the worst-case bound (75) for add-one estimator. We analyze the terms separately as follows.

For the second term, given any i such that $\pi_i\ge c_n$, we have, by definition in (85), $n_i^-\ge 9n\pi_i/10$ and $n_i^+-n_i^-\le n\pi_i/5$, which implies

$$\sum_{i:\pi_i\ge c_n}\pi_i\max_{n_i^-\le m\le n_i^+}\epsilon(m)$$
$$\le\sum_{i:\pi_i\ge c_n}\pi_i\left(\frac{2k}{0.9n\pi_i}+\frac{10}{9}\frac{c_0(\log n)^3\sqrt{k}}{n\pi_i}\right)$$
$$\lesssim\frac{k^2}{n}+\frac{(\log n)^3k^{3/2}}{n}. \tag{87}$$

For the third term, applying [9, Lemma 16] (which, in turn, is based on the Bernstein inequality in [18]), we get $\mathbb{P}\left[N_i>n_i^+\right]+\mathbb{P}\left[N_i<n_i^-\right]\le 2n^{\frac{-\tau^2}{4+10\tau}}$.

To bound the first term in (86), we follow the method in [5] and [9] of representing the sample path of the Markov chain using independent samples generated from $M(\cdot|i)$ which we describe below. Consider a random variable $X_1\sim\pi$ and an array $W=\{W_{i\ell}: i=1,\dots,k \text{ and } \ell=1,2,\dots\}$ of independent random variables, such that X and W are independent and $W_{i\ell}\overset{\text{iid}}{\sim}M(\cdot|i)$ for each i. Starting with generating X_1 from π, at every step $i\ge 2$ we set X_i as the first element in the X_{i-1}-th row of W that has not been sampled yet. Then one can verify that $\{X_1,\dots,X_n\}$ is a Markov chain with initial distribution π and transition matrix M. Furthermore, the transition counts satisfy $N_{ij}=\sum_{\ell=1}^{N_i}\mathbf{1}_{\{W_{i\ell}=j\}}$, where N_i be the number of elements sampled from the ith row of W.

Note the conditioned on $N_i=m$, the random variables $\{W_{i1},\dots,W_{im}\}$ are no longer iid. Instead, we apply a union bound. Note that for each fixed m, the estimator

$$\widehat{M}^{+1}(j|i)=\frac{\sum_{\ell=1}^m\mathbf{1}_{\{W_{i\ell}=j\}}+1}{m+k}\triangleq\widehat{M}_m^{+1}(j|i),\quad j\in[k]$$

is an add-one estimator for $M(j|i)$ based on an iid sample of size m. Lemma 17 below provides a high-probability bound for the add-one estimator in this iid setting. Using this result and the union bound, we have

$$\sum_{i:\pi_i\ge c_n}\mathbb{P}\left[D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i))>\epsilon(N_i), n_i^-\le N_i\le n_i^+\right]$$
$$\le\sum_{i:\pi_i\ge c_n}\left(n_i^+-n_i^-\right)\max_{n_i\le m\le n_i^+}\mathbb{P}\left[\begin{array}{c}D(M(\cdot|i)\|\widehat{M}_m^{+1}(\cdot|i))\\ >\epsilon(m)\end{array}\right]$$
$$\le\sum_{i:\pi_i\ge c_n}\frac{1}{n^2}\le\frac{k}{n^2}$$

where the second inequality applies Lemma 17 with $t=n\ge n_i^+\ge m$ and uses $n_i^+-n_i^-\le n\pi_i/5$ for $\pi_i\ge c_n$.

Combining the above with (87), we continue (86) with $\tau=25$ to get

$$\mathbb{E}\left[\sum_{i=1}^k\mathbf{1}_{\{X_n=i\}}D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right]$$
$$\lesssim\frac{k^2}{n}+\frac{(\log n)^3k^{3/2}}{n}+\frac{k(\log(n+k))^2}{n\gamma}$$

which is $O\left(\frac{k^2}{n}\right)$ whenever $k\ge(\log n)^6$ and $\gamma\ge\frac{(\log(n+k))^2}{k}$. □

Lemma 17 (KL risk bound for add-one estimator): Let $V_1,\dots,V_m\overset{iid}{\sim}Q$ for some distribution $Q=\{Q_i\}_{i=1}^k$ on $[k]$. Consider the add-one estimator $\widehat{Q}^{+1}$ with $\widehat{Q}_i^{+1}=\frac{1}{m+k}(\sum_{j=1}^m\mathbf{1}_{\{V_j=i\}}+1)$. There exists an absolute constant c_0 such that for any $t\ge m$,

$$\mathbb{P}\left[D(Q\|\widehat{Q}^{+1})\ge\frac{2k}{m}+\frac{c_0(\log t)^3\sqrt{k}}{m}\right]\le\frac{1}{t^3}.$$

Proof: Let $\widehat{Q}$ be the empirical estimator $\widehat{Q}_i=\frac{1}{m}\sum_{j=1}^m\mathbf{1}_{\{V_j=i\}}$. Then $\widehat{Q}_i^{+1}=\frac{m\widehat{Q}_i+1}{m+k}$ and hence

$$D(Q\|\widehat{Q}^{+1})$$
$$=\sum_{i=1}^k\left(Q_i\log\frac{Q_i}{\widehat{Q}_i^{+1}}-Q_i+\widehat{Q}_i^{+1}\right)$$
$$=\sum_{i=1}^k\left(Q_i\log\frac{Q_i(m+k)}{m\widehat{Q}_i+1}-Q_i+\frac{m\widehat{Q}_i+1}{m+k}\right)$$
$$=\sum_{i=1}^k\left(Q_i\log\frac{Q_i}{\widehat{Q}_i+\frac{1}{m}}-Q_i+\widehat{Q}_i+\frac{1}{m}\right)$$
$$+\sum_{i=1}^k\left(Q_i\log\frac{m+k}{m}-\frac{k\widehat{Q}_i}{m+k}-\frac{k}{m(m+k)}\right)$$
$$\le\sum_{i=1}^k\left(Q_i\log\frac{Q_i}{\widehat{Q}_i+\frac{1}{m}}-Q_i+\widehat{Q}_i+\frac{1}{m}\right)+\frac{k}{m} \tag{88}$$

with last equality following by $0\le\log\left(\frac{m+k}{m}\right)\le k/m$.

To control the sum in the above display it suffices to consider its Poissonized version. Specifically, we aim to show

$$\mathbb{P}\left[\begin{array}{c}\sum_{i=1}^k\left(Q_i\log\frac{Q_i}{\widehat{Q}_i^{\text{poi}}+\frac{1}{m}}-Q_i+\widehat{Q}_i^{\text{poi}}+\frac{1}{m}\right)\\ >\frac{k}{m}+\frac{c_0(\log t)^3\sqrt{k}}{m}\end{array}\right]\le\frac{1}{t^4} \tag{89}$$

where $m\widehat{Q}_i^{\text{poi}}, i=1,\dots,k$ are distributed independently as $\text{Poi}(mQ_i)$. (Here and below $\text{Poi}(\lambda)$ denotes the Poisson distribution with mean λ.) To see why (89) implies the desired result, letting $w=\frac{k}{m}+\frac{c_0(\log t)^3\sqrt{k}}{m}$ and $Y=\sum_{i=1}^k m\widehat{Q}_i^{\text{poi}}\sim$

$\text{Poi}(m)$, we have

$$\mathbb{P}\left[\sum_{i=1}^{k}\left(Q_i\log\frac{Q_i}{\widehat{Q}_i+\frac{1}{m}}-Q_i+\widehat{Q}_i+\frac{1}{m}\right)>w\right]$$
$$\overset{(a)}{=}\mathbb{P}\left[\sum_{i=1}^{k}\begin{pmatrix}Q_i\log\frac{Q_i}{\widehat{Q}_i^{\mathsf{poi}}+\frac{1}{m}}\\ -Q_i+\widehat{Q}_i^{\mathsf{poi}}+\frac{1}{m}\end{pmatrix}>w\middle|\sum_{i=1}^{k}Q_i^{\mathsf{poi}}=1\right]$$
$$\overset{(b)}{\leq}\frac{1}{t^4\mathbb{P}[Y=m]}=\frac{m!}{t^4e^{-m}m^m}\overset{(c)}{\lesssim}\frac{\sqrt{m}}{t^4}\leq\frac{1}{t^3}. \tag{90}$$

where (a) followed from the fact that conditioned on their sum independent Poisson random variables follow a multinomial distribution; (b) applies (89); (c) follows from Stirling's approximation.

To prove (89) we rely on concentration inequalities for sub-exponential distributions. A random variable X is called sub-exponential with parameters $\sigma^2, b > 0$, denoted as $\mathsf{SE}(\sigma^2, b)$ if

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]\leq e^{\frac{\lambda^2\sigma^2}{2}},\quad \forall|\lambda|<\frac{1}{b}. \tag{91}$$

Sub-exponential random variables satisfy the following properties [50, Sec. 2.1.3]:

- If X is $\mathsf{SE}(\sigma^2, b)$ for any $t > 0$

$$\mathbb{P}\left[|X-\mathbb{E}[X]|\geq v\right]\leq\begin{cases}2e^{-v^2/(2\sigma^2)}, & 0<v\leq\frac{\sigma^2}{b}\\ 2e^{-v/(2b)}, & v>\frac{\sigma^2}{b}.\end{cases} \tag{92}$$

- Bernstein condition: A random variable X is $\mathsf{SE}(\sigma^2, b)$ if it satisfies

$$\mathbb{E}\left[|X-\mathbb{E}[X]|^\ell\right]\leq\frac{\ell!\sigma^2b^{\ell-2}}{2},\ \ell=2,3,\ldots. \tag{93}$$

- If $X_1,\ldots,X_k$ are independent $\mathsf{SE}(\sigma^2, b)$, then $\sum_{i=1}^k X_i$ is $\mathsf{SE}(k\sigma^2, b)$.

Define $X_i = Q_i\log\frac{Q_i}{\widehat{Q}_i^{\mathsf{poi}}+\frac{1}{m}} - Q_i + \widehat{Q}_i^{\mathsf{poi}} + \frac{1}{m}, i\in[k]$. Then Lemma 18 below shows that X_i's are independent $\mathsf{SE}(\sigma^2, b)$ with $\sigma^2 = \frac{c_1(\log m)^4}{m^2}, b = \frac{c_2(\log m)^2}{m}$ for absolute constants c_1, c_2, and hence $\sum_{i=1}^k(X_i-\mathbb{E}[X_i])$ is $\mathsf{SE}(k\sigma^2, b)$. In view of (92) for the choice $c_0 = 8(c_1+c_2)$ this implies

$$\mathbb{P}\left[\sum_{i=1}^{k}(X_i-\mathbb{E}[X_i])\geq c_0\frac{(\log t)^3\sqrt{k}}{m}\right]$$
$$\leq 2e^{-\frac{c_0^2k(\log t)^6}{2m^2\sigma^2}}+2e^{-\frac{c_0\sqrt{k}(\log t)^3}{2mb}}\leq\frac{1}{t^3}. \tag{94}$$

Using $0\leq y\log y-y+1\leq(y-1)^2, y>0$ and $\mathbb{E}\left[\frac{\lambda}{\text{Poi}(\lambda)+1}\right] = \sum_{v=0}^{\infty}\frac{e^{-\lambda}\lambda^{v+1}}{(v+1)!} = 1-e^{-\lambda}$

$$\mathbb{E}\left[\sum_{i=1}^{k}X_i\right]\leq\mathbb{E}\left[\sum_{i=1}^{k}\frac{\left(Q_i-\left(\widehat{Q}_i^{\mathsf{poi}}+\frac{1}{m}\right)\right)^2}{\widehat{Q}_i^{\mathsf{poi}}+\frac{1}{m}}\right]$$
$$=\sum_{i=1}^{k}mQ_i^2\mathbb{E}\left[\frac{1}{m\widehat{Q}_i^{\mathsf{poi}}+1}\right]-1+\frac{k}{m}$$
$$=\sum_{i=1}^{k}Q_i\left(1-e^{-mQ_i}\right)-1+\frac{k}{m}\leq\frac{k}{m}.$$

Combining the above with (94) we get (89) as required. □

Lemma 18: There exist absolute constants c_1, c_2 such that the following holds. For any $p\in(0,1)$ and $nY\sim\text{Poi}(np)$, $X = p\log\frac{p}{Y+\frac{1}{n}} - p + Y + \frac{1}{n}$ is $\mathsf{SE}\left(\frac{c_1(\log n)^4}{n^2}, \frac{c_2(\log n)^2}{n}\right)$.

Proof: Note that X is a non-negative random variable. Since $\mathbb{E}\left[(X-\mathbb{E}[X])^\ell\right]\leq 2^\ell\mathbb{E}\left[X^\ell\right]$, by the Bernstein condition (93), it suffices to show $\mathbb{E}[X^\ell]\leq\left(\frac{c_3\ell(\log n)^2}{n}\right)^\ell, \ell=2,3,\ldots$ for some absolute constant c_3. guarantees the desired sub-exponential behavior. The analysis is divided into following two cases for some absolute constant $c_4\geq 24$.

I. $\boldsymbol{p\geq\frac{c_4\ell\log n}{n}}$:

Using Chernoff bound for Poisson [51, Theorem 3]

$$\mathbb{P}\left[|\text{Poi}(\lambda)-\lambda|>x\right]\leq 2e^{-\frac{x^2}{2(\lambda+x/3)}},\quad\lambda,x>0, \tag{95}$$

we get

$$\mathbb{P}\left[|Y-p|>\sqrt{\frac{c_4\ell p\log n}{4n}}\right]$$
$$\leq 2\exp\left(-\frac{c_4n\ell p\log n}{8np+2\sqrt{c_4n\ell p\log n}}\right)$$
$$\leq 2\exp\left(-\frac{c_4\ell\log n}{8+2\sqrt{c_4\ell\log n/np}}\right)\leq\frac{1}{n^{2\ell}} \tag{96}$$

which implies $p/2\leq Y\leq 2p$ with probability at least $1-n^{-2\ell}$. Since $0\leq X\leq\frac{(Y-p+\frac{1}{n})^2}{Y+\frac{1}{n}}$, we get $\mathbb{E}[X^\ell]\lesssim\frac{\left(\sqrt{c_4\ell p\log n/4n}\right)^{2\ell}}{(p/2)^\ell}+\frac{n^\ell}{n^{2\ell}}\lesssim\left(\frac{c_4\ell\log n}{n}\right)^\ell$.

II. $\boldsymbol{p<\frac{c_4\ell\log n}{n}}$:

- On the event $\{Y>p\}$, we have $X\leq Y+\frac{1}{n}\leq 2Y$, where the last inequality follows because nY takes non-negative integer values. Since $X\geq 0$, we have $X^\ell\mathbf{1}_{\{Y>p\}}\leq(2Y)^\ell\mathbf{1}_{\{Y>p\}}$ for any $\ell\geq 2$.
  Using the Chernoff bound (95), we get $Y\leq\frac{2c_4\ell\log n}{n}$ with probability at least $1-n^{-2\ell}$, which implies

$$\mathbb{E}\left[X^\ell\mathbf{1}_{\{Y\geq p\}}\right]$$
$$\leq\mathbb{E}\left[(2Y)^\ell\mathbf{1}_{\left\{Y>p,Y\leq\frac{2c_4\ell\log n}{n}\right\}}\right]$$
$$+\mathbb{E}\left[(2Y)^\ell\mathbf{1}_{\left\{Y>p,Y>\frac{2c_4\ell\log n}{n}\right\}}\right]$$
$$\leq\left(\frac{4c_4\ell\log n}{n}\right)^\ell+2^\ell\left(\mathbb{E}[Y^{2\ell}]\mathbb{P}\left[Y>\frac{2c_4\ell\log n}{n}\right]\right)^{\frac{1}{2}}$$
$$\leq\left(\frac{c_5\ell\log n}{n}\right)^\ell$$

for absolute constant c_5. Here, the last inequality follows from Cauchy-Schwarz and using the Poisson moment bound [52, Theorem 2.1]:[3] $\mathbb{E}[(nY)^{2\ell}]\leq$

[3]For a result with less precise constants, see also [52, Eq. (1)] based on [53, Corollary 1].

$\left(\frac{2\ell}{\log\left(1+\frac{2\ell}{np}\right)}\right)^{2\ell} \le (c_6\ell\log n)^{2\ell}$ for some absolute constant c_6, with the second inequality applying the assumption $p < \frac{c_4\ell\log n}{n}$.

- As $X\mathbf{1}_{\{Y\le p\}} \le p\log n + \frac{1}{n} \lesssim \frac{\ell(\log n)^2}{n}$, we get $\mathbb{E}\left[X^\ell \mathbf{1}_{\{Y\le p\}}\right] \le \left(\frac{c_7\ell(\log n)^2}{n}\right)^\ell$ for some absolute constant c_7.

□

1) Proof of Corollary 4: We show the following monotonicity result of the prediction risk. In view of this result, Corollary 4 immediately follows from Theorem 2 and Theorem 3(i). Intuitively, the optimal prediction risk is monotonically increasing with the number of states; this, however, does not follow immediately due to the extra assumptions of irreducibility, reversibility, and prescribed spectral gap.

Lemma 19: $\mathsf{Risk}_{k+1,n}(\gamma_0) \ge \mathsf{Risk}_{k,n}(\gamma_0)$ for all $\gamma_0 \in (0,1), k \ge 2$.

Proof: Fix an $M \in \mathcal{M}_k(\gamma_0)$ such that $\gamma_*(M) > \gamma_0$. Denote the stationary distribution π such that $\pi M = \pi$. Fix $\delta \in (0,1)$ and define a transition matrix $\widetilde{M}$ with $k+1$ states as follows:

$$\widetilde{M} = \begin{pmatrix} (1-\delta)M & \delta\mathbf{1} \\ (1-\delta)\pi & \delta \end{pmatrix}$$

One can verify the following:

- $\widetilde{M}$ is irreducible and reversible;
- The stationary distribution for $\widetilde{M}$ is $\widetilde{\pi} = ((1-\delta)\pi, \delta)$
- The absolute spectral gap of $\widetilde{M}$ is $\gamma_*(\widetilde{M}) = (1-\delta)\gamma_*(M)$, so that $\widetilde{M} \in \mathcal{M}_{k+1}(\gamma_0)$ for all sufficiently small δ.
- Let $(X_1,\dots,X_n)$ and $(\widetilde{X}_1,\dots,\widetilde{X}_n)$ be stationary Markov chains with transition matrices M and $\widetilde{M}$, respectively. Then as $\delta \to 0$, $(X_1,\dots,X_n)$ converges to $(\widetilde{X}_1,\dots,\widetilde{X}_n)$ in law, i.e., the joint probability mass function converges pointwise.

Next fix any estimator $\widehat{M}$ for state space $[k+1]$. Note that without loss of generality we can assume $\widehat{M}(j|i) > 0$ for all $i,j \in [k+1]$ for otherwise the KL risk is infinite. Define $\widehat{M}^{\mathrm{trunc}}$ as $\widehat{M}$ without the $k+1$-th row and column, and denote by $\widehat{M}'$ its normalized version, namely, $\widehat{M}'(\cdot|i) = \frac{\widehat{M}^{\mathrm{trunc}}(\cdot|i)}{1-\widehat{M}^{\mathrm{trunc}}(k+1|i)}$ for $i = 1,\dots,k$. Then

$$\begin{aligned}
&\mathbb{E}_{\widetilde{X}^n}\left[D(\widetilde{M}(\cdot|\widetilde{X}_n)\|\widehat{M}(\cdot|\widetilde{X}_n))\right] \\
&\xrightarrow{\delta\to 0} \mathbb{E}_{X^n}\left[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))\right] \\
&\ge \mathbb{E}_{X^n}\left[D(M(\cdot|X_n)\|\widehat{M}'(\cdot|X_n))\right] \\
&\ge \inf_{\widehat{M}} \mathbb{E}_{X^n}\left[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))\right]
\end{aligned}$$

where in the first step we applied the convergence in law of $\widetilde{X}^n$ to X^n and the continuity of $P \mapsto D(P\|Q)$ for fixed componentwise positive Q; in the second step we used the fact that for any sub-probability measure $Q = (q_i)$ and its normalized version $\bar{Q} = Q/\alpha$ with $\alpha = \sum q_i \le 1$, we have $D(P\|Q) = D(P\|\bar{Q}) + \log\frac{1}{\alpha} \ge D(P\|\bar{Q})$. Taking the supremum over $M \in \mathcal{M}_k(\gamma_0)$ on the LHS and the supremum over $\widetilde{M} \in \mathcal{M}_{k+1}(\gamma_0)$ on the RHS, and finally the infimum over $\widehat{M}$ on the LHS, we conclude $\mathsf{Risk}_{k+1,n}(\gamma_0) \ge \mathsf{Risk}_{k,n}(\gamma_0)$. □

V. Higher-Order Markov Chains

In this section we extend our spectral gap-independent results in Section II and III to higher-order chains. While the upper bound is mostly a straightforward application of the risk-redundancy relationship in Lemma 6, the lower bound exhibits some distinction. First, there is no separation between the $k = 2$ case and $k \ge 3$ case: even for the second-order Markov chain with $k = 2$, the number of free parameters is 4 which allows for similar lower bound construction (but more difficult computation) as those in the proof of Theorem 8. Second, to apply the mutual information based arguments, the counterpart of Lemma 12 in higher-order chains no longer follows from the simple mixing condition, as the first-order chain $\{X_t,\cdots,X_{t+m-1}\}_{t=1}^{n-m+1}$ is typically not reversible. To this end, we carefully construct the prior and lower bound a more challenging quantity - the *pseudo spectral gap*. The high-level idea is to add *laziness* to the chain motivated by [29, Corollary 1.15], which in turn exhibits a simple proof in our construction shown in Lemma 31.

A. Basic Setups

We start with some basic definitions for higher-order Markov chains. Let $m \ge 1$. Let $X_1, X_2, \dots$ be an m^{th}-order Markov chain with state space $\mathcal{S}$ and transition matrix $M \in \mathbb{R}^{\mathcal{S}^m \times \mathcal{S}}$ so that $\mathbb{P}\left[X_{t+1} = x_{t+1} | X_{t-m+1}^t = x_{t-m+1}^t\right] = M(x_{t+1}|x_{t-m+1}^t)$ for all $t \ge m$. Clearly, the joint distribution of the process is specified by the transition matrix and the initial distribution, which is a joint distribution for $(X_1,\dots,X_m)$.

A distribution π on $\mathcal{S}^m$ is a *stationary* distribution if $\{X_t : t \ge 1\}$ with $(X_1,\dots,X_m) \sim \pi$ is a stationary process, that is,

$$(X_{i_1+t},\dots,X_{i_n+t}) \overset{\text{law}}{=} (X_{i_1},\dots,X_{i_n}), \quad \forall n, i_1,\dots,i_n,t \in \mathbb{N}. \tag{97}$$

It is clear that (97) is equivalent to $(X_1,\dots,X_m) \overset{\text{law}}{=} (X_2,\dots,X_{m+1})$. In other words, π is the solution to the linear system:

$$\begin{aligned}
&\pi(x_1,\dots,x_m) \\
&= \sum_{x_0\in\mathcal{S}} \pi(x_0,x_1,\dots,x_{m-1}) M(x_m|x_1,\dots,x_{m-1}),
\end{aligned} \tag{98}$$

for all $x_1,\dots,x_m \in \mathcal{S}$. Note that this implies, in particular, that π as a joint distribution of m-tuples itself must satisfy those symmetry properties required by stationarity, such as all marginals being identical, etc.

Next we discuss reversibility. A random process $\{X_t\}$ is *reversible* if for any n,

$$X^n \overset{\text{law}}{=} \overline{X^n}, \tag{99}$$

where $\overline{X^n} \triangleq (X_n,\dots,X_1)$ denotes the time reversal of $X^n = (X_1,\dots,X_n)$. Note that a reversible m^{th}-order Markov chain must be stationary. Indeed,

$$(X_2,\dots,X_{m+1}) \overset{\text{law}}{=} (X_m,\dots,X_1) \overset{\text{law}}{=} (X_1,\dots,X_m), \tag{100}$$

where the first equality follows from $(X_1,\ldots,X_{m+1})\overset{\text{law}}{=}(X_{m+1},\ldots,X_1)$. The following lemma gives a characterization for reversibility:

Lemma 20: An m^{th}-order stationary Markov chain is reversible if and only if (99) holds for $n=m+1$, namely

$$\begin{aligned}&\pi(x_1,\ldots,x_m)M(x_{m+1}|x_1,\ldots,x_m)\\&=\pi(x_{m+1},\ldots,x_2)M(x_1|x_{m+1},\ldots,x_2),\end{aligned}\tag{101}$$

for all $x_1,\ldots,x_{m+1}\in\mathcal{S}$

Proof: First, we show that (99) for $n=m+1$ implies that for $n\le m$. Indeed,

$$(X_1,\ldots,X_n)\overset{\text{law}}{=}(X_{m+1},\ldots,X_{m-n+2})\overset{\text{law}}{=}(X_n,\ldots,X_1)\tag{102}$$

where the first equality follows from $(X_1,\ldots,X_{m+1})\overset{\text{law}}{=}(X_{m+1},\ldots,X_1)$ and the second applies stationarity.

Next, we show (99) for $n=m+2$ and the rest follows from induction on n. Indeed,

$$\begin{aligned}&\mathbb{P}\left[(X_1,\ldots,X_{m+2})=(x_1,\ldots,x_{m+2})\right]\\&=\pi(x_1,\ldots,x_m)M(x_{m+1}|x_1,\ldots,x_m)\\&\quad M(x_{m+2}|x_2,\ldots,x_{m+1})\\&\overset{(a)}{=}\pi(x_{m+1},\ldots,x_2)M(x_1|x_{m+1},\ldots,x_2)\\&\quad M(x_{m+2}|x_2,\ldots,x_{m+1})\\&\overset{(b)}{=}\pi(x_2,\ldots,x_{m+1})M(x_1|x_{m+1},\ldots,x_2)\\&\quad M(x_{m+2}|x_2,\ldots,x_{m+1})\\&\overset{(c)}{=}\pi(x_{m+2},\ldots,x_3)M(x_2|x_{m+2},\ldots,x_3)\\&\quad M(x_1|x_{m+1},\ldots,x_2)\\&=\mathbb{P}\left[(X_1,\ldots,X_{m+2})=(x_{m+2},\ldots,x_1)\right]\\&=\mathbb{P}\left[(X_{m+2},\ldots,X_1)=(x_1,\ldots,x_{m+2})\right].\end{aligned}$$

where (a) and (c) apply (99) for $n=m+1$, namely, (101); (b) applies (99) for $n=m$. □

In view of the proof of (100), we note that any distribution π on $\mathcal{S}^m$ and m^{th}-order transition matrix M satisfying $\pi(x^m)=\pi(\overleftarrow{x^m})$ and (101) also satisfy (98). This implies such a π is a stationary distribution for M. In view of Lemma 20 the above conditions also guarantee reversibility. This observation can be summarized in the following lemma, which will be used to prove the reversibility of specific Markov chains later.

Lemma 21: Let M be a $k^m\times k$ stochastic matrix describing transitions from $\mathcal{S}^m$ to $\mathcal{S}$. Suppose that π is a distribution on $\mathcal{S}^m$ such that $\pi(x^m)=\pi(\overleftarrow{x^m})$ and $\pi(x^m)M(x_{m+1}|x^m)=\pi(x_2^{m+1})M(x_1|x_2^{m+1})$. Then π is the stationary distribution of M and the resulting chain is reversible.

For m^{th}-order stationary Markov chains, the optimal prediction risk is defined as as

$$\begin{aligned}\mathsf{Risk}_{k,n,m}&\triangleq\inf_{\widehat{M}}\sup_{M}\mathbb{E}[D(M(\cdot|X_{n-m+1}^n)\|\widehat{M}(\cdot|X_{n-m+1}^n))]\\&=\inf_{\widehat{M}}\sup_{M}\sum_{x^m\in\mathcal{S}^m}\mathbb{E}\begin{bmatrix}D(M(\cdot|x^m)\|\widehat{M}(\cdot|x^m))\\\mathbf{1}_{\{X_{n-m+1}^n=x^m\}}\end{bmatrix}\end{aligned}\tag{103}$$

where the supremum is taken over all $k^m\times k$ stochastic matrices M and the trajectory is initiated from the stationary distribution. In the remainder of this section we will show the following result, completing the proof of Theorem 5 previously announced in Section I.

Theorem 22: For all $m\ge 2$, there exist a constant $C_m>0$ such that for all $2\le k\le n^{\frac{1}{m+1}}/C_m$,

$$\frac{k^{m+1}}{C_m n}\log\left(\frac{n}{k^{m+1}}\right)\le\mathsf{Risk}_{k,n,m}\le\frac{C_m k^{m+1}}{n}\log\left(\frac{n}{k^{m+1}}\right).$$

Furthermore, the lower bound holds even when the Markov chains are required to be reversible.

B. Upper Bound

We prove the upper bound part of the preceding theorem, using only stationarity (not reversibility). We rely on techniques from [22, Chapter 6, Page 486] for proving redundancy bounds for the m^{th}-order chains. Let Q be the probability assignment given by

$$Q(x^n)=\frac{1}{k^m}\prod_{a^m\in\mathcal{S}^m}\frac{\prod_{j=1}^k N_{a^mj}!}{k\cdot(k+1)\cdots(N_{a^m}+k-1)},\tag{104}$$

where N_{a^mj} denotes the number of times the block a^mj occurs in x^n, and $N_{a^m}=\sum_{j=1}^k N_{a^mj}$ is the number of times the block a^m occurs in x^{n-1}. This probability assignment corresponds to the add-one rule

$$Q(j|x^n)=\widehat{M}_{x^n}^{+1}(j|x_{n-m+1}^n)=\frac{N_{x_{n-m+1}^n j}+1}{N_{x_{n-m+1}^n}+k}.\tag{105}$$

Then in view of Lemma 6, the following lemma proves the desired upper bound in Theorem 22.

Lemma 23: Let $\mathsf{Red}(Q_{X^n})$ be the redundancy of the m^{th}-order Markov chain, as defined in Section II-A, and X^m be the corresponding observed trajectory. Then

$$\begin{aligned}\mathsf{Red}(Q_{X^n})\le&\frac{1}{n-m}\left\{k^m(k-1)\left[\log\left(1+\frac{n-m}{k^m(k-1)}\right)+1\right]\right.\\&\left.+m\log k\right\}\end{aligned}$$

Proof: We show that for every Markov chain with transition matrix M and initial distribution π on $\mathcal{S}^m$, and every trajectory $(x_1,\cdots,x_n)$, it holds that

$$\begin{aligned}&\log\frac{\pi(x_1^m)\prod_{t=m}^{n-1}M(x_{t+1}|x_{t-m+1}^t)}{Q(x_1,\cdots,x_n)}\\&\le k^m(k-1)\left[\log\left(1+\frac{n-m}{k^m(k-1)}\right)+1\right]+m\log k,\end{aligned}\tag{106}$$

where $M(x_{t+1}|x_{t-m+1}^t)$ the transition probability of going from x_{t-m+1}^t to x_{t+1}. Note that

$$\begin{aligned}\prod_{t=m}^{n-1}M(x_{t+1}|x_{t-m+1}^t)&=\prod_{a^{m+1}\in\mathcal{S}^{m+1}}M(a_{m+1}|a^m)^{N_{a^{m+1}}}\\&\le\prod_{a^{m+1}\in\mathcal{S}^{m+1}}(N_{a^{m+1}}/N_{a^m})^{N_{a^{m+1}}},\end{aligned}$$

where the last inequality follows from $\sum_{a_{m+1}\in\mathcal{S}} \frac{N_{a^{m+1}}}{N_{a^m}} \log \frac{N_{a^{m+1}}}{N_{a^m} M(a_{m+1}|a^m)} \geq 0$ for each a^m, by the non-negativity of the KL divergence. Therefore, we have

$$\begin{aligned}
&\frac{\pi(x_1^m)\prod_{t=m}^{n-1} M(x_{t+1}|x_{t-m+1}^t)}{Q(x_1,\cdots,x_n)} \\
&\leq k^m \cdot \prod_{a^m\in\mathcal{S}^m} \frac{k\cdot(k+1)\cdot\cdots\cdot(N_{a^m}+k-1)}{N_{a^m}^{N_{a^m}}} \\
&\quad \prod_{a_{m+1}\in\mathcal{S}} \frac{N_{a^{m+1}}^{N_{a^{m+1}}}}{N_{a^{m+1}}!}. \qquad (107)
\end{aligned}$$

Using (33) we continue (107) to get

$$\begin{aligned}
&\log \frac{\pi(x_1)\prod_{t=m}^{n-1} M(x_{t+1}|x_t)}{Q(x_1,\cdots,x_n)} \\
&\leq m\log k + \sum_{a^m\in\mathcal{S}^m} \log \frac{k\cdot(k+1)\cdot\cdots\cdot(N_{a^m}+k-1)}{N_{a^m}!} \\
&= m\log k + \sum_{a^m\in\mathcal{S}^m}\sum_{\ell=1}^{N_{a^m}} \log\left(1+\frac{k-1}{\ell}\right) \\
&\leq m\log k + \sum_{a^m\in\mathcal{S}^m}\int_0^{N_{a^m}} \log\left(1+\frac{k-1}{x}\right)dx \\
&= m\log k + \sum_{a^m\in\mathcal{S}^m}\left((k-1)\log\left(1+\frac{N_{a^m}}{k-1}\right)\right. \\
&\quad \left. + N_{a^m}\log\left(1+\frac{k-1}{N_{a^m}}\right)\right) \\
&\overset{(a)}{\leq} k^m(k-1)\log\left(1+\frac{n-m}{k^m(k-1)}\right) + k^m(k-1) \\
&\quad + m\log k,
\end{aligned}$$

where (a) follows from the concavity of $x \mapsto \log x$, $\sum_{a^m\in\mathcal{S}^m} N_{a^m} = n-m+1$, and $\log(1+x)\leq x$. □

Remark 6: The computational complexity of the estimator $M^{+1}(\cdot|X_{n-m+1}^n)$ for any given value of m is $O(nmk)$. This is because given any realization x_{n-m+1}^n of X_{n-m+1}^n, for any $j\in[k]$ and $t=1,\ldots,n-m$, it takes $O(m)$ time to check if X_t^{t+m} equals $x_{n-m+1}^n j$. Summing over t, j the $O(nmk)$ time complexity follows.

C. Lower Bound

The lower bound proof is divided into two cases: $m\geq 2, k=2$ and $m\geq 2, k\geq 3$. The idea for $m\geq 2, k=2$ is similar to the proof of Theorem 8. Same as before, it turns out that we can achieve the desired lower bound even with the smaller risk regime of squared error loss. We use the Bayesian strategy, with a prior that has the uniform stationary distribution on the second-order state space $\{11,12,21,22\}$, and identify the key trajectories with a significant contribution. Remember that the proof in the $k=3$ case for the first-order chains uses the following trajectory set

$$\mathcal{X} = \cup_{t=1}^{n-1}\left\{\begin{array}{c} x^n : x_1=\cdots=x_t=1, x_i\in\{2,3\}, \\ i=t+1,\ldots,n \end{array}\right\},$$

and study the probabilities related to it. A similar analysis using the transitions between the states $\{2,3\}$ as in the above trajectories is the key difficulty for the binary states space, which is where we modify our approach. In the current setup, we use the trajectory set

$$\mathcal{V} = \left\{\begin{array}{c} 1^{n-t}z^t : z_1=z_2=z_t=2, z_i^{i+1}\neq 11, \\ i\in[t-1], t=4,\ldots,n-2 \end{array}\right\}. \qquad (108)$$

The prior distribution class we study is uniform on the set one-parametric family of transition matrices

$$\widetilde{\mathcal{M}} = \left\{ M_p = \begin{array}{c} \\ 11 \\ 21 \\ 12 \\ 22 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 1-\frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & 1-\frac{1}{n} \\ 1-p & p \\ p & 1-p \end{bmatrix} \end{array} : 0\leq p\leq 1 \right\}.$$

We show that the z^t part of the trajectories in (108) can be represented using the transitions between the coupled states $\{22, 212\}$, which is similar to the transition between the states $\{2,3\}$ in the first-order chain analysis. Given the above, we achieve trajectory probabilities similar to (36), enabling us to imitate similar analyses of the first-order case.

The proof technique for $k\geq 3$ closely follows a similar strategy of using mutual information based lower bound as in Theorem 3, with the following characteristics in the constructed chain:

- the transition matrix is reversible;
- the trajectory initializes on the m-tuple state 1^m with a constant probability that depends only on m;
- the transition matrix includes an embedded matrix on the states space $\{2,\ldots,k\}$ that is a counterpart of the symmetric matrix T in (42). We show that a sufficient condition on a $(k-1)^m\times(k-1)$ dimensional Markov chain T is $T(x_{m+1}|x^m) = T(x_1|\overline{x_2^{m+1}})$ for every $x^{m+1}\in\{2,\ldots,k\}^m$, where $\overline{x_2^{m-1}} = (x_m, x_{m-1},\cdots,x_2)$ is the time reversal.

Similar to Section III-B, for the purpose of achieving the risk lower bound we only focus on the trajectories in

$$\begin{aligned}
\mathcal{X} &= \cup_{t=m}^{n-1}\mathcal{X}_t, \\
\mathcal{X}_t &= \{x^n : x_1,\ldots,x_t=1, x_{t+1},\ldots,x_n\neq 1\}.
\end{aligned}$$

Conditioned on each trajectory set $\mathcal{X}_t$ one can show that the smaller chain $(X_{t+1},\ldots,X_n)$ has the same distribution as a length-$(n-t)$ trajectory of a stationary m^{th}-order Markov chain with state space $\{2,\ldots,m\}$. Given this we demonstrate a mutual information based minimax risk lower bound (cf. Lemma 29)

$$\mathsf{Risk}_{k,n,m} \gtrsim \frac{c_m}{n}\left(I(T;Y^{n-m}) - m\log(k-1)\right), \qquad (109)$$

for some absolute constant c_m depending on m and Y^n is a stationary m^{th}-order Markov chain on $\{2,\ldots,k\}$. We proceed to lower bound the mutual information, which in turn asks for an upper bound of the mean squared error of a certain estimator similar to Lemma 12. However, the key difficulty in mimicking the previous proof is the first-order chain $\{Y_1^m, Y_2^{1+m},\ldots,Y_{n-m+1}^n\}$ is not reversible, and, as a result,

the analysis based on spectral decomposition and absolute spectral gap fails. To resolve this issue, we introduce the pseudo spectral gap. Although much more complicated to compute in general, we carefully add laziness to the prior construction of T so that the pseudo spectral gap could be easily controlled (cf. Lemma 31). This enables us to derive good estimation guarantees on T based on the trajectory Y^n, providing us with the desired lower bound.

Below we provide the details of the proofs.

Proof of the Lower Bound:

1) Special Case: $m = 2, k = 2$: The transition matrix for second-order chains is given by a $k^2 \times k$ stochastic matrices M that gives the transition probability from the ordered pairs $(i, j) \in \mathcal{S} \times \mathcal{S}$ to some state $\ell \in \mathcal{S}$:

$$M(\ell|ij) = \mathbb{P}\left[X_3 = \ell | X_1 = i, X_2 = j\right]. \tag{110}$$

Our result is the following.

Theorem 24: $\mathsf{Risk}_{2,n,2} = \Theta\left(\frac{\log n}{n}\right)$.

Proof: The upper bound part has been shown in Lemma 23. For the lower bound, consider the following one-parametric family of transition matrices (we replace $\mathcal{S}$ by $\{1, 2\}$ for simplicity of the notation)

$$\widetilde{\mathcal{M}} = \left\{ M_p = \begin{array}{c} \\ 11 \\ 21 \\ 12 \\ 22 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 1-\frac{1}{n} & \frac{1}{n} \\ \frac{1}{n} & 1-\frac{1}{n} \\ 1-p & p \\ p & 1-p \end{bmatrix} \end{array} : 0 \le p \le 1 \right\} \tag{111}$$

and place a uniform prior on $p \in [0, 1]$. One can verify that each M_p has the uniform stationary distribution over the set $\{1, 2\} \times \{1, 2\}$ and the chains are reversible.

Next we introduce the set of trajectories based on which we will lower bound the prediction risk. Analogous to the set $\mathcal{X} = \cup_{t=1}^{n} \mathcal{X}_t$ defined in (35) for analyzing the first-order chains, we define

$$\mathcal{V} = \left\{ \begin{array}{c} 1^{n-t} z^t : z_1 = z_2 = z_t = 2, z_i^{i+1} \neq 11, \\ i \in [t-1], t = 4, \ldots, n-2 \end{array} \right\} \subset \{1, 2\}^n. \tag{112}$$

In other words, the sequences in $\mathcal{V}$ start with a string of 1's before transitioning into two consecutive 2's, are forbidden to have no consecutive 1's thereafter, and finally end with 2.

To compute the probability of sequences in $\mathcal{V}$, we need the following preparations. Denote by $\oplus$ the the operation that combines any two blocks from $\{22, 212\}$ via merging the last symbol of the first block and the first symbol of the second block, for example, $22 \oplus 212 = 2212, 22 \oplus 22 \oplus 22 = 2222$. Then for any $x^n \in \mathcal{V}$ we can write it in terms of the initial all-1 string, followed by alternating run of blocks from $\{22, 212\}$ with the first run being of the block 22 (all the runs have positive lengths), combined with the merging operation $\oplus$:

$$\begin{aligned} x^n = & \underbrace{1 \ldots 1}_{\text{all ones}} \underbrace{22 \oplus 22 \cdots \oplus 22}_{p_1 \text{ many } 22} \oplus \underbrace{212 \oplus 212 \cdots \oplus 212}_{p_2 \text{ many } 212} \\ & \oplus \underbrace{22 \oplus 22 \cdots \oplus 22}_{p_3 \text{ many } 22} \oplus \underbrace{212 \oplus 212 \cdots \oplus 212}_{p_4 \text{ many } 212} \oplus 22 \oplus \ldots . \end{aligned} \tag{113}$$

Let the vector $(q_{22\to22}, q_{22\to212}, q_{212\to22}, q_{212\to212})$ denotes the transition probabilities between blocks in $\{22, 212\}$ (recall the convention that the two blocks overlap in the symbol 2). Namely, according to (111),

$$\begin{aligned} q_{22\to22} &= \mathbb{P}\left[X_3 = 2, X_2 = 2 | X_2 = 2, X_1 = 2\right] \\ &= M(2|22) = 1 - p \\ q_{22\to212} &= \mathbb{P}\left[X_4 = 2, X_3 = 1, X_2 = 2 | X_2 = 2, X_1 = 2\right] \\ &= M(2|21)M(1|22) = \left(1 - \frac{1}{n}\right) p \\ q_{212\to22} &= \mathbb{P}\left[X_4 = 2, X_3 = 2 | X_3 = 2, X_2 = 1, X_1 = 2\right] \\ &= M(2|12) = p \\ q_{212\to212} &= \mathbb{P}\left[\begin{array}{c|c} X_5 = 2, X_4 = 1, & X_3 = 2, X_2 = 1, \\ X_3 = 2 & X_1 = 2 \end{array} \right] \\ &= M(2|21)M(1|12) = \left(1 - \frac{1}{n}\right)(1 - p). \end{aligned}$$

Given any $x^n \in \mathcal{V}$ we can calculate its probability under the law of M_p using frequency counts $\boldsymbol{F}(x^n) = (F_{111}, F_{22\to22}, F_{22\to212}, F_{212\to22}, F_{212\to212})$, defined as

$$\begin{aligned} F_{111} &= \sum_i \mathbf{1}_{\{x_i=1, x_{i+1}=1, x_{i+2}=1\}}, \\ F_{22\to22} &= \sum_i \mathbf{1}_{\{x_i=2, x_{i+1}=2, x_{i+2}=2\}}, \\ F_{22\to212} &= \sum_i \mathbf{1}_{\{x_i=2, x_{i+1}=2, x_{i+2}=1, x_{i+3}=2\}}, \\ F_{212\to22} &= \sum_i \mathbf{1}_{\{x_i=2, x_{i+1}=1, x_{i+2}=2, x_{i+3}=2\}}, \\ F_{212\to212} &= \sum_i \mathbf{1}_{\{x_i=2, x_{i+1}=1, x_{i+2}=2, x_{i+3}=1, x_{i+4}=2\}}. \end{aligned}$$

Denote $\mu(x^n|p) = \mathbb{P}\left[X^n = x^n | p\right]$. Then for each $x^n \in \mathcal{V}$ with $\boldsymbol{F}(x^n) = \boldsymbol{F}$ we have

$$\begin{aligned} &\mu(x^n|p) \\ &= \mathbb{P}(X^{F_{111}+2} = 1^{F_{111}+2}) M(2|11) M(2|12) \\ &\quad \prod_{a,b \in \{22,212\}} q_{a\to b}^{F_{a\to b}} \\ &= \frac{1}{4}\left(1 - \frac{1}{n}\right)^{F_{111}} \frac{1}{n} \cdot p \cdot p^{F_{212\to22}} \left\{ p\left(1 - \frac{1}{n}\right) \right\}^{F_{22\to212}} \\ &\quad (1-p)^{F_{22\to22}} \left\{ (1-p)\left(1 - \frac{1}{n}\right) \right\}^{F_{212\to212}} \\ &= \frac{1}{4}\left(1 - \frac{1}{n}\right)^{F_{111}+F_{22\to212}+F_{212\to212}} \frac{1}{n} p^{y+1} (1-p)^{f-y} \end{aligned} \tag{114}$$

where $y = F_{212\to22} + F_{22\to212}$ denotes the number of times the chain alternates between runs of 22 and runs of 212, and

$f = F_{212\to 22} + F_{22\to 212} + F_{212\to 212} + F_{22\to 22}$ denotes the number of times the chain jumps between blocks in $\{22, 212\}$.

Note that the range of f includes all the integers in between 1 and $(n-6)/2$. This follows from the definition of $\mathcal{V}$ and the fact that if we merge either 22 or 212 using the operation $\oplus$ at the end of any string z^t with $z_t = 2$, it increases the length of the string by at most 2. Also, given any value of f the value of y ranges from 0 to f.

Lemma 25: The number of sequences in $\mathcal{V}$ corresponding to a fixed pair (y, f) is $\binom{f}{y}$.

Proof: Fix $x^n \in \mathcal{V}$ and let that p_{2i-1} is the length of the i-th run of 22 blocks and p_{2i} is the length of the i-th run of 212 blocks in x^n as depicted in (113). The p_i's are all non-negative integers. There are total $y+1$ such runs and the p_i's satisfy $\sum_{i=1}^{y+1} p_i = f+1$, as the total number of blocks is one more than the total number of transitions. Each positive integer solution to this equation $\{p_i\}_{i=1}^{y+1}$ corresponds to a sequence $x^n \in \mathcal{V}$ and vice versa. The total number of such sequences is $\binom{f}{y}$. □

We are now ready to compute the Bayes estimator and risk. For any $x^n \in \mathcal{V}$ with a given (y, f), the Bayes estimator of p with prior $p \sim \text{Uniform}[0,1]$ is

$$\widehat{p}(x^n) = \mathbb{E}[p|x^n] = \frac{\mathbb{E}[p \cdot \mu(x^n|p)]}{\mathbb{E}[\mu(x^n|p)]} \overset{(114)}{=} \frac{y+2}{f+3}.$$

Note that the probabilities $\mu(x^n|p)$ in (114) can be bounded from below by $\frac{1}{4en} p^{y+1}(1-p)^{f-y}$. Using this, for each $x^n \in \mathcal{V}$ with given y, f we get the following bound on the integrated squared error for a particular sequence x^n

$$\begin{aligned}
&\int_0^1 \mu(x^n|p)(p - \widehat{p}(x^n))^2 dp \\
&\geq \frac{1}{4en} \int_0^1 p^{y+1}(1-p)^{f-y} \left(p - \frac{y+2}{f+3}\right)^2 dp \\
&= \frac{1}{4en} \frac{(y+1)!(f-y)!}{(f+2)!} \frac{(y+2)(f-y+1)}{(f+3)^2(f+4)} \qquad (115)
\end{aligned}$$

where the last equality followed by noting that the integral is the variance of a Beta$(y+2, f-y+1)$ random variable without its normalizing constant.

Next we bound the risk of any predictor by the Bayes error. Consider any predictor $\widehat{M}(\cdot|ij)$ (as a function of the sample path X) for transition from ij, $i, j \in \{1, 2\}$. By the Pinsker's inequality, we conclude that

$$\begin{aligned}
D(M(\cdot|12)\|\widehat{M}(\cdot|12)) &\geq \frac{1}{2}\|M(\cdot|12) - \widehat{M}(\cdot|12)\|_{\ell_1}^2 \\
&\geq \frac{1}{2}(p - \widehat{M}(2|12))^2 \qquad (116)
\end{aligned}$$

and similarly, $D(M(\cdot|22)\|\widehat{M}(\cdot|22)) \geq \frac{1}{2}(p - \widehat{M}(1|22))^2$. Abbreviate $\widehat{M}(2|12) \equiv \widehat{p}_{12}$ and $\widehat{M}(1|22) \equiv \widehat{p}_{22}$, both functions of X. Using (115) and Lemma 25, we have

$$\begin{aligned}
&\sum_{i,j=1}^{3} \mathbb{E}[D(M(\cdot|ij)\|\widehat{M}(\cdot|ij)))\mathbf{1}_{\{X_{n-1}^n = ij\}}] \\
&\geq \frac{1}{2}\mathbb{E}\begin{bmatrix} (p-\widehat{p}_{12})^2 \mathbf{1}_{\{X_{n-1}^n = 12, X^n \in \mathcal{V}\}} \\ +(p-\widehat{p}_{22})^2 \mathbf{1}_{\{X_{n-1}^n = 22, X^n \in \mathcal{V}\}} \end{bmatrix} \\
&\geq \frac{1}{2}\int_0^1 \begin{bmatrix} \sum_{\boldsymbol{F}} \sum_{x^n \in \mathcal{V}: \boldsymbol{F}(x^n) = \boldsymbol{F}} \\ \mu(x^n|p) \begin{pmatrix} (p-\widehat{p}_{12})^2 \mathbf{1}_{\{x_{n-1}^n = 12\}} \\ +(p-\widehat{p}_{22})^2 \mathbf{1}_{\{x_{n-1}^n = 22\}} \end{pmatrix} \end{bmatrix} dp \\
&\geq \frac{1}{2}\int_0^1 \left[\sum_{\boldsymbol{F}} \sum_{x^n \in \mathcal{V}: \boldsymbol{F}(x^n) = \boldsymbol{F}} \mu(x^n|p)(p - \widehat{p}(x^n))^2 \right] dp \\
&\geq \frac{1}{2} \sum_{f=1}^{\frac{n-6}{2}} \sum_{y=0}^{f} \binom{f}{y} \frac{1}{4en} \left\{ \frac{(y+1)!(f-y)!}{(f+2)!} \frac{(y+2)(f-y+1)}{(f+3)^2(f+4)} \right\} \\
&\geq \frac{1}{8en} \sum_{f=1}^{\frac{n-6}{2}} \sum_{y=0}^{f} \frac{y+1}{(f+2)(f+1)} \frac{(y+2)(f-y+1)}{(f+3)^2(f+4)} \\
&\geq \Theta\left(\frac{1}{n}\right) \sum_{f=1}^{\frac{n-6}{2}} \sum_{y=\frac{f}{4}}^{\frac{f}{3}} \frac{1}{f^2} = \Theta\left(\frac{\log n}{n}\right). \qquad (117)
\end{aligned}$$

□

2) General Case: $m \geq 2, k \geq 3$: We will prove the following.

Theorem 26: For absolute constant C, we have

$$\begin{aligned}
&\text{Risk}_{k,n,m} \\
&\geq \frac{1}{2^{m+4}} \left(\frac{1}{2} - \frac{2^m - 2}{n}\right) \left(1 - \frac{1}{n}\right)^{n-2m+1} \frac{(k-1)^{m+1}}{n} \\
&\quad \log\left(\frac{1}{2^{2m+8} \cdot 3\pi e(m+1)} \cdot \frac{n-m}{(k-1)^{m+1}}\right).
\end{aligned}$$

For ease of notation let $\mathcal{S} = \{1, \ldots, k\}$. Denote $\widetilde{\mathcal{S}} = \{2, \ldots, k\}$. Consider an m^{th}-order transition matrix M of the following form (with $b = \frac{1}{2} - \frac{2^m - 2}{n}$):

$M(s|x^m)$

Starting string x^m	Next state	
	$s = 1$	$s \in \{2, \ldots, k\}$
1^m	$1 - \frac{1}{n}$	$\frac{1}{n(k-1)}$
$1x^{m-1}$, $x^{m-1} \in \widetilde{\mathcal{S}}^{m-1}$	$1 - b$	$\frac{b}{(k-1)}$
$x^m \in \widetilde{\mathcal{S}}^m$	$\frac{1}{n}$	$\left(1 - \frac{1}{n}\right) T(s\|x^m)$
$x^m \notin \left\{ \begin{matrix} 1^m, 1\widetilde{\mathcal{S}}^{m-1}, \\ \widetilde{\mathcal{S}}^m \end{matrix} \right\}$	$\frac{1}{2}$	$\frac{1}{2(k-1)}$

(118)

Here T is a $(k-1)^m \times (k-1)$ transition matrix for an m^{th}-order Markov chain with state space $\widetilde{\mathcal{S}}$, satisfying the following property:

(P) $T(x_{m+1}|x^m) = T(x_1|\overline{x_2^{m+1}}), \quad \forall x^{m+1} \in \widetilde{\mathcal{S}}^{m+1}$.

Lemma 27: Under the condition (P), the transition matrix T has a stationary distribution that is uniform on $\widetilde{\mathcal{S}}^m$. Furthermore, the resulting m^{th}-order Markov chain is reversible (and hence stationary).

Proof: We prove this result using Lemma 21. Let π denote the uniform distribution on $\widetilde{\mathcal{S}}^m$, i.e., $\pi(x^m) = \frac{1}{(k-1)^m}$ for all $x^m \in \widetilde{\mathcal{S}}^m$. Then for any $x^m \in \widetilde{\mathcal{S}}^m$ the condition $\pi(\underline{x^m}) = \pi(\overline{x^m})$ follows directly and $\pi(x^m)T(x_{m+1}|x^m) = \pi(\overline{x_2^{m+1}})T(x_1|\overline{x_2^{m+1}})$ follows from the assumption (P). □

Next we address the stationarity and reversibility of the chain with the bigger transition matrix M in (118):

Lemma 28: Let M be defined in (118), wherein the transition matrix T satisfies the condition (P). Then M has a stationary distribution given by

$$\pi(x^m) = \begin{cases} \frac{1}{2} & x^m = 1^m \\ \frac{b}{(k-1)^m} & x^m \in \widetilde{\mathcal{S}}^m \\ \frac{1}{n(k-1)^{d(x^m)}} & \text{otherwise} \end{cases} \tag{119}$$

where $d(x^m) \triangleq \sum_{i=1}^m \mathbf{1}_{\{x_i \in \widetilde{\mathcal{S}}\}}$ and $b = \frac{1}{2} - \frac{2^m-2}{n}$ as in (118). Furthermore, the m^{th}-order Markov chain with initial distribution π and transition matrix M is reversible.

Proof: Note that the choice of b guarantees that $\sum_{x^m \in \mathcal{S}^m} \pi(x^m) = 1$. Next we again apply Lemma 21 to verify stationarity and reversibility. First of all, since $d(x^m) = d(\overline{x^m})$, we have $\pi(x^m) = \pi(\overline{x^m})$ for all $x^m \in \mathcal{S}^m$. Next we check the condition $\pi(x^m)M(x_{m+1}|x^m) = \pi(\overline{x_2^{m+1}})M(x_1|\overline{x_2^{m+1}})$. For the sequence 1^{m+1} the claim is easily verified. For the rest of the sequences we have the following.

- **Case 1** ($x^{m+1} \in \widetilde{\mathcal{S}}^{m+1}$)**:** Note that $x^{m+1} \in \widetilde{\mathcal{S}}^{m+1}$ if and only if $x^m, \overline{x_2^{m+1}} \in \widetilde{\mathcal{S}}^m$. This implies

$$\begin{aligned} \pi(x^m)M(x_{m+1}|x^m) &= \frac{b}{(k-1)^m}\left(1-\frac{1}{n}\right)T(x_{m+1}|x^m) \\ &= \frac{b}{(k-1)^m}\left(1-\frac{1}{n}\right)T(x_1|\overline{x_2^{m+1}}) \\ &= \pi(\overline{x_2^{m+1}})M(x_1|\overline{x_2^{m+1}}). \end{aligned}$$

- **Case 2** ($x^{m+1} \in 1\widetilde{\mathcal{S}}^m$ **or** $x^{m+1} \in \widetilde{\mathcal{S}}^m 1$)**:** By symmetry it is sufficient to analyze the case $x^{m+1} \in 1\widetilde{\mathcal{S}}^m$. Note that in the sub-case $x^{m+1} \in 1\widetilde{\mathcal{S}}^m$, $x^m \in 1\widetilde{\mathcal{S}}^{m-1}$ and $\overline{x_2^{m+1}} \in \widetilde{\mathcal{S}}^m$. This implies

$$\begin{aligned} \pi(x^m) &= \frac{1}{n(k-1)^{m-1}}, \quad M(x_{m+1}|x^m) = \frac{b}{k-1}, \\ \pi(\overline{x_2^{m+1}}) &= \frac{b}{(k-1)^m}, \quad M(x_1|\overline{x_2^{m+1}}) = \frac{1}{n}. \end{aligned} \tag{120}$$

In view of this we get

$$\pi(x^m)M(x_{m+1}|x^m) = \pi(\overline{x_2^{m+1}})M(x_1|\overline{x_2^{m+1}}).$$

- **Case 3** ($x^{m+1} \notin 1^{m+1} \cup \widetilde{\mathcal{S}}^{m+1} \cup 1\widetilde{\mathcal{S}}^m \cup \widetilde{\mathcal{S}}^m 1$)**:** Suppose that x^{m+1} has d many elements from $\widetilde{\mathcal{S}}$. Then $x^m, x_2^{m+1} \notin \left\{1^m, \widetilde{\mathcal{S}}^m\right\}$. We have the following sub-cases.
  - If $x_1 = x_{m+1} = 1$, then both x^m, x_2^{m+1} have exactly d elements from $\widetilde{\mathcal{S}}$. This implies $\pi(x^m) = \pi(\overline{x_2^{m+1}}) = \frac{1}{n(k-1)^d}$ and $M(x_{m+1}|x^m) = M(x_1|\overline{x_2^{m+1}}) = \frac{1}{2}$.
  - If $x_1, x_{m+1} \in \widetilde{\mathcal{S}}$, then both x^m, x_2^{m+1} have exactly $d-1$ elements from $\widetilde{\mathcal{S}}$. This implies $\pi(x^m) = \pi(\overline{x_2^{m+1}}) = \frac{1}{n(k-1)^{d-1}}$ and $M(x_{m+1}|x^m) = M(x_1|\overline{x_2^{m+1}}) = \frac{1}{2(k-1)}$.
  - If $x_1 = 1, x_{m+1} \in \widetilde{\mathcal{S}}$, then x^m has $d-1$ elements from $\widetilde{S}$ and x_2^{m+1} has d elements from $\mathcal{S}$. This implies $\pi(x^m) = \frac{1}{n(k-1)^{d-1}}, \pi(\overline{x_2^{m+1}}) = \frac{1}{n(k-1)^d}$ and $M(x_{m+1}|x^m) = \frac{1}{2(k-1)}, M(x_1|\overline{x_2^{m+1}}) = \frac{1}{2}$.
  - If $x_1 \in \widetilde{\mathcal{S}}, x_{m+1} = 1$, then x^m has d elements from $\widetilde{S}$ and x_2^{m+1} has $d-1$ elements from $\mathcal{S}$ then $\pi(x^m) = \frac{1}{n(k-1)^d}, \pi(\overline{x_2^{m+1}}) = \frac{1}{n(k-1)^{d-1}}$ and $M(x_{m+1}|x^m) = \frac{1}{2}, M(x_1|\overline{x_2^{m+1}}) = \frac{1}{2(k-1)}$.

  For all these sub-cases we have $\pi(x^m)M(x_{m+1}|x^m) = \pi(\overline{x_2^{m+1}})M(x_1|\overline{x_2^{m+1}})$ as required.

This finishes the proof. □

Let $(X_1, \ldots, X_n)$ be the trajectory of a stationary Markov chain with transition matrix M as in (118). We observe the following properties:

(R1) This Markov chain is irreducible and reversible. Furthermore, the stationary distribution π assigns probability $\frac{1}{2}$ to the initial state 1^m.

(R2) For $m \le t \le n-1$, let $\mathcal{X}_t$ denote the collections of trajectories x^n such that $x_1, x_2, \cdots, x_t = 1$ and $x_{t+1}, \cdots, x_n \in \widetilde{\mathcal{S}}$. Then using Lemma 28

$$\begin{aligned} &\mathbb{P}(X^n \in \mathcal{X}_t) \\ &= \mathbb{P}(X_1 = \cdots = X_t = 1) \cdot \mathbb{P}(X_{t+1} \neq 1 | X_{t-m+1}^t = 1^m) \\ &\quad \cdot \prod_{i=2}^{m-1} \mathbb{P}(X_{t+i} \neq 1 | X_{t-m+i}^t = 1^{m-i+1}, X_{t+1}^{t+i-1} \in \widetilde{\mathcal{S}}^{i-1}) \\ &\quad \cdot \mathbb{P}(X_{t+m} \neq 1 | X_t = 1, X_{t+1}^{t+m-1} \in \widetilde{\mathcal{S}}^{m-1}) \\ &\quad \cdot \prod_{s=t+m}^{n-1} \mathbb{P}(X_{s+1} \neq 1 | X_{s-m+1}^s \in \widetilde{\mathcal{S}}^m) \\ &= \frac{1}{2} \cdot \left(1-\frac{1}{n}\right)^{t-m} \cdot \frac{b}{n2^{m-2}} \cdot \left(1-\frac{1}{n}\right)^{n-m-t} \\ &= \frac{b}{n2^{m-1}}\left(1-\frac{1}{n}\right)^{n-2m}. \end{aligned} \tag{121}$$

Moreover, this probability does not depend of the choice of T;

(R3) Conditioned on the event that $X^n \in \mathcal{X}_t$, the trajectory $(X_{t+1}, \cdots, X_n)$ has the same distribution as a length-$(n-t)$ trajectory of a stationary m^{th}-order Markov chain with state space $\widetilde{\mathcal{S}}$ and transition probability T, and the uniform initial distribution. Indeed,

$$\begin{aligned} &\mathbb{P}\left[X_{t+1} = x_{t+1}, \ldots, X_n = x_n | X^n \in \mathcal{X}_t\right] \\ &= \left(\frac{\frac{1}{2} \cdot \left(1-\frac{1}{n}\right)^{t-m} \cdot \frac{b}{n2^{m-2}(k-1)^m}}{\frac{b}{n2^{m-1}}\left(1-\frac{1}{n}\right)^{n-2m}}\right). \end{aligned}$$

$$\prod_{s=t+m}^{n-1}\left(1-\frac{1}{n}\right)T(x_{s+1}|x_{s-m+1}^{s})$$
$$=\frac{1}{(k-1)^m}\prod_{s=t+m}^{n-1}T(x_{s+1}|x_{s-m+1}^{s}).$$

a) Reducing the Bayes prediction risk to mutual information: Consider the following Bayesian setting, we first draw T from some prior satisfying property (P), then generate the stationary m^{th}-order Markov chain $X^n=(X_1,\ldots,X_n)$ with state space $[k]$ and transition matrix M in (118) and stationary distribution π in (119). The following lemma lower bounds the Bayes prediction risk.

Lemma 29: Conditioned on T, let $Y^n=(Y_1,\ldots,Y_n)$ denote an m^{th}-order stationary Markov chain on state space $\widetilde{\mathcal{S}}=\{2,\ldots,k\}$ with transition matrix T and uniform initial distribution. Then

$$\inf_{\widehat{M}}\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_{n-m+1}^{n})\|\widehat{M}(\cdot|X_{n-m+1}^{n}))]\right]$$
$$\geq\frac{b(n-1)}{n^2 2^{m-1}}\left(1-\frac{1}{n}\right)^{n-2m}\left(I(T;Y^{n-m})-m\log(k-1)\right).$$

Proof: We first relate the Bayes estimator of M and T (given the X and Y chain respectively). For each $m\leq t\leq n$, denote by $\widehat{M}_t=\widehat{M}_t(\cdot|x^t)$ the Bayes estimator of $M(\cdot|x_{t-m+1}^{t})$ given $X^t=x^t$, and $\widehat{T}_t(\cdot|y^t)$ the Bayes estimator of $T(\cdot|y_{t-m+1}^{t})$ given $Y^t=y^t$. For each $t=1,\ldots,n-1$ and for each trajectory $x^n=(1,\ldots,1,x_{t+1},\ldots,x_n)\in\mathcal{X}_t$, recalling the form (21) of the Bayes estimator, we use the relation between M and T to get, for each $j\in\widetilde{\mathcal{S}}$,

$$\begin{aligned}
&\widehat{M}_n(j|x^n)\\
&=\frac{\mathbb{P}\left[X^{n+1}=(x^n,j)\right]}{\mathbb{P}\left[X^n=x^n\right]}\\
&=\frac{(1-\frac{1}{n})\mathbb{E}[\frac{1}{(k-1)^m}\prod_{s=t+m}^{n-1}T(x_{s+1}|x_{s-m+1}^{s})T(j|x_{n-m+1}^{n})]}{\mathbb{E}[\frac{1}{(k-1)^m}\prod_{s=t+m}^{n-1}T(x_{s+1}|x_{s-m+1}^{s})]}\\
&=\left(1-\frac{1}{n}\right)\frac{\mathbb{P}\left[Y^{n-t+1}=(x_{t+1}^{n},j)\right]}{\mathbb{P}\left[Y^{n-t}=x_{t+1}^{n}\right]}\\
&=\left(1-\frac{1}{n}\right)\widehat{T}_{n-t}(j|x_{t+1}^{n}).
\end{aligned}$$

Furthermore, since $M(1|x^m)=1/n$ for all $x^m\in\widetilde{\mathcal{S}}$ in the construction (118), the Bayes estimator also satisfies $\widehat{M}_n(1|x^n)=1/n$ for $x^n\in\mathcal{X}_t$ and $t\leq n-m$. In all, we have for $x^n\in\mathcal{X}_t, t\leq n-m$

$$\widehat{M}_n(\cdot|x^n)=\frac{1}{n}\delta_1+\left(1-\frac{1}{n}\right)\widehat{T}_{n-t}(\cdot|x_{t+1}^{n}). \tag{122}$$

with δ_1 denoting the point mass at state 1, which parallels the fact that

$$M(\cdot|y^m)=\frac{1}{n}\delta_1+\left(1-\frac{1}{n}\right)T(\cdot|y^m),\quad y^m\in\widetilde{\mathcal{S}}^m. \tag{123}$$

By (R2), each event $\{X^n\in\mathcal{X}_t\}$ occurs with probability at least $\frac{b}{n2^{m-1}}\left(1-\frac{1}{n}\right)^{n-2m}$, and is independent of T. Therefore,

$$\begin{aligned}
&\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_{n-1}X_n)\|\widehat{M}(\cdot|X^n))]\right]\\
&\geq\frac{b}{n2^{m-1}}\left(1-\frac{1}{n}\right)^{n-2m}\sum_{t=m}^{n-m}\\
&\quad\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_{n-m+1}^{n})\|\widehat{M}(\cdot|X^n))|X^n\in\mathcal{X}_t]\right]. \qquad (124)
\end{aligned}$$

By (R3), the conditional joint law of $(T,X_{t+1},\ldots,X_n)$ on the event $\{X^n\in\mathcal{X}_t\}$ is the same as the joint law of $(T,Y_1,\ldots,Y_{n-t})$. Thus, we may express the Bayes prediction risk in the X chain as

$$\begin{aligned}
&\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_{n-m+1}^{n})\|\widehat{M}(\cdot|X^n))|X^n\in\mathcal{X}_t]\right]\\
&\overset{(a)}{=}\left(1-\frac{1}{n}\right)\cdot\mathbb{E}_T\left[\mathbb{E}[D(T(\cdot|Y_{n-t-m+1}^{n-t})\|\widehat{T}(\cdot|Y^{n-t}))]\right]\\
&\overset{(b)}{=}\left(1-\frac{1}{n}\right)\cdot I(T;Y_{n-t+1}|Y^{n-t}), \qquad (125)
\end{aligned}$$

where (a) follows from (122), (123), and the fact that for distributions P,Q supported on $\widetilde{\mathcal{S}}$, $D(\epsilon\delta_1+(1-\epsilon)P\|\epsilon\delta_1+(1-\epsilon)Q)=(1-\epsilon)D(P\|Q)$; (b) is the mutual information representation (20) of the Bayes prediction risk. Finally, the lemma follows from (124), (125), and the chain rule

$$\begin{aligned}
&\sum_{t=m}^{n-m}I(T;Y_{n-t+1}|Y^{n-t})\\
&=I(T;Y^{n-m})-I(T;Y^m)\geq I(T;Y^{n-m})-m\log(k-1),
\end{aligned}$$

as $I(T;Y^m)\leq H(Y^m)\leq m\log(k-1)$. □

b) Prior construction and lower bounding the mutual information: We assume that $k=2k_0+1$ for some integer k_0. For simplicity of notation we replace $\widetilde{S}$ by $\mathcal{Y}=1,\ldots,k-1$. This does not affect the lower bound. Define an equivalent relation on $|\mathcal{Y}|^{m-1}$ given by the following rule: x^{m-1} and y^{m-1} are related if and only if $x^{m-1}=y^{m-1}$ or $x^{m-1}=\overline{y^{m-1}}$. Let R_{m-1} be a subset of $\mathcal{Y}^{m-1}$ that consists of exactly one representative from each of the equivalent classes. As each of the equivalent classes under this relation will have at most two elements the total number of equivalent classes is at least $\frac{|\mathcal{Y}|^{m-1}}{2}$, i.e., $|R_{m-1}|\geq\frac{(k-1)^{m-1}}{2}$. We consider the following prior: let $u=\left\{u_{ix^{m-1}j}\right\}_{i\leq j\in[k_0],x^{m-1}\in R_{m-1}}$ be iid and uniformly distributed in $[1/(4k_0),3/(4k_0)]$ and for each $i\leq j,x^{m-1}\in R_{m-1}$ define $u_{jx^{m-1}i},u_{i\overline{x^{m-1}}j},u_{j\overline{x^{m-1}}i}$ to be same as $u_{ix^{m-1}j}$. Let the transition matrix T be given by

$$\begin{aligned}
&T(2j-1|2i-1,x^{m-1})=T(2j|2i,x^{m-1})=u_{ix^{m-1}j},\\
&T(2j|2i-1,x^{m-1})=T(2j-1|2i,x^{m-1})=\frac{1}{k_0}-u_{ix^{m-1}j},\\
&\quad i,j\in\mathcal{Y},x^{m-1}\in\mathcal{Y}^{m-1}. \qquad (126)
\end{aligned}$$

One can check that the constructed T is a stochastic matrix and satisfies the property (P), which enforces uniform stationary distribution. Also each entry of T belongs to the interval $[\frac{1}{2(k-1)},\frac{3}{2(k-1)}]$.

Next we use the following lemma to derive estimation guarantees on T.

Lemma 30: Suppose that T is an $\ell^m \times \ell$ transition matrix, on state space $\mathcal{Y}^m$ with $|\mathcal{Y}| = \ell$, satisfying $T(x_{m+1}|x^m) = T(x_1|x_2^{m+1}), \quad \forall x^{m+1} \in [\ell]^{m+1}$ and $T(y_{m+1}|y^m) \in [\frac{c_1}{\ell}, \frac{c_2}{\ell}]$ with $0 < c_1 < 1 < c_2$ for all $y^{m+1} \in [\ell]^{m+1}$. Then there is an estimator $\widehat{T}$ based on stationary trajectory Y^n simulated from T such that

$$\mathbb{E}[\|\widehat{T}-T\|_{\mathsf{F}}^2] \le \frac{4c_1^{2m+3}(m+1)\ell^{2m}}{c_2(n-m)},$$

where

$$\|\widehat{T}-T\|_{\mathsf{F}} = \sqrt{\sum_{y^{m+1}} (\widehat{T}(y_{m+1}|y^m) - T(y_{m+1}|y^m))^2}$$

denotes the Frobenius norm.

For our purpose we will use the above lemma on T with $\ell = k-1, c_1 = \frac{1}{2}, c_2 = \frac{3}{2}$. Therefore it follows that there exist estimators $\widehat{T}(Y^n)$ and $\widehat{u}(Y^n)$ such that

$$\begin{aligned}\mathbb{E}[\|\widehat{u}(Y^n) - u\|_2^2] \le& \mathbb{E}[\|\widehat{T}(Y^n)-T\|_{\mathsf{F}}^2] \\ \le& \frac{4c_2(m+1)(k-1)^{2m}}{c_1^{2m+3}(n-m)}. \end{aligned} \tag{127}$$

Here and below, we identify $u = \{u_{ix^{m-1}j}\}_{i\le j, x^{m-1}\in R_{m-1}}$ and $\widehat{u} = \{\widehat{u}_{ix^{m-1}j}\}_{i\le j, x^{m-1}\in R_{m-1}}$ as $\frac{|R_{m-1}|k_0(k_0+1)}{2} = \frac{|R_{m-1}|(k^2-1)}{8}$-dimensional vectors.

Let $h(X) = \int -f_X(x)\log f_X(x)dx$ denote the differential entropy of a continuous random vector X with density f_X w.r.t the Lebesgue measure and $h(X|Y) = \int -f_{XY}(xy)\log f_{X|Y}(x|y)dxdy$ the conditional differential entropy (cf. e.g. [44]). Then

$$\begin{aligned} h(u) =& \sum_{i\le j\in[k_0], x^{m-1}\in R_{m-1}} h(u_{ix^{m-1}j}) \\ =& -\frac{|R_{m-1}|(k^2-1)}{8}\log(k-1). \end{aligned} \tag{128}$$

Then

$$\begin{aligned} & I(T;Y^n) \overset{(a)}{=} I(u;Y^n) \\ & \overset{(b)}{\ge} I(u;\widehat{u}(Y^n)) = h(u) - h(u|\widehat{u}(Y^n)) \\ & \overset{(c)}{\ge} h(u) - h(u-\widehat{u}(Y^n)) \\ & \overset{(d)}{\ge} \frac{|R_{m-1}|(k^2-1)}{16} \\ & \quad \log\left(\frac{c_1^{2m+3}|R_{m-1}|(k^2-1)(n-m)}{64\pi e c_2(m+1)(k-1)^{2m+2}}\right) \\ & \ge \frac{(k-1)^{m+1}}{32}\log\left(\frac{n-m}{c_m(k-1)^{m+1}}\right). \end{aligned}$$

for constant $c_m = \frac{128\pi e c_2(m+1)}{c_1^{2m+3}}$, where (a) is because u and T are in one-to-one correspondence by (126); (b) follows from the data processing inequality; (c) is because $h(\cdot)$ is translation invariant and concave; (d) follows from the maximum entropy principle [44]: $h(u-\widehat{u}(Y^n)) \le \frac{|R_{m-1}|(k^2-1)}{16}\log\left(\frac{2\pi e}{|R_{m-1}|(k^2-1)/8}\cdot\mathbb{E}[\|\widehat{u}(Y^n)-u\|_2^2]\right)$, which in turn is bounded by (127). Plugging this lower bound into Lemma 29 completes the lower bound proof of Theorem 22.

3) Proof of Lemma 30 via Pseudo Spectral Gap: In view of Lemma 27 we get that the stationary distribution of T is uniform over $\mathcal{Y}^m$, and there is a one-to-one correspondence between the joint distribution of Y^{m+1} and the transition probabilities

$$\mathbb{P}\left[Y^{m+1} = y^{m+1}\right] = \frac{1}{\ell^m}T(y_{m+1}|y^m). \tag{129}$$

Consider the following estimator $\widehat{T}$: for $y_{m+1} \in [\ell]^{m+1}$, let

$$\widehat{T}(y_{m+1}|y^m) = \ell^m \cdot \frac{\sum_{t=1}^{n-m}\mathbf{1}_{\{Y_t^{t+m}=y^{m+1}\}}}{n-m}.$$

Clearly $\mathbb{E}[\widehat{T}(y_{m+1}|y^m)] = \ell^m\mathbb{P}[y_{m+1}|y^m] = T(y_{m+1}|y^m)$. Next we observe that the sequence of random variables $\{Y_t^{t+m}\}_{t=1}^{n-m}$ is a first-order Markov chain on $[\ell]^{m+1}$. Let us denote its transition matrix by T_{m+1} and note that its stationary distribution is given by $\pi(a^{m+1}) = \ell^{-m}T(a_{m+1}|a^m), a^{m+1} \in [\ell]^{m+1}$. For the transition matrix T_{m+1}, which must be non-reversible, the *pseudo spectral gap* $\gamma_{\mathrm{ps}}(T_{m+1})$ is defined as

$$\gamma_{\mathrm{ps}}(T_{m+1}) = \max_{r\ge 1}\frac{\gamma((T_{m+1}^*)^r T_{m+1}^r)}{r},$$

where T_{m+1}^* is the adjoint of T_{m+1} defined as $T_{m+1}^*(b^{m+1}|a^{m+1}) = \pi(b^{m+1})T(a^{m+1}|b^{m+1})/\pi(a^{m+1})$. With these notations, the concentration inequality of [18, Theorem 3.2] gives the following variance bound:

$$\begin{aligned}\mathrm{Var}(\widehat{T}(y_{m+1}|y^m)) \le& \ell^{2m}\cdot\frac{4\mathbb{P}\left[Y^{m+1}=y^{m+1}\right]}{\gamma_{\mathrm{ps}}(T_{m+1})(n-m)} \\ \le& \ell^{2m}\cdot\frac{4T(y_{m+1}|y^m)\ell^{-m}}{\gamma_{\mathrm{ps}}(T_{m+1})(n-m)}.\end{aligned}$$

The following lemma bounds the pseudo spectral gap from below.

Lemma 31: Let $T \in \mathbb{R}^{\ell^m\times\ell}$ be the transition matrix of an m-th order Markov chain $(Y_t)_{t\ge1}$ over a discrete state space $\mathcal{Y}$ with $|\mathcal{Y}| = \ell$, and assume that

- all the entries of T lie in the interval $[\frac{c_1}{\ell}, \frac{c_2}{\ell}]$ for some absolute constants $0 < c_1 < c_2$;
- T has the uniform stationary distribution on $[\ell]^m$.

Let $T_{m+1} \in \mathbb{R}^{\ell^{m+1}\times\ell^{m+1}}$ be the transition matrix of the first-order Markov chain $((Y_t, Y_{t+1}, \cdots, Y_{t+m}))_{t\ge1}$. Then we have

$$\gamma_{\mathrm{ps}}(T_{m+1}) \ge \frac{c_1^{2m+3}}{c_2(m+1)}.$$

Consequently, we have

$$\begin{aligned}\mathbb{E}[\|\widehat{T}-T\|_{\mathsf{F}}^2] =& \sum_{y^{m+1}\in[\ell]^{m+1}}\mathrm{Var}(\widehat{T}(y_{m+1}|y^m)) \\ \le& \sum_{y^{m+1}\in[\ell]^{m+1}}\frac{4c_2(m+1)\ell^m}{c_1^{2m+3}}\cdot\frac{T(y_{m+1}|y^m)}{n-m} \\ =& \frac{4c_2(m+1)\ell^{2m}}{c_1^{2m+3}(n-m)},\end{aligned}$$

completing the proof.

Proof of Lemma 31: As T_{m+1} is a first-order Markov chain, the stochastic matrix T_{m+1}^{m+1} defines the probabilities of transition from $(Y_t, Y_{t+1}, \cdots, Y_{t+m})$ to $(Y_{t+m+1}, Y_{t+m+2}, \cdots, Y_{t+2m+1})$. By our assumption on T

$$\min_{a^{2m+2}\in\mathcal{Y}^{2m+2}} T_{m+1}^{m+1}(a_{m+2}^{2m+2}|a^{m+1}) \geq \prod_{t=0}^{m} T(a_{2m+2-t}|a_{m+2-t}^{2m+1-t}) \geq \frac{c_1^{m+1}}{\ell^{m+1}}. \tag{130}$$

Given any $a^{m+1}, b^{m+1} \in \mathcal{Y}^{m+1}$, using the above inequality we have

$$\begin{aligned}
&(T_{m+1}^*)^{m+1}(b^{m+1}|a^{m+1}) \\
&= \sum_{\boldsymbol{y}_1\in\mathcal{Y}^{m+1},\ldots,\boldsymbol{y}_m\in\mathcal{Y}^{m+1}} T_{m+1}^*(b^{m+1}|\boldsymbol{y}_m) \\
&\quad\cdot\left\{\prod_{t=1}^{m-1} T_{m+1}^*(\boldsymbol{y}_{m-t+1}|\boldsymbol{y}_{m-t})\right\} \\
&\quad\cdot T_{m+1}^*(\boldsymbol{y}_1|a^{m+1}) \\
&= \sum_{\boldsymbol{y}_1\in\mathcal{Y}^{m+1},\ldots,\boldsymbol{y}_m\in\mathcal{Y}^{m+1}} \frac{\pi(b^{m+1})T_{m+1}(\boldsymbol{y}_m|b^{m+1})}{\pi(\boldsymbol{y}_m)} \\
&\quad\cdot\left\{\prod_{t=1}^{m-1} \frac{\pi(\boldsymbol{y}_{m-t+1})T_{m+1}(\boldsymbol{y}_{m-t}|\boldsymbol{y}_{m-t+1})}{\pi(\boldsymbol{y}_{m-t})}\right\} \\
&\quad\cdot\frac{\pi(\boldsymbol{y}_1)T_{m+1}(a^{m+1}|\boldsymbol{y}_1)}{\pi(a^{m+1})} \\
&= \frac{\pi(b^{m+1})}{\pi(a^{m+1})} \sum_{\boldsymbol{y}_1\in\mathcal{Y}^{m+1},\ldots,\boldsymbol{y}_m\in\mathcal{Y}^{m+1}} T_{m+1}(\boldsymbol{y}_m|b^{m+1}) \\
&\quad\cdot\left\{\prod_{t=1}^{m-1} T_{m+1}(\boldsymbol{y}_{m-t}|\boldsymbol{y}_{m-t+1})\right\} T_{m+1}(a^{m+1}|\boldsymbol{y}_1) \\
&= \frac{\pi(b^{m+1})}{\pi(a^{m+1})} T_{m+1}^{m+1}(a^{m+1}|b^{m+1}) \\
&= \frac{\pi(b^m)T(b_{m+1}|b^m)}{\pi(b^m)T(a_{m+1}|a^m)} T_{m+1}^{m+1}(a^{m+1}|b^{m+1}) \geq \frac{c_1}{c_2}\cdot\frac{c_1^{m+1}}{\ell^{m+1}}.
\end{aligned} \tag{131}$$

Using (130),(131) we get

$$\begin{aligned}
&\min_{a^{m+1},b^{m+1}\in\mathcal{Y}^{m+1}} \left\{(T_{m+1}^*)^{m+1}T_{m+1}^{m+1}\right\}(b^{m+1}|a^{m+1}) \\
&\geq \sum_{d^{m+1}\in\mathcal{Y}^{m+1}} \left(\min_{a^{m+1},d^{m+1}\in\mathcal{Y}^{m+1}} (T_{m+1}^*)^{m+1}(d^{m+1}|a^{m+1})\right) \\
&\quad\left(\min_{b^{m+1},d^{m+1}\in\mathcal{Y}^{m+1}} T_{m+1}^{m+1}(b^{m+1}|d^{m+1})\right) \\
&\geq \sum_{d^{m+1}\in\mathcal{Y}^{m+1}} \frac{c_1^{2m+3}}{c_2\ell^{2m+2}} \geq \frac{c_1^{2m+3}}{c_2\ell^{m+1}}.
\end{aligned} \tag{132}$$

As $(T_{m+1}^*)^{m+1}T_{m+1}^{m+1}$ is an $\ell^{m+1}\times\ell^{m+1}$ stochastic matrix, we can use Lemma 32 to get the lower bound on its spectral gap $\gamma((T_{m+1}^*)^{m+1}T_{m+1}^{m+1}) \geq \frac{c_1^{2m+3}}{c_2}$. Hence we get

$$\gamma_{\text{ps}}(T_{m+1}) \geq \frac{\gamma((T_{m+1}^*)^{m+1}T_{m+1}^{m+1})}{m+1} \geq \frac{c_1^{2m+3}}{c_2(m+1)} \tag{133}$$

as required. A more generalized version of Lemma 32 can be found in from [54].

Lemma 32: Suppose that A is a $d\times d$ stochastic matrix with $\min_{i,j} A_{ij} \geq \epsilon$. Then for any eigenvalue λ of A other than 1 we have $|\lambda| \leq 1 - d\epsilon$.

Proof: Suppose that λ is an eigenvalue of A other than 1 with non-zero left eigenvector $\boldsymbol{v}$, i.e. $\lambda v_j = \sum_{i=1}^d v_i A_{ij}, j = 1,\ldots,d$. As A is a stochastic matrix we know that $\sum_j A_{ij} = 1$ for all i and hence $\sum_{i=1}^d v_i = 0$. This implies

$$\begin{aligned}
|\lambda v_j| &= \left|\sum_{i=1}^d v_i A_{ij}\right| = \left|\sum_{i=1}^d v_i(A_{ij}-\epsilon)\right| \\
&\leq \sum_{i=1}^d |v_i(A_{ij}-\epsilon)| = \sum_{i=1}^d |v_i|(A_{ij}-\epsilon)
\end{aligned} \tag{134}$$

with the last equality following from $A_{ij} \geq \epsilon$. Summing over $j = 1,\ldots d$ in the above equation and dividing by $\sum_{i=1}^d |v_i|$ we get $|\lambda| \leq 1 - d\epsilon$ as required. □

□

VI. Extensions to Other Loss Functions

As mentioned in Section I-A, standard arguments based on concentration inequalities inevitably rely on mixing conditions such as the spectral gap. In contrast, the risk bound in Theorem 1, which is free of any mixing condition, is enabled by powerful techniques from universal compression which bound the redundancy by the pointwise maximum over all trajectories combined with information-theoretic or combinatorial argument. This program only relies on the Markovity of the process rather than stationarity or spectral gap assumptions. The limitation of this approach, however, is that the reduction between prediction and redundancy crucially depends on the form of the KL loss function in (1), which allows one to use the mutual information representation and the chain rule to relate individual risks to the cumulative risk.

In this section we present some preliminary results where the prediction risk are measure by other f-divergences. Let us first mention that for certain f-divergences the minimax risk is infinite. For example, consider the reverse KL loss, where the prediction loss is assessed by $D(\widehat{M}(\cdot|X_n)\|M(\cdot|X_n))$ as opposed to $D(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))$ in (1). This is somewhat surprising because for iid data, it is easy to show that the optimal rate of the reverse KL risk is $\Theta(\frac{k}{n})$ for $k = O(n)$, achieved simply by the empirical distribution. To see why for Markov chains the reserve KL risk is infinite, consider a chain which visits state 1 at time n *for the first time*, which happens with some positive probability, and we are tasked to estimate the first row of the transition matrix $P(\cdot) \equiv M(\cdot|1)$, despite that no sample from P is available whatsoever. Since $\sup_P \inf_{\widehat{P}} D(\widehat{P}\|P) = \infty$, the worst-case reverse KL loss is also infinite. In contrast, $\sup_P D(P\|\text{Uniform}([k])) = \log k < \infty$, and the add-one estimator (5) will precisely output uniform if the state 1 was never observed previously; as such, the usual KL prediction loss is finite, as characterized

Next, we focus on the *total variation* and the χ^2*-divergence*. For simplicity, we focus on stationary reversible first-order Markov chains.

A. Total Variation

Consider the counterpart for the minimax risk (1), when we replace the KL prediction loss by squared total variation, namely,

$$\begin{aligned}\mathsf{Risk}_{k,n}^{\mathsf{TV}} &\triangleq \inf_{\widehat{M}} \sup_{\pi,M} \mathbb{E}[\mathrm{TV}(\widehat{M}(\cdot|X_n), M(\cdot|X_n))^2] \\ &= \inf_{\widehat{M}} \sup_{\pi,M} \sum_{i=1}^{k} \mathbb{E}[\mathrm{TV}(M(\cdot|i), \widehat{M}(\cdot|i))^2 \mathbf{1}_{\{X_n=i\}}],\end{aligned} \tag{135}$$

where for distributions P, Q on $[k]$, $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{i=1}^{k}|P(i)-Q(i)|$. The following bounds on the optimal total variation risk are readily obtainable from the characterization of the KL risk: For all $3 \le k \lesssim \sqrt{n}$,

$$\frac{\log n}{n} \lesssim \mathsf{Risk}_{k,n}^{\mathsf{TV}} \lesssim \frac{k^2}{n}\log\left(\frac{n}{k^2}\right). \tag{136}$$

Indeed, the upper bound follows from Theorem 1 plus Pinsker's inequality. For the lower bound, we apply the construction for $k=3$ states from Section III-A and simply notice from (40) that the lower bound therein is shown for the squared error of estimating individual transition probabilities and hence applies to the squared total variation.

The bound (136) shows that for small state space with $3 \le k = O(1)$, the optimal prediction risk measured in squared total variation is $\Theta(\frac{\log n}{n})$, which is strictly slower than $\Theta(\frac{1}{n})$ for iid processes on small alphabets. Determining the optimal rate in total variation for binary or large state space is an outstanding question. In particular, the lower bound strategy in Section III-B based on embedding a $(k-1)$-state chain in a k-state chain is no longer viable due to the lack of chain rule for total variation.

B. χ^2-Loss

Next we extend the spectral gap dependent bound in Theorem 3 to for the χ^2-divergence loss. Given any two discrete distributions $P = (P_1, \dots P_k) \ll Q = (Q_1, \dots Q_k)$ on $[k]$, their χ^2-divergence is defined as

$$\chi^2(P\|Q) = \sum_{i=1}^{k} \frac{(P_i - Q_i)^2}{Q_i} \tag{137}$$

and ∞ if $P \not\ll Q$. Let $\mathsf{Risk}_{k,n}^{\chi^2}(\gamma_0)$ denote the counterpart of (9) with χ^2-loss, namely

$$\mathsf{Risk}_{k,n}^{\chi^2}(\gamma_0) \triangleq \inf_{\widehat{M}} \sup_{M \in \mathcal{M}_k(\gamma_0)} \mathbb{E}\left[\chi^2(M(\cdot|X_n)\|\widehat{M}(\cdot|X_n))\right] \tag{138}$$

where the supremum is taken over all reversible Markov chains with absolute spectral gap at least γ_0. Then we have the following result.

Theorem 33: Given any $k \ge 2$, $\mathsf{Risk}_{k,n}^{\chi^2}(\gamma_0) \le \frac{C}{\gamma_0^4}\frac{k^2}{n}$.

Proof: The proof is similar to that of Theorem 3(i). Fix $\gamma_0 \in (0,1)$. Assuming the chain terminates in state 1 we bound the risk of estimating the first row by the add-one estimator $\widehat{M}^{+1}(j|1) = \frac{N_{1j}+1}{N_1+k}$ with $O\left(\frac{k}{n}\right)$. By symmetry then the overall risk bound becomes $O\left(\frac{k^2}{n}\right)$. In particular, under the absolute spectral gap condition of $\gamma_* \ge \gamma_0$, we show

$$\mathbb{E}\left[\mathbf{1}_{\{X_n=1\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] \lesssim \frac{k}{n\gamma_0^4}. \tag{139}$$

We decompose the left hand side in (139) based on N_1 as

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{X_n=1\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}_{\{A^{\le}\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] \\ &\quad + \mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]\end{aligned}$$

where the atypical set $A^{\le}$ and the typical set $A^{>}$are defined as before in (74)

$$\begin{aligned}A^{\le} &\triangleq \{X_n = 1, N_1 \le (n-1)\pi_1/2\}, \\ A^{>} &\triangleq \{X_n = 1, N_1 > (n-1)\pi_1/2\}.\end{aligned}$$

For the atypical case, note the following deterministic property of the add-one estimator. Let $\widehat{Q}$ be an add-one estimator with sample size n and alphabet size k of the form $\widehat{Q}_i = \frac{n_i+1}{n+k}$, where $\sum n_i = n$. Since $\widehat{Q}$ is bounded below by $\frac{1}{n+k}$ everywhere, for any distribution P, we have

$$\chi^2(P\|\widehat{Q}) \le (n+k). \tag{140}$$

Applying this bound on the event $A^{\le}$, we have for any $s > 0$

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{A^{\le}\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] \\ &\le (n\pi_1 + k)\,\mathbb{P}\left[X_n = 1, N_1 \le (n-1)\pi_1/2\right] \\ &\overset{(a)}{\le} \mathbf{1}_{\{n\pi_1\gamma_* \le s\}}\pi_1\,(n\pi_1+k) \\ &\quad + \mathbf{1}_{\{n\pi_1\gamma_* > s\}}\pi_1\,(n\pi_1+k)\,\frac{\mathbb{E}\left[(N_1-(n-1)\pi_1)^4\,|X_n=1\right]}{n^4\pi_1^4/16} \\ &\overset{(b)}{\lesssim} \mathbf{1}_{\{n\pi_1\gamma_* \le s\}}\frac{s}{n\gamma_*}\left(\frac{s}{\gamma_*}+k\right) \\ &\quad + \mathbf{1}_{\{n\pi_1\gamma_* > s\}}\,(n\pi_1+k)\left(\frac{1}{n^2\pi_1\gamma_*^2} + \frac{1}{n^4\pi_1^3\gamma_*^4}\right) \\ &\overset{(c)}{\lesssim} \frac{1}{n}\left\{\frac{s^2}{\gamma_*^2} + \frac{sk}{\gamma_*}\right\} + \frac{\mathbf{1}_{\{n\pi_1\gamma_* > s\}}}{n}\left(\frac{1}{\gamma_*^2} + \frac{1}{n^2\pi_1^2\gamma_*^4}\right. \\ &\quad \left. + \frac{k}{n\pi_1\gamma_*^2} + \frac{k}{n^3\pi_1^3\gamma_*^4}\right) \\ &\lesssim \frac{1}{n}\left\{\frac{s^2}{\gamma_*^2} + \frac{sk}{\gamma_*} + \frac{1}{\gamma_*^2} + \frac{1}{s^2\gamma_*^2} + \frac{k}{s\gamma_*} + \frac{k}{s^3\gamma_*}\right\}\end{aligned}$$

where we got (a) from Markov inequality, (b) from Lemma 16(iii). Plugging $s = 10$ we get

$$\mathbb{E}\left[\mathbf{1}_{\{A^{\le}\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] \lesssim \frac{k}{n\gamma_*^2} \le \frac{k}{n\gamma_0^2}. \tag{141}$$

Next we bound $\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}\chi^2\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]$. Define

$$\Delta_i = \frac{(M(i|1)N_1 - N_{1i} + M(i|1)k - 1)^2}{(N_1+k)\,(N_{1i}+1)}.$$

As $\chi^2(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)) = \sum_{i=1}^{k}\Delta_i$ it suffices to bound $\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}\Delta_i\right]$ for each i. For some $r \ge 1$ to be optimized later consider the following cases separately

(a) $n\pi_1 \le r$ or $n\pi_1 M(i|1) \le 10$: We have

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\Delta_i\right] &= \mathbb{E}\left[\frac{\mathbf{1}_{\{A^>\}}\left(M(i|1)N_1 - N_{1i} + M(i|1)k - 1\right)^2}{(N_1+k)(N_{1i}+1)}\right]\\ &\overset{(a)}{\lesssim} \frac{\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\left(M(i|1)N_1 - N_{1i}\right)^2\right] + k^2\pi_1 M(i|1)^2 + \pi_1}{n\pi_1 + k}\\ &\overset{(b)}{\lesssim} \frac{\pi_1 \mathbb{E}\left[\left(M(i|1)N_1 - N_{1i}\right)^2 \middle| X_n = 1\right]}{n\pi_1 + k} + \frac{1 + rkM(i|1)}{n}\end{aligned} \tag{142}$$

where (a) follows from $N_1 > \frac{(n-1)\pi_1}{2}$ in $A^>$ and the fact that $(x+y+z)^2 \le 3(x^2+y^2+z^2)$; (b) uses the assumption that either $n\pi_1 \le r$ or $n\pi_1 M(i|1) \le 10$. Applying Lemma 16(i) and the fact that $x + x^2 \le 2(1+x^2)$, with $x = \frac{\sqrt{M(i|1)}}{\gamma_*}$ continuing the last display we get

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\Delta_i\right] &\lesssim \frac{n\pi_1 M(i|1) + \left(1 + \frac{M(i|1)}{\gamma_*^2}\right)}{n} + \frac{1 + rkM(i|1)}{n}\\ &\lesssim \frac{1 + rkM(i|1)}{n} + \frac{M(i|1)}{n\gamma_0^2}.\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\chi^2(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right]\\ &= \sum_{i=1}^{k}\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\Delta_i\right] \lesssim \frac{rk}{n} + \frac{1}{\gamma_0^2}.\end{aligned} \tag{143}$$

(b) $n\pi_1 > r$ and $n\pi_1 M(i|1) > 10$:
We decompose $A^>$ based on count of N_{1i} into atypical part $B^\le$ and typical part $B^>$

$$\begin{aligned}B^\le &\triangleq \{X_n = 1, N_1 > (n-1)\pi_1/2, N_{1i} \le (n-1)\pi_1 M(i|1)/4\}\\ B^> &\triangleq \{X_n = 1, N_1 > (n-1)\pi_1/2, N_{1i} > (n-1)\pi_1 M(i|1)/4\}\end{aligned}$$

and bound each of $\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right]$ and $\mathbb{E}\left[\mathbf{1}_{\{B^>\}}\Delta_i\right]$ separately.

(i) **Bound on $\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right]$:** We have

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right]\\ &= \mathbb{E}\left[\frac{\mathbf{1}_{\{B^\le\}}\left(M(i|1)N_1 - N_{1i} + M(i|1)k - 1\right)^2}{(N_1+k)(N_{1i}+1)}\right]\\ &\lesssim \frac{\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\left\{\left(M(i|1)N_1 - N_{1i}\right)^2 + k^2 M(i|1)^2\right\}\right] + \pi_1}{n\pi_1 + k}\\ &\lesssim \frac{\pi_1 \mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\left(M(i|1)N_1 - N_{1i}\right)^2 \middle| X_n = 1\right]}{n}\\ &\quad + \frac{k^2\pi_1 M(i|1)^2 \mathbb{P}[B^\le | X_n = 1]}{n\pi_1 + k} + \frac{1}{n}\\ &\lesssim \frac{\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\left(M(i|1)N_1 - N_{1i}\right)^2 \middle| X_n = 1\right]}{n}\\ &\quad + k\pi_1 M(i|1)^2 \mathbb{P}[B^\le | X_n = 1] + \frac{1}{n}.\end{aligned} \tag{144}$$

For the set $B^\le$, using the fourth moment bound from Lemma 16(ii) and Markov inequality

$$\begin{aligned}&\mathbb{P}[B^\le | X_n = 1]\\ &\le \mathbb{P}\left[N_1 M(i|1) - N_{1i} > \frac{(n-1)\pi_1 M(i|1)}{4} | X_n = 1\right]\\ &\le \frac{256\mathbb{E}\left[\left(N_1 M(i|1) - N_{1i}\right)^4 | X_n = 1\right]}{\left((n-1)\pi_1 M(i|1)\right)^4}.\end{aligned} \tag{145}$$

We use the following bound on the above fourth moment from Lemma 16(ii) with the inequality $x + x^4 \le 2(1+x^4)$, with $x = \frac{\sqrt{M(i|1)}}{\gamma_*}$

$$\begin{aligned}&\mathbb{E}\left[\left(N_1 M(i|1) - N_{1i}\right)^4 | X_n = 1\right]\\ &\lesssim \left(n\pi_1 M(i|1)\right)^2 + \frac{\sqrt{M(i|1)}}{\gamma_*} + \frac{M(i|1)^2}{\gamma_*^4}\\ &\le \left(n\pi_1 M(i|1)\right)^2 + 2\left(1 + \frac{M(i|1)^2}{\gamma_*^4}\right).\end{aligned} \tag{146}$$

Then, for the first term in (144), using Cauchy-Schwarz inequality, (145), and (146) we get

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\left(M(i|1)N_1 - N_{1i}\right)^2 \middle| X_n = 1\right]\\ &\le \sqrt{\mathbb{P}[B^\le | X_n = 1]}\\ &\sqrt{\mathbb{E}\left[\left(M(i|1)N_1 - N_{1i}\right)^4 | X_n = 1\right]}\\ &\lesssim \frac{\mathbb{E}\left[\left(M(i|1)N_1 - N_{1i}\right)^4 | X_n = 1\right]}{\left(n\pi_1 M(i|1)\right)^2} \lesssim 1 + \frac{1}{r^2\gamma_*^4}\end{aligned} \tag{147}$$

For the second term, using Markov's inequality with the second moment to bound $\mathbb{P}[B^\le | X_n = 1]$ we get

$$\begin{aligned}&\pi_1 M(i|1)\mathbb{P}[B^\le | X_n = 1]\\ &\lesssim \pi_1 M(i|1)\frac{\mathbb{E}\left[\left(M(i|1)N_1 - N_{1i}\right)^2 | X_n = 1\right]}{\left(n\pi_1 M(i|1)\right)^2}\\ &\lesssim \frac{1}{n}\left(1 + \frac{1}{r\gamma_*^2}\right).\end{aligned} \tag{148}$$

Combining (147) and (148), in view of (144) we get

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right]\\ &\lesssim \frac{1}{n}\left(1 + \frac{1}{r^2\gamma_*^4}\right) + \frac{kM(i|1)}{n}\left(1 + \frac{1}{r\gamma_*^2}\right)\\ &\lesssim \frac{1 + kM(i|1)}{n}\left(1 + \frac{1}{r^2\gamma_0^4}\right)\end{aligned} \tag{149}$$

(ii) **Bound on $\mathbb{E}\left[\mathbf{1}_{\{B^>\}}\Delta_i\right]$:** We have

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\{B^>\}}\Delta_i\right]\\ &= \mathbb{E}\left[\frac{\mathbf{1}_{\{B^>\}}\left(M(i|1)N_1 - N_{1i} + M(i|1)k - 1\right)^2}{(N_1+k)(N_{1i}+1)}\right]\\ &\overset{(a)}{\lesssim} \frac{\mathbb{E}\left[\mathbf{1}_{\{B^>\}}\left\{\left(M(i|1)N_1 - N_{1i}\right)^2\right\}\right]}{(n\pi_1 + k)(n\pi_1 M(i|1) + 1)}\\ &\quad + \frac{k^2\pi_1 M(i|1)^2 + \pi_1}{(n\pi_1 + k)(n\pi_1 M(i|1) + 1)}\end{aligned}$$

$$\lesssim \frac{\pi_1 \mathbb{E}\left[(M(i|1)N_1 - N_{1i})^2 \middle| X_n = 1\right]}{(n\pi_1 + k)(n\pi_1 M(i|1) + 1)} + \frac{kM(i|1)}{n}$$

where (a) follows using properties of the set $B^>$ along with $(x+y+z)^2 \le 3(x^2+y^2+z^2)$. Using Lemma 16(i) we get

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{B^>\}}\Delta_i\right] \\
&\lesssim \frac{n\pi_1 M(i|1) + \left(1 + \frac{M(i|1)}{\gamma_*^2}\right)}{n(n\pi_1 M(i|1) + 1)} + \frac{kM(i|1)}{n} \\
&\lesssim \frac{1 + kM(i|1)}{n} + \frac{M(i|1)}{n\gamma_0^2}.
\end{aligned}$$

Combining the last bound and (149) and summing them over $i \in [k]$, we get for $n\pi_1 > r, n\pi_1 M(i|1) > 10$

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\chi^2(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right] \\
&= \sum_{i=1}^{k}\left[\mathbb{E}\left[\mathbf{1}_{\{B^\le\}}\Delta_i\right] + \mathbb{E}\left[\mathbf{1}_{\{B^>\}}\Delta_i\right]\right] \lesssim \frac{k}{n}\left(1 + \frac{1}{r\gamma_0^4}\right).
\end{aligned}$$

Combining this with (143) we obtain

$$\mathbb{E}\left[\mathbf{1}_{\{A^>\}}\chi^2(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right] \lesssim \frac{k}{n}\left(r + \frac{1}{r\gamma_0^4}\right).$$

where we chose $r = 10$ for the last inequality. In view of (141) this implies the required bound.

□

VII. Discussions and Open Problems

We discuss the assumptions and implications of our results as well as related open problems.

a) Very large state space: Theorem 1 determines the optimal prediction risk under the assumption of $k \lesssim \sqrt{n}$. When $k \gtrsim \sqrt{n}$, Theorem 1 shows that the KL risk is bounded away from zero. However, as the KL risk can be as large as $\log k$, it is a meaningful question to determine the optimal rate in this case, which, thanks to the general reduction in (11), reduces to determining the redundancy for symmetric and general Markov chains. For iid data, the minimax *pointwise* redundancy is known to be $n \log \frac{k}{n} + O(\frac{n^2}{k})$ [48, Theorem 1] when $k \gg n$. Since the average and pointwise redundancy usually behave similarly, for Markov chains it is reasonable to conjecture that the redundancy is $\Theta(n \log \frac{k^2}{n})$ in the large alphabet regime of $k \gtrsim \sqrt{n}$, which, in view of (11), would imply the optimal prediction risk is $\Theta(\log \frac{k^2}{n})$ for $k \gg \sqrt{n}$. In comparison, we note that the prediction risk is at most $\log k$, achieved by the uniform distribution.

b) Stationarity: As mentioned above, the redundancy result in Lemma 7 (see also [26], [27]) holds for nonstationary Markov chains as well. However, our redundancy-based risk upper bound in Lemma 6 crucially relies on stationarity. It is unclear whether the result of Theorem 1 carries over to nonstationary chains.

Appendix A
Mutual Information Representation of Prediction Risk

The following lemma justifies the representation (22) for the prediction risk as maximal conditional mutual information. Unlike (17) for redundancy which holds essentially without any condition [45], here we impose certain compactness assumptions which hold finite alphabets such as finite-state Markov chains studied in this paper.

Lemma 34: Let $\mathcal{X}$ be finite and let Θ be a compact subset of $\mathbb{R}^d$. Given $\{P_{X^{n+1}|\theta} : \theta \in \Theta\}$, define the prediction risk

$$\mathsf{Risk}_n \triangleq \inf_{Q_{X_{n+1}|X^n}} \sup_{\theta\in\Theta} D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n|\theta}), \tag{150}$$

Then

$$\mathsf{Risk}_n = \sup_{P_\theta \in \mathcal{M}(\Theta)} I(\theta; X_{n+1}|X^n). \tag{151}$$

where $\mathcal{M}(\Theta)$ denotes the collection of all (Borel) probability measures on Θ.

Note that for stationary Markov chains, (22) follows from Lemma 34 since one can take θ to be the joint distribution of $(X_1, \ldots, X_{n+1})$ itself which forms a compact subset of the probability simplex on $\mathcal{X}^{n+1}$.

Proof: It is clear that (150) is equivalent to

$$\begin{aligned}
&\mathsf{Risk}_n \\
&= \inf_{Q_{X_{n+1}|X^n}} \sup_{P_\theta\in\mathcal{M}(\Theta)} \\
&\quad D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n,\theta}).
\end{aligned} \tag{152}$$

By the variational representation (14) of conditional mutual information, we have

$$\begin{aligned}
&I(\theta; X_{n+1}|X^n) \\
&= \inf_{Q_{X_{n+1}|X^n}} D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n,\theta}).
\end{aligned} \tag{153}$$

Thus (151) amounts to justifying the interchange of infimum and supremum in (150). It suffices to prove the upper bound.

Let $|\mathcal{X}| = K$. For $\epsilon \in (0,1)$, define an auxiliary quantity:

$$\begin{aligned}
\mathsf{Risk}_{n,\epsilon} \triangleq &\inf_{Q_{X_{n+1}|X^n} \ge \frac{\epsilon}{K}} \sup_{P_\theta\in\mathcal{M}(\Theta)} \\
&D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n,\theta}),
\end{aligned} \tag{154}$$

where the constraint in the infimum is pointwise, namely, $Q_{X_{n+1}=x_{n+1}|X^n=x^n} \ge \frac{\epsilon}{K}$ for all $x_1, \ldots, x_{n+1} \in \mathcal{X}$. By definition, we have $\mathsf{Risk}_n \le \mathsf{Risk}_{n,\epsilon}$. Furthermore, $\mathsf{Risk}_{n,\epsilon}$ can be equivalently written as

$$\begin{aligned}
&\mathsf{Risk}_{n,\epsilon} \\
&= \inf_{Q_{X_{n+1}|X^n}} \sup_{P_\theta\in\mathcal{M}(\Theta)} \\
&\quad D(P_{X_{n+1}|X^n,\theta}\|(1-\epsilon)Q_{X_{n+1}|X^n} + \epsilon U|P_{X^n,\theta}),
\end{aligned} \tag{155}$$

where U denotes the uniform distribution on $\mathcal{X}$.

We first show that the infimum and supremum in (155) can be interchanged. This follows from the standard minimax theorem. Indeed, note that $D(P_{X_{n+1}|X^n,\theta}\|(1-\epsilon)Q_{X_{n+1}|X^n} + \epsilon U|P_{X^n,\theta})$ is convex in $Q_{X_{n+1}|X^n}$, affine in P_θ, continuous

in each argument, and takes values in $[0, \log \frac{K}{\epsilon}]$. Since $\mathcal{M}(\Theta)$ is convex and weakly compact (by Prokhorov's theorem) and the collection of conditional distributions $Q_{X_{n+1}|X^n}$ is convex, the minimax theorem (see, e.g., [55, Theorem 2]) yields

$$\begin{aligned}&\mathsf{Risk}_{n,\epsilon}\\&= \sup_{\pi\in\mathcal{M}(\Theta)} \inf_{Q_{X_{n+1}|X^n}}\\&D(P_{X_{n+1}|X^n,\theta}\|(1-\epsilon)Q_{X_{n+1}|X^n}+\epsilon U|P_{X^n,\theta}).\end{aligned} \tag{156}$$

Finally, by the convexity of the KL divergence, for any P on $\mathcal{X}$, we have

$$\begin{aligned}D(P\|(1-\epsilon)Q+\epsilon U) \le&(1-\epsilon)D(P\|Q)+\epsilon D(P\|U)\\ \le&(1-\epsilon)D(P\|Q)+\epsilon\log K,\end{aligned} \tag{157}$$

which, in view of (153) and (156), implies

$$\mathsf{Risk}_n \le \mathsf{Risk}_{n,\epsilon} \le \sup_{P_\theta\in\mathcal{M}(\Theta)} I(\theta; X_{n+1}|X^n)+\epsilon\log K.$$

By the arbitrariness of ϵ, (151) follows. □

Appendix B
Proof of Lemma 16

Recall that for any irreducible and reversible finite states transition matrix M with stationary distribution π the followings are satisfied:

1) $\pi_i > 0$ for all i.
2) $M(j|i)\pi_i = M(i|j)\pi_j$ for all i, j.

The following is a direct consequence of the Markov property.

Lemma 35: For any $1 \le t_1 < \cdots < t_m < \cdots < t_k$ and any $Z_2 = f\left(X_{t_k}, \ldots, X_{t_m}\right), Z_1 = g\left(X_{t_{m-1}}, \ldots, X_{t_1}\right)$ we have

$$\begin{aligned}&\mathbb{E}\left[Z_2\mathbf{1}_{\{X_{t_m}=j\}}Z_1|X_1=i\right]\\&=\mathbb{E}\left[Z_2|X_{t_m}=j\right]\mathbb{E}\left[\mathbf{1}_{\{X_{t_m}=j\}}Z_1|X_1=i\right]\end{aligned} \tag{158}$$

For $t \ge 0$, denote the t-step transition probability by $\mathbb{P}\left[X_{t+1}=j|X_1=i\right] = M^t(j|i)$, which is the ijth entry of M^t. The following result is standard (see, e.g., [17, Chap. 12]). We include the proof mainly for the purpose of introducing the spectral decomposition.

Lemma 36: Define $\lambda_* \triangleq 1-\gamma_* = \max\{|\lambda_i| : i \ne 1\}$. For any $t \ge 0$, $|M^t(j|i)-\pi_j| \le \lambda_*^t\sqrt{\frac{\pi_j}{\pi_i}}$.

Proof: Throughout the proof all vectors are column vectors except for π. Let D_π denote the diagonal matrix with entries $D_\pi(i,i) = \pi_i$. By reversibility, $D_\pi^{\frac{1}{2}}MD_\pi^{-\frac{1}{2}}$, which shares the same spectrum with M, is a symmetric matrix and admits the spectral decomposition $D_\pi^{\frac{1}{2}}MD_\pi^{-\frac{1}{2}} = \sum_{a=1}^k \lambda_a u_a u_a^\top$ for some orthonormal basis $\{u_1, \ldots, u_k\}$; in particular, $\lambda_1 = 1$ and $u_{1i} = \sqrt{\pi_i}$. Then for each $t \ge 1$,

$$M^t = \sum_{a=1}^k \lambda_a^t D_\pi^{-\frac{1}{2}}u_a u_a^\top D_\pi^{\frac{1}{2}} = \mathbf{1}\pi + \sum_{a=2}^k \lambda_a^t D_\pi^{-\frac{1}{2}}u_a u_a^\top D_\pi^{\frac{1}{2}}. \tag{159}$$

where $\mathbf{1}$ is the all-ones vector. As u_a's satisfy $\sum_{a=1}^k u_a u_a^\top = I$ we get $\sum_{a=2}^k u_{ab}^2 = 1-u_{a1}^2 \le 1$ for any $b = 1, \ldots, k$. Using this along with Cauchy-Schwarz inequality we get

$$\begin{aligned}\left|M^t(j|i)-\pi_j\right| &\le \sqrt{\frac{\pi_j}{\pi_i}}\sum_{a=2}^k |\lambda_a|^t |u_{ai}u_{aj}|\\ &\le \lambda_*^t\sqrt{\frac{\pi_j}{\pi_i}}\left(\sum_{a=2}^k u_{ai}^2\right)^{\frac{1}{2}}\left(\sum_{a=2}^k u_{aj}^2\right)^{\frac{1}{2}} \le \lambda_*^t\sqrt{\frac{\pi_j}{\pi_i}}\end{aligned}$$

as required. □

Lemma 37: Fix states i, j. For any integers $a \ge b \ge 1$, define for $s = 1, 2, 3, 4$

$$h_s(a,b) = \left|\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i\}}\left(\mathbf{1}_{\{X_a=j\}}-M(j|i)\right)^s|X_b=i\right]\right|.$$

Then

(i) $h_1(a,b) \le 2\sqrt{M(j|i)}\lambda_*^{a-b}$
(ii) $|h_2(a,b)-\pi_i M(j|i)(1-M(j|i))| \le 4\sqrt{M(j|i)}\lambda_*^{a-b}$.
(iii)

$$\begin{aligned}&h_3(a,b), h_4(a,b)\\&\le \pi_i M(j|i)(1-M(j|i))+4\sqrt{M(j|i)}\lambda_*^{a-b}.\end{aligned}$$

Proof: We apply Lemma 36 and time reversibility:

(i)

$$\begin{aligned}h_1(a,b) =& \left|\mathbb{P}\left[X_{a+1}=i, X_a=j|X_b=i\right]\right.\\&\left.-M(j|i)\mathbb{P}\left[X_{a+1}=i|X_b=i\right]\right|\\=&\left|M(i|j)M^{a-b}(j|i)-M(j|i)M^{a-b+1}(i|i)\right|\\ \le& M(i|j)\left|M^{a-b}(j|i)-\pi_j\right|\\&+M(j|i)\left|M^{a-b+1}(i|i)-\pi_i\right|\\ \le& \lambda_*^{a-b}M(i|j)\sqrt{\frac{\pi_j}{\pi_i}}+M(j|i)\lambda_*^{a-b+1}\\=&\lambda_*^{a-b}\sqrt{M(j|i)M(i|j)}+M(j|i)\lambda_*^{a-b+1}\\ \le& 2\sqrt{M(j|i)}\lambda_*^{a-b}.\end{aligned}$$

(ii)

$$\begin{aligned}&|h_2(a,b)-\pi_i M(j|i)(1-M(j|i))|\\=&\left|\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i,X_a=j\}}|X_b=i\right]-\pi_i M(j|i)\right.\\&+(M(j|i))^2\left(\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i\}}|X_b=i\right]-\pi_i\right)\\&\left.-2M(j|i)(\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i,X_a=j\}}|X_b=i\right]-\pi_i M(j|i))\right|\\ \le& \left|\mathbb{P}\left[X_{a+1}=i,X_a=j|X_b=i\right]-\pi_j M(i|j)\right|\\&+(M(j|i))^2\left|\mathbb{P}\left[X_{a+1}=i|X_b=i\right]-\pi_i\right|\\&+2M(j|i)\left|\mathbb{P}\left[X_{a+1}=i,X_a=j|X_b=i\right]-\pi_j M(i|j)\right|\\=&M(i|j)\left|M^{a-b}(j|i)-\pi_j\right|\\&+(M(j|i))^2\left|M^{a-b+1}(i|i)-\pi_i\right|\\&+2M(j|i)M(i|j)\left|M^{a-b}(j|i)-\pi_j\right|\\ \le& M(i|j)\sqrt{\frac{\pi_j}{\pi_i}}\lambda_*^{a-b}+(M(j|i))^2\lambda_*^{a-b+1}\\&+2M(j|i)M(i|j)\sqrt{\frac{\pi_j}{\pi_i}}\lambda_*^{a-b}\end{aligned}$$

$$\le \lambda_*^{a-b}\left(\sqrt{M(i|j)}\sqrt{\frac{M(i|j)\pi_j}{\pi_i}}+(M(j|i))^2\right.$$
$$\left.+2M(j|i)\sqrt{M(i|j)}\sqrt{\frac{M(i|j)\pi_j}{\pi_i}}\right)$$
$$\le 4\sqrt{M(j|i)}\lambda_*^{a-b}.$$

(iii) $h_3(a,b), h_4(a,b) \le h_2(a,b)$.

□

Proof of Lemma 16(i): For ease of notation we use c_0 to denote an absolute constant whose value may vary at each occurrence. Fix $i,j \in [k]$. Note that the empirical count defined in (4) can be written as $N_i = \sum_{a=1}^{n-1} \mathbf{1}_{\{X_{n-a}=i\}}$ and $N_{ij} = \sum_{a=1}^{n-1} \mathbf{1}_{\{X_{n-a}=i, X_{n-a+1}=j\}}$. Then

$$\mathbb{E}\left[(M(j|i)N_i - N_{ij})^2 | X_n = i\right]$$
$$= \mathbb{E}\left[\left(\sum_{a=1}^{n-1}\mathbf{1}_{\{X_{n-a}=i\}}\begin{pmatrix}\mathbf{1}_{\{X_{n-a+1}=j\}}\\ -M(j|i)\end{pmatrix}\right)^2 \middle| X_n = i\right]$$
$$\overset{(a)}{=} \mathbb{E}\left[\left(\sum_{a=1}^{n-1}\mathbf{1}_{\{X_{a+1}=i\}}\begin{pmatrix}\mathbf{1}_{\{X_a=j\}}\\ -M(j|i)\end{pmatrix}\right)^2 \middle| X_1 = i\right]$$
$$\overset{(b)}{=} \left|\sum_{a,b}\mathbb{E}\left[\eta_a\eta_b | X_1 = i\right]\right| \le 2\sum_{a\ge b}\left|\mathbb{E}\left[\eta_a\eta_b | X_1 = i\right]\right|,$$

where (a) is due to time reversibility; in (b) we defined $\eta_a \triangleq \mathbf{1}_{\{X_{a+1}=i\}}\left(\mathbf{1}_{\{X_a=j\}} - M(j|i)\right)$. We divide the summands into different cases and apply Lemma 37.

A. Case I: Two Distinct Indices

For any $a > b$, using Lemma 35 we get

$$\left|\mathbb{E}\left[\eta_a\eta_b | X_1 = i\right]\right| = \left|\mathbb{E}\left[\eta_a | X_{b+1} = i\right]\right|\left|\mathbb{E}\left[\eta_b | X_1 = 1\right]\right|$$
$$= h_1(a, b+1)h_1(b,1) \tag{160}$$

which implies

$$\sum_{n-1\ge a>b\ge 1}\sum\left|\mathbb{E}\left[\eta_a\eta_b | X_1 = i\right]\right|$$
$$= \sum_{n-1\ge a>b\ge 1}\sum h_1(a,b+1)h_1(b,1)$$
$$\lesssim M(j|i)\sum_{n-1\ge a>b\ge 1}\sum\lambda_*^{a-2} \lesssim \frac{M(j|i)}{\gamma_*^2}.$$

Here the last inequality (and similar sums in later deductions) can be explained as follows. Note that for $\gamma_* \ge \frac{1}{2}$ (i.e. $\lambda_* \le \frac{1}{2}$), the sum is clearly bounded by an absolute constant; for $\gamma_* < \frac{1}{2}$ (i.e. $\lambda_* > \frac{1}{2}$), we compare the sum with the mean (or higher moments in other calculations) of a geometric random variable.

B. Case II: Single Index

$$\sum_{a=1}^{n-1}\mathbb{E}\left[\eta_a^2 | X_1 = i\right] = \sum_{a=1}^{n-1} h_2(a,1)$$
$$\lesssim n\pi_i M(j|i)(1-M(j|i)) + \frac{\sqrt{M(j|i)}}{\gamma_*}. \tag{161}$$

Combining the above we get

$$\mathbb{E}\left[(N_{ij} - M(j|i)N_i)^2 | X_n = i\right]$$
$$\lesssim n\pi_i M(j|i)(1-M(j|i)) + \frac{\sqrt{M(j|i)}}{\gamma_*} + \frac{M(j|i)}{\gamma_*^2}$$

as required. □

Proof of Lemma 16(ii): We first note that due to reversibility we can write (similar as in proof of Lemma 16(i)) with $\eta_a = \mathbf{1}_{\{X_{a+1}=i\}}\left(\mathbf{1}_{\{X_a=j\}} - M(j|i)\right)$

$$\mathbb{E}\left[(M(j|i)N_i - N_{ij})^4 | X_n = i\right]$$
$$= \mathbb{E}\left[\left(\sum_{a=1}^{n-1}\mathbf{1}_{\{X_{a+1}=i\}}\left(\mathbf{1}_{\{X_a=j\}} - M(j|i)\right)\right)^4 \middle| X_1 = i\right]$$
$$= \left|\sum_{a,b,d,e}\mathbb{E}\left[\eta_a\eta_b\eta_d\eta_e | X_1 = i\right]\right|$$
$$\le \sum_{a,b,d,e}\left|\mathbb{E}\left[\eta_a\eta_b\eta_d\eta_e | X_1 = i\right]\right| \lesssim \sum_{a\ge b\ge d\ge e}\left|\mathbb{E}\left[\eta_a\eta_b\eta_d\eta_e | X_1 = i\right]\right|. \tag{162}$$

We bound the sum over different combinations of $a \ge b \ge d \ge e$ to come up with a bound on the required fourth moment. We first divide the η's into groups depending on how many distinct indices of η there are. We use the following identities which follow from Lemma 35: for indices $a > b > d > e$

- $\left|\mathbb{E}\left[\eta_a\eta_b\eta_d\eta_e | X_1 = i\right]\right| = h_1(a,b+1)h_1(b,d+1)h_1(d,e+1)h_1(e,1)$
- For $s_1, s_2, s_3 \in \{1,2\}$, $\left|\mathbb{E}\left[\eta_a^{s_1}\eta_b^{s_2}\eta_d^{s_3} | X_1 = i\right]\right| = h_{s_1}(a,b+1)h_{s_2}(b,d+1)h_{s_3}(d,1)$
- For $s_1, s_2 \in \{1,2,3\}$, $\left|\mathbb{E}\left[\eta_a^{s_1}\eta_b^{s_2} | X_1 = i\right]\right| = h_{s_1}(a,b+1)h_{s_2}(b,1)$
- $\mathbb{E}\left[\eta_a^4 | X_1 = 1\right] = h_4(a,1)$

and then use Lemma 37 to bound the h functions.

C. Case I: Four Distinct Indices

Using Lemma 37 we have

$$\sum_{n-1\ge a>b>d>e\ge 1}\sum\left|\mathbb{E}\left[\eta_a\eta_b\eta_d\eta_e | X_1 = i\right]\right|$$
$$= \sum_{n-1\ge a>b>d>e\ge 1}\sum h_1(a,b+1)h_1(b,d+1)h_1(d,e+1)h_1(e,1)$$
$$\le M(j|i)^2\sum_{n-1\ge a>b>d>e\ge 1}\sum\lambda_*^{a-4} \lesssim \frac{M(j|i)^2}{\gamma_*^4}.$$

D. Case II: Three Distinct Indices

There are three cases, namely $\eta_a^2\eta_b\eta_d, \eta_a\eta_b^2\eta_d$ and $\eta_a\eta_b\eta_d^2$.

1) Bounding $\sum\sum\sum_{n-1\geq a>b>d\geq 1}\left|\mathbb{E}\left[\eta_a^2\eta_b\eta_d|X_1=i\right]\right|$:

$$\begin{aligned}
&\sum_{n-1\geq a>b>d\geq 1}\sum\sum\left|\mathbb{E}\left[\eta_a^2\eta_b\eta_d|X_1=i\right]\right|\\
&=\sum_{n-1\geq a>b>d\geq 1}\sum\sum h_2(a,b+1)h_1(b,d+1)h_1(d,1)\\
&\lesssim \sum_{n-1\geq a>b>d\geq 1}\sum\sum (\pi_i M(j|i)(1-M(j|i))\\
&\qquad\qquad +\sqrt{M(j|i)}\lambda_*^{a-b-1}\Big)M(j|i)\lambda_*^{b-2}\\
&\lesssim \frac{M(j|i)}{\gamma_*^2}n\pi_i M(j|i)(1-M(j|i))+\frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3}\\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2+\frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3}+\frac{M(j|i)^2}{\gamma_*^4}
\end{aligned}$$

where the last inequality followed by using $xy\leq x^2+y^2$.

2) Bounding $\sum\sum\sum_{n-2\geq a>b>d\geq 1}\left|\mathbb{E}\left[\eta_a\eta_b^2\eta_d|X_1=i\right]\right|$:

$$\begin{aligned}
&\sum_{n-2\geq a>b>d\geq 1}\sum\sum\left|\mathbb{E}\left[\eta_a\eta_b^2\eta_d|X_1=i\right]\right|\\
&=\sum_{n-2\geq a>b>d\geq 1}\sum\sum h_1(a,b+1)h_2(b,d+1)h_1(d,1)\\
&\lesssim \sum_{n-2\geq a>b>d\geq 1}\sum\sum (\pi_i M(j|i)(1-M(j|i))\\
&\quad +\sqrt{M(j|i)}\lambda_*^{b-d-1}\Big)M(j|i)\lambda_*^{a-b+d-2}\\
&\lesssim \frac{M(j|i)}{\gamma_*^2}n\pi_i M(j|i)(1-M(j|i))+\frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3}\\
&\lesssim n\pi_i M(j|i)(1-M(j|i))^2+\frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3}+\frac{M(j|i)^2}{\gamma_*^4}.
\end{aligned}$$

3) Bounding $\sum\sum\sum_{n-2\geq a>b>d\geq 1}\left|\mathbb{E}\left[\eta_a\eta_b\eta_d^2|X_1=i\right]\right|$:

$$\begin{aligned}
&\sum_{n-2\geq a>b>d\geq 1}\sum\sum\left|\mathbb{E}\left[\eta_a\eta_b\eta_d^2|X_1=i\right]\right|\\
&=\sum_{n-2\geq a>b>d\geq 1}\sum\sum h_1(a,b+1)h_1(b,d+1)h_2(d,1)\\
&\lesssim \sum_{n-2\geq a>b>d\geq 1}\sum\sum (\pi_i M(j|i)(1-M(j|i))\\
&\quad +\sqrt{M(j|i)}\lambda_*^{d-1}\Big)M(j|i)\lambda_*^{a-d-2}\\
&\lesssim \frac{M(j|i)}{\gamma_*^2}n\pi_i M(j|i)(1-M(j|i))+\frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3}\\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2+\frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3}+\frac{M(j|i)^2}{\gamma_*^4}.
\end{aligned}$$

E. Case III: Two Distinct Indices

There are three different cases, namely $\eta_a^2\eta_b^2$, $\eta_a^3\eta_b$ and $\eta_a\eta_b^3$.

1) Bounding $\sum\sum_{n-2\geq a>b\geq 1}\left|\mathbb{E}\left[\eta_a^2\eta_b^2|X_1=i\right]\right|$:

$$\begin{aligned}
&\sum_{n-2\geq a>b\geq 1}\sum\mathbb{E}\left[\eta_a^2\eta_b^2|X_1=i\right]\\
&=\sum_{n-2\geq a>b\geq 1}\sum h_2(a,b+1)h_2(b,1)\\
&\lesssim \sum_{n-2\geq a>b\geq 1}\sum\Big(\pi_i M(j|i)(1-M(j|i))+\sqrt{M(j|i)}\lambda_*^{a-b-1}\Big)\\
&\quad\Big(\pi_i M(j|i)(1-M(j|i))+\sqrt{M(j|i)}\lambda_*^{b-1}\Big)\\
&\lesssim \sum_{n-2\geq a>b\geq 1}\sum\Big\{\pi_i M(j|i)(1-M(j|i))\sqrt{M(j|i)}(\lambda_*^{a-b-1}+\lambda_*^{b-1})\\
&\quad +(\pi_i M(j|i)(1-M(j|i)))^2+M(j|i)\lambda_*^{a-2}\Big\}\\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2\\
&\quad +\frac{\sqrt{M(j|i)}}{\gamma_*}n\pi_i M(j|i)(1-M(j|i))+\frac{M(j|i)}{\gamma_*^2}\\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2+\frac{M(j|i)}{\gamma_*^2}.
\end{aligned}$$

2) Bounding $\sum\sum_{n-2\geq a>b\geq 1}\left|\mathbb{E}\left[\eta_a^3\eta_b|X_1=i\right]\right|$:

$$\begin{aligned}
&\sum_{n-2\geq a>b\geq 1}\sum\left|\mathbb{E}\left[\eta_a^3\eta_b|X_1=i\right]\right|\\
&=\sum_{n-2\geq a>b\geq 1}\sum h_3(a,b+1)h_1(b,1)\\
&\lesssim \sum_{n-2\geq a>b\geq 1}\sum\Big(\pi_i M(j|i)(1-M(j|i))+\sqrt{M(j|i)}\lambda_*^{a-b-1}\Big)\\
&\qquad\qquad \sqrt{M(j|i)}\lambda_*^{b-1}\\
&\lesssim \frac{\sqrt{M(j|i)}}{\gamma_*}n\pi_i M(j|i)(1-M(j|i))+\frac{M(j|i)}{\gamma_*^2}\\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2+\frac{M(j|i)}{\gamma_*^2}.
\end{aligned}$$

3) Bounding $\sum\sum_{n-2\geq a>b\geq 1}\left|\mathbb{E}\left[\eta_a\eta_b^3|X_1=i\right]\right|$:

$$\begin{aligned}
&\sum_{n-2\geq a>b\geq 1}\sum\left|\mathbb{E}\left[\eta_a\eta_b^3|X_1=i\right]\right|\\
&=\sum_{n-2\geq a>b\geq 1}\sum h_1(a,b+1)h_3(b,1)\\
&\lesssim \sum_{n-2\geq a>b\geq 1}\sum\Big(\pi_i M(j|i)(1-M(j|i))+\sqrt{M(j|i)}\lambda_*^{b-1}\Big)\\
&\qquad\qquad \sqrt{M(j|i)}\lambda_*^{a-b-1}\\
&\lesssim \frac{\sqrt{M(j|i)}}{\gamma_*}n\pi_i M(j|i)(1-M(j|i))+\frac{M(j|i)}{\gamma_*^2}\\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2+\frac{M(j|i)}{\gamma_*^2}.
\end{aligned}$$

F. Case IV: Single Index

Bounding $\sum_{a=1}^{n-1}\mathbb{E}[\eta_a^4|X_1=i]$:

$$\begin{aligned}
&\sum_{a=1}^{n-1}\mathbb{E}\left[\eta_a^4|X_1=i\right]\\
&=\sum_{a=1}^{n-1}h_4(a,1)\leq n\pi_i M(j|i)(1-M(j|i))+\frac{\sqrt{M(j|i)}}{\gamma_*}.
\end{aligned}$$

Combining all cases we get

$$\begin{aligned}
&\mathbb{E}\left[(M(j|i)N_i - N_{ij})^4 \,|X_n = i\right] \\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2 + \frac{\sqrt{M(j|i)}}{\gamma_*} + \frac{M(j|i)}{\gamma_*^2} \\
&\quad + \frac{M(j|i)^{\frac{3}{2}}}{\gamma_*^3} + \frac{M(j|i)^2}{\gamma_*^4} \\
&\lesssim (n\pi_i M(j|i)(1-M(j|i)))^2 + \frac{\sqrt{M(j|i)}}{\gamma_*} + \frac{M(j|i)^2}{\gamma_*^4}
\end{aligned}$$

as required. □

Proof of Lemma 16(iii): Throughout our proof we repeatedly use the spectral decomposition (159) applied to the diagonal elements:

$$M^t(i|i) = \pi_i + \sum_{v\geq 2} \lambda_v^t u_{vi}^2, \quad \sum_{v\geq 2} u_{vi}^2 \leq 1.$$

Write $N_i - (n-1)\pi_i = \sum_{a=1}^{n-1} \xi_a$ where $\xi_a = \mathbf{1}_{\{X_a=i\}} - \pi_i$. For $a \geq b \geq d \geq e$,

$$\begin{aligned}
&\mathbb{E}\left[\xi_a\xi_b\xi_d\xi_e|X_1 = i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b \begin{pmatrix} \mathbf{1}_{\{X_d=i,X_e=i\}} - \pi_i \mathbf{1}_{\{X_d=i\}} \\ -\pi_i \mathbf{1}_{\{X_e=i\}} + \pi_i^2 \end{pmatrix} \middle| X_1 = i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b \mathbf{1}_{\{X_d=i,X_e=i\}}|X_1 = i\right] \\
&\quad - \pi_i \mathbb{E}\left[\xi_a\xi_b \mathbf{1}_{\{X_d=i\}}|X_1 = i\right] \\
&\quad - \pi_i \mathbb{E}\left[\xi_a\xi_b \mathbf{1}_{\{X_e=i\}}|X_1 = i\right] + \pi_i^2 \mathbb{E}\left[\xi_a\xi_b|X_1 = i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b|X_d = i\right] \mathbb{P}\left[X_d = i|X_e = i\right] \mathbb{P}[X_e = i|X_1 = i] \\
&\quad - \pi_i \mathbb{E}\left[\xi_a\xi_b|X_d = i\right] \mathbb{P}[X_d = i|X_1 = i] \\
&\quad - \pi_i \mathbb{E}\left[\xi_a\xi_b|X_e = i\right] \mathbb{P}[X_e = i|X_1 = i] \\
&\quad + \pi_i^2 \mathbb{E}\left[\xi_a\xi_b|X_1 = i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b|X_d = i\right] \left\{M^{d-e}(i|i)M^{e-1}(i|i) - \pi_i M^{d-1}(i|i)\right\} \\
&\quad - \left\{\pi_i \mathbb{E}\left[\xi_a\xi_b|X_e = i\right] M^{e-1}(i|i) - \pi_i^2 \mathbb{E}\left[\xi_a\xi_b|X_1 = i\right]\right\}
\end{aligned} \tag{163}$$

Using the Markov property for any $d \leq b \leq a$, we get

$$\begin{aligned}
&\left|\mathbb{E}[\xi_a\xi_b|X_d = i] - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right| \\
&= \left|\mathbb{E}\left[\mathbf{1}_{\{X_a=i,X_b=i\}} - \pi_i \mathbf{1}_{\{X_a=i\}} - \pi_i \mathbf{1}_{\{X_b=i\}} + \pi_i^2|X_d = i\right]\right. \\
&\quad \left. -\pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right| \\
&= \left|M^{a-b}(i|i)M^{b-d}(i|i) - \pi_i M^{a-d}(i|i) - \pi_i M^{b-d}(i|i)\right. \\
&\quad \left. +\pi_i^2 - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right| \\
&= \left|\left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right)\left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{b-d}\right)\right. \\
&\quad \left. -\pi_i \left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-d}\right) - \pi_i \left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{b-d}\right)\right. \\
&\quad \left. +\pi_i^2 - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right| \\
&= \left|\left(\sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right)\left(\sum_{v\geq 2} u_{vi}^2 \lambda_v^{b-d}\right) - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-d}\right| \\
&\leq \lambda_*^{a-d}\left(\sum_{v\geq 2} u_{vi}^2\right)\left(\sum_{v\geq 2} u_{vi}^2\right) + \lambda_*^{a-d}\pi_i \sum_{v\geq 2} u_{vi}^2 \leq 2\lambda_*^{a-d}.
\end{aligned} \tag{164}$$

We also get for $d \geq e$

$$\begin{aligned}
&\left|M^{d-e}(i|i)M^{e-1}(i|i) - \pi_i M^{d-1}(i|i)\right| \\
&= \left|\left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{d-e}\right)\left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{e-1}\right)\right. \\
&\quad \left. -\pi_i \left(\pi_i + \sum_{v\geq 2} u_{vi}^2 \lambda_v^{d-1}\right)\right| \\
&= \left|\pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{e-1} + \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{d-e}\right. \\
&\quad \left. + \left(\sum_{v\geq 2} u_{vi}^2 \lambda_v^{e-1}\right)\left(\sum_{v\geq 2} u_{vi}^2 \lambda_v^{d-e}\right) - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{d-1}\right| \\
&\leq 2\lambda_*^{d-1} + \pi_i \lambda_*^{e-1} + \pi_i \lambda_*^{d-e}.
\end{aligned} \tag{165}$$

This implies

$$\begin{aligned}
&\left|\mathbb{E}\left[\xi_a\xi_b|X_d = i\right]\right| \left|M^{d-e}(i|i)M^{e-1}(i|i) - \pi_i M^{d-1}(i|i)\right| \\
&\leq \left(\pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b} + 2\lambda_*^{a-d}\right)\left(2\lambda_*^{d-1} + \pi_i \lambda_*^{e-1} + \pi_i \lambda_*^{d-e}\right) \\
&\leq \left(\pi_i \lambda_*^{a-b} + 2\lambda_*^{a-d}\right)\left(2\lambda_*^{d-1} + \pi_i \lambda_*^{e-1} + \pi_i \lambda_*^{d-e}\right) \\
&\leq 4\left[\pi_i^2 \lambda_*^{a-b+d-e} + \pi_i^2 \lambda_*^{a-b+e-1}\right. \\
&\quad \left. +\pi_i\left(\lambda_*^{a-b+d-1} + \lambda_*^{a-d+e-1} + \lambda_*^{a-e}\right) + \lambda_*^{a-1}\right]
\end{aligned} \tag{166}$$

Using (164) along with Lemma 36 for any $e \leq b \leq a$ we get

$$\begin{aligned}
&\left|\pi_i \mathbb{E}\left[\xi_a\xi_b|X_e = i\right] M^{e-1}(i|i) - \pi_i^2 \mathbb{E}\left[\xi_a\xi_b|X_1 = i\right]\right| \\
&\leq \pi_i \left|\mathbb{E}\left[\xi_a\xi_b|X_e = i\right]\right| \left|M^{e-1}(i|i) - \pi_i\right| \\
&\quad + \pi_i^2 \left|\mathbb{E}\left[\xi_a\xi_b|X_e = i\right] - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right| \\
&\quad + \pi_i^2 \left|\mathbb{E}\left[\xi_a\xi_b|X_1 = i\right] - \pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b}\right| \\
&\leq \pi_i \left[\pi_i \sum_{v\geq 2} u_{vi}^2 \lambda_v^{a-b} + 2\lambda_*^{a-e}\right] 2\lambda_*^{e-1} + 2\pi_i^2 \lambda_*^{a-e} + 2\pi_i^2 \lambda_*^{a-1} \\
&\leq 2\pi_i^2 \lambda_*^{a-b+e-1} + 4\pi_i^2 \lambda_*^{a-e} + 4\pi_i^2 \lambda_*^{a-1}.
\end{aligned} \tag{167, 168}$$

This together with (166) and (163) implies

$$\begin{aligned}
\left|\mathbb{E}\left[\xi_a\xi_b\xi_d\xi_e|X_1 = i\right]\right| &\lesssim \pi_i^2\left(\lambda_*^{a-b+d-e} + \lambda_*^{a-b+e-1}\right) + \lambda_*^{a-1} \\
&\quad + \pi_i\left(\lambda_*^{a-b+d-1} + \lambda_*^{a-d+e-1} + \lambda_*^{a-e}\right).
\end{aligned} \tag{169}$$

To bound the sum over $n-1 \geq a \geq b \geq d \geq e \geq 1$, we divide the analysis according to the number of distinct ordered indices related variations in terms.

G. Case I: four Distinct Indices

We sum (169) over all possible $a > b > d > e$.

- For the first term,

$$\pi_i^2 \sum_{n-1\geq a>b>d>e\geq 1} \lambda_*^{a-b+d-e} \lesssim \frac{n\pi_i^2}{\gamma_*} \sum_{n-1\geq a>b\geq 3} \lambda_*^{a-b} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2}.$$

- For the second term,

$$\pi_i^2 \sum_{n-1\geq a>b>d>e\geq 1} \lambda_*^{a-b+e-1} \lesssim \frac{n\pi_i^2}{\gamma_*} \sum_{n-1\geq a>b\geq 3} \lambda_*^{a-b} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2}.$$

- For the third term,

$$\sum_{n-1\geq a>b>d>e\geq 1} \lambda_*^{a-1} \lesssim \sum_{n-1\geq a\geq 4} a^3\lambda_*^{a-1} \lesssim \frac{1}{\gamma_*^4}.$$

- For the fourth term,

$$\pi_i \sum_{n-1\geq a>b>d>e\geq 1} \lambda_*^{a-b+d-1} \leq \frac{\pi_i}{\gamma_*^2} \sum_{n-1\geq a>b\geq 3} \lambda_*^{a-b} \lesssim \frac{n\pi_i}{\gamma_*^3}.$$

- For the fifth term,

$$\pi_i \sum_{n-1\geq a>b>d>e\geq 1} \lambda_*^{a-d+e-1} \lesssim \frac{\pi_i}{\gamma_*} \left(\sum_{n-1\geq a>b\geq 3} \lambda_*^{a-b}\right) \left(\sum_{d\geq 2}^{b-1} \lambda_*^{b-d}\right) \lesssim \frac{n\pi_i}{\gamma_*^3}.$$

- For the sixth term,

$$\pi_i \sum_{n-1\geq a>b>d>e\geq 1} \lambda_*^{a-e} \lesssim \pi_i \left(\sum_{n-1\geq a>b\geq 3} \lambda_*^{a-b}\right) \left(\sum_{d\geq 2}^{b-1} \lambda_*^{b-d}\right) \left(\sum_{e\geq 1}^{d-1} \lambda_*^{d-e}\right) \lesssim \frac{n\pi_i}{\gamma_*^3}.$$

Combining the above bounds and using the fact that $ab \leq a^2 + b^2$, we obtain

$$\sum_{n-1\geq a>b>d>e\geq 1} |\mathbb{E}\left[\xi_a\xi_b\xi_d\xi_e|X_1 = i\right]| \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{n\pi_i}{\gamma_*^3} + \frac{1}{\gamma_*^4} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{1}{\gamma_*^4}. \tag{170}$$

H. Case II: Three Distinct Indices

There are three cases, namely, $\xi_a\xi_b^2\xi_e$, $\xi_a\xi_b\xi_e^2$, and $\xi_a^2\xi_b\xi_e$.

1) Bounding $\sum\sum\sum_{n-1\geq a>b>e\geq 1} \left|\mathbb{E}\left[\xi_a\xi_b^2\xi_e|X_1 = i\right]\right|$: We specialize (169) with $b = d$ to get

$$\left|\mathbb{E}\left[\xi_a\xi_b^2\xi_e|X_1 = i\right]\right| \lesssim \pi_i\left(\lambda_*^{a-b+e-1} + \lambda_*^{a-e}\right) + \lambda_*^{a-1}.$$

Summing over a, b, e we have

$$\begin{aligned}
&\sum_{n-1\geq a>b>e\geq 1} \left|\mathbb{E}\left[\xi_a\xi_b^2\xi_e|X_1 = i\right]\right| \\
&\lesssim \sum_{n-1\geq a>b>e\geq 1} \left\{\pi_i\left(\lambda_*^{a-b+e-1} + \lambda_*^{a-e}\right) + \lambda_*^{a-1}\right\} \\
&\lesssim \frac{\pi_i}{\gamma_*} \sum_{n-1\geq a>b\geq 2} \lambda_*^{a-b} \\
&\quad + \pi_i \left(\sum_{n-1\geq a>b\geq 2} \lambda_*^{a-b}\right) \left(\sum_{e\geq 1}^{b-1} \lambda_*^{b-e}\right) + \sum_{n-1\geq a\geq 3} a^3\lambda_*^{a-1} \\
&\lesssim \frac{n\pi_i}{\gamma_*^2} + \frac{1}{\gamma_*^3} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{1}{\gamma_*^3}
\end{aligned} \tag{171}$$

with last inequality following from $xy \leq x^2 + y^2$.

2) Bounding $\sum\sum\sum_{n-1\geq a>b>e\geq 1} \left|\mathbb{E}\left[\xi_a\xi_b\xi_e^2|X_1 = i\right]\right|$: We specialize (169) with $e = d$ to get

$$\begin{aligned}
&\left|\mathbb{E}\left[\xi_a\xi_b\xi_e^2|X_1 = i\right]\right| \\
&\lesssim \pi_i^2\lambda_*^{a-b} + \pi_i\left(\lambda_*^{a-b+e-1} + \lambda_*^{a-e}\right) + \lambda_*^{a-1}.
\end{aligned}$$

Summing over a, b, e and applying (171), we get

$$\begin{aligned}
&\sum_{n-1\geq a>b>e\geq 1} \left|\mathbb{E}\left[\xi_a\xi_b\xi_e^2|X_1 = i\right]\right| \\
&\lesssim \sum_{n-1\geq a>b>e\geq 1} \left\{\begin{array}{c}\pi_i^2\lambda_*^{a-b} + \lambda_*^{a-1} \\ +\pi_i\left(\lambda_*^{a-b+e-1} + \lambda_*^{a-e}\right)\end{array}\right\} \\
&\lesssim n\pi_i^2 \sum_{n-1\geq a>b\geq 2} \lambda_*^{a-b} + \frac{n\pi_i}{\gamma_*^2} + \frac{1}{\gamma_*^3} \\
&\lesssim \frac{n^2\pi_i^2}{\gamma_*} + \frac{n\pi_i}{\gamma_*^2} + \frac{1}{\gamma_*^3} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{1}{\gamma_*^3}.
\end{aligned} \tag{172}$$

3) Bounding $\sum\sum\sum_{n-1\geq a>b>e\geq 1} \left|\mathbb{E}\left[\xi_a^2\xi_b\xi_e|X_1 = i\right]\right|$: Specializing (169) with $a = b$ we get

$$\begin{aligned}
&\left|\mathbb{E}\left[\xi_b^2\xi_d\xi_e|X_1 = i\right]\right| \\
&\lesssim \pi_i^2\left(\lambda_*^{d-e} + \lambda_*^{e-1}\right) \\
&\quad + \lambda_*^{b-1} + \pi_i\left(\lambda_*^{d-1} + \lambda_*^{b-d+e-1} + \lambda_*^{b-e}\right),
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\left|\mathbb{E}\left[\xi_a^2\xi_b\xi_e|X_1 = i\right]\right| \\
&\lesssim \pi_i^2\left(\lambda_*^{b-e} + \lambda_*^{e-1}\right) \\
&\quad + \lambda_*^{a-1} + \pi_i\left(\lambda_*^{b-1} + \lambda_*^{a-b+e-1} + \lambda_*^{a-e}\right).
\end{aligned}$$

For the first, second and fourth terms

$$\sum_{n-1\geq a>b>e\geq 1} \left\{\pi_i^2\left(\lambda_*^{b-e} + \lambda_*^{e-1}\right) + \pi_i\lambda_*^{b-1}\right\} \lesssim \frac{\pi_i^2}{\gamma_*} \sum_{n-1\geq a>b\geq 2} 1 + \frac{n\pi_i}{\gamma_*^2} \lesssim \frac{n^2\pi_i^2}{\gamma_*} + \frac{n\pi_i}{\gamma_*^2},$$

and for summing the remaining terms we use (171), which implies

$$\sum_{n-1\geq a>b>e\geq 1} \left|\mathbb{E}\left[\xi_a^2\xi_b\xi_e|X_1 = i\right]\right| \lesssim \frac{n^2\pi_i^2}{\gamma_*} + \frac{n\pi_i}{\gamma_*^2} + \frac{1}{\gamma_*^3} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{1}{\gamma_*^3}. \tag{173}$$

I. Case III: Two Distinct Indices

There are three cases, namely, $\eta_a^2\eta_e^2, \eta_a\eta_e^3$ and $\eta_a^3\eta_e$.

1) Bounding $\sum\sum_{n-1\geq a>e\geq 1}\mathbb{E}\left[\xi_a^2\xi_e^2|X_1=i\right]$: Specializing (169) for $a=b$ and $e=d$ we get

$$\mathbb{E}\left[\xi_a^2\xi_e^2|X_1=i\right]\lesssim \pi_i^2+\pi_i\left(\lambda_*^{e-1}+\lambda_*^{a-e}\right)+\lambda_*^{a-1}.$$

Summing up over a,e we have

$$\begin{aligned}&\sum_{n-1\geq a>e\geq 1}\sum \mathbb{E}\left[\xi_a^2\xi_e^2|X_1=i\right]\\ &\lesssim \sum_{n-1\geq a>e\geq 1}\sum\left\{\pi_i^2+\pi_i\left(\lambda_*^{e-1}+\lambda_*^{a-e}\right)+\lambda_*^{a-1}\right\}\\ &\lesssim n^2\pi_i^2+\frac{n\pi_i}{\gamma_*}+\frac{1}{\gamma_*^2}.\end{aligned}\tag{174}$$

2) Bounding $\sum\sum_{n-1\geq a>e\geq 1}\left|\mathbb{E}\left[\xi_a\xi_e^3|X_1=i\right]\right|$: Specializing (169) for $e=b=d$ we get

$$\left|\mathbb{E}\left[\xi_a\xi_e^3|X_1=i\right]\right|\lesssim \pi_i\lambda_*^{a-e}+\lambda_*^{a-1}$$

which sums up to

$$\begin{aligned}&\sum_{n-1\geq a>e\geq 1}\sum\left|\mathbb{E}\left[\xi_a\xi_e^3|X_1=i\right]\right|\\ &\lesssim \pi_i\sum_{n-1\geq a>e\geq 1}\sum\lambda_*^{a-e}+\sum_{n-1\geq a>e\geq 1}\sum\lambda_*^{a-1}\lesssim\frac{n\pi_i}{\gamma_*}+\frac{1}{\gamma_*^2}.\end{aligned}\tag{175}$$

3) Bounding $\sum\sum_{n-1\geq a>e\geq 1}\left|\mathbb{E}\left[\xi_a^3\xi_e|X_1=i\right]\right|$: Specializing (169) for $a=b=d$ we get

$$\left|\mathbb{E}\left[\xi_a^3\xi_e|X_1=i\right]\right|\lesssim \pi_i\left(\lambda_*^{a-e}+\lambda_*^{e-1}\right)+\lambda_*^{a-1}$$

which sums up to

$$\begin{aligned}&\sum_{n-1\geq a>e\geq 1}\sum\left|\mathbb{E}\left[\xi_a^3\xi_e|X_1=i\right]\right|\\ &\lesssim \sum_{n-1\geq a>e\geq 1}\sum\left\{\pi_i\left(\lambda_*^{a-e}+\lambda_*^{e-1}\right)+\lambda_*^{a-1}\right\}\lesssim\frac{n\pi_i}{\gamma_*}+\frac{1}{\gamma_*^2}.\end{aligned}\tag{176}$$

J. Case IV: Single Distinct Index

We specialize (169) to $a=b=d=e$ to get

$$\mathbb{E}\left[\xi_a^4|X_1=i\right]\lesssim\pi_i+\lambda_*^{a-1}.$$

Summing the above over a

$$\sum_{a=1}^{n-1}\mathbb{E}\left[\xi_a^4|X_1=i\right]\lesssim n\pi_i+\frac{1}{\gamma_*}.\tag{177}$$

Combining (170)–(177) and using $\frac{n\pi_i}{\gamma_*}\lesssim\frac{n^2\pi_i^2}{\gamma_*^2}+\frac{1}{\gamma_*^4}$, we get

$$\mathbb{E}\left[\left(N_i-(n-1)\pi_i\right)^4|X_1=i\right]\lesssim\frac{n^2\pi_i^2}{\gamma_*^2}+\frac{1}{\gamma_*^4}.$$

□

ACKNOWLEDGMENT

The authors are grateful to Alon Orlitsky for helpful and encouraging comments and to Dheeraj Pichapati for providing the full version of [1]. They also thank David Pollard for insightful discussions on Markov chains at the initial stages of the project.

REFERENCES

[1] M. Falahatgar, A. Orlitsky, V. Pichapati, and A. T. Suresh, "Learning Markov distributions: Does estimation Trump compression?" in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 2689–2693.
[2] Y. Hao, A. Orlitsky, and V. Pichapati, "On learning Markov chains," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 648–657.
[3] M. S. Bartlett, "The frequency goodness of fit test for probability chains," *Math. Proc. Cambridge Philos. Soc.*, vol. 47, no. 1, pp. 86–95, 1951.
[4] T. W. Anderson and L. A. Goodman, "Statistical inference about Markov chains," *Ann. Math. Statist.*, vol. 28, no. 1, pp. 89–110, Mar. 1957.
[5] P. Billingsley, "Statistical methods in Markov chains," *Ann. Math. Statist.*, vol. 32, no. 1, pp. 12–40, Mar. 1961.
[6] G. Wolfer and A. Kontorovich, "Minimax learning of ergodic Markov chains," in *Proc. 30th Int. Conf. Algorithmic Learn. Theory*, 2019, pp. 904–930.
[7] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Ann. Statist.*, vol. 28, no. 6, pp. 1601–1619, 2000.
[8] S. Kamath and S. Verdu, "Estimation of entropy rate and Rényi entropy rate for Markov chains," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 685–689.
[9] Y. Han, J. Jiao, C. Lee, T. Weissman, Y. Wu, and T. Yu, "Entropy rate estimation for Markov chains with large state space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9781–9792.
[10] D. Hsu, A. Kontorovich, D. A. Levin, Y. Peres, C. Szepesvári, and G. Wolfer, "Mixing time estimation in reversible Markov chains from a single sample path," *Ann. Appl. Probab.*, vol. 29, no. 4, pp. 2439–2480, Aug. 2019.
[11] D. Braess, J. Forster, T. Sauer, and H. U. Simon, "How to achieve minimax expected Kullback–Leibler distance from an unknown finite distribution," in *Proc. 13th Int. Conf. Algorithmic Learn. Theory (ALT)*, Lübeck, Germany. Berlin, Germany: Springer, Nov. 2002, pp. 380–394.
[12] L. Paninski, "Variational minimax estimation of discrete distributions under KL loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004, pp. 1033–1040.
[13] S. Kamath, A. Orlitsky, D. Pichapati, and A. Suresh, "On learning distributions from their samples," in *Proc. Conf. Learn. Theory*, Jun. 2015, pp. 1066–1100.
[14] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, Oct. 1999.
[15] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Proc. Conf. Learn. Theory*, 2018, pp. 439–473.
[16] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Found. Comput. Math.*, vol. 20, no. 4, pp. 1–47, 2019.
[17] D. Levin and Y. Peres, *Markov Chains and Mixing Times*, vol. 107. Providence, RI, USA: American Mathematical Society, 2017.
[18] D. Paulin, "Concentration inequalities for Markov chains by Marton couplings and spectral methods," *Electron. J. Probab.*, vol. 20, pp. 1–20, Jan. 2015.
[19] P. Lezaud, "Chernoff-type bound for finite Markov chains," *Ann. Appl. Probab.*, vol. 8, no. 3, pp. 849–867, Aug. 1998.
[20] P. Whittle, "Some distribution and moment formulae for the Markov chain," *J. Roy. Stat. Soc., B Methodol.*, vol. 17, no. 2, pp. 235–242, Jul. 1955.
[21] L. Davisson, R. McEliece, M. Pursley, and M. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 3, pp. 269–279, May 1981.
[22] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Found. Trends Commun. Inf. Theory*, vol. 1, no. 4, pp. 417–527, 2004.
[23] J. Mardia, J. Jiao, E. Tánczos, R. D. Nowak, and T. Weissman, "Concentration inequalities for the empirical distribution of discrete distributions: Beyond the method of types," *Inf. Inference, A J. IMA*, vol. 9, no. 4, pp. 813–850, Dec. 2020.
[24] R. Agrawal, "Finite-sample concentration of the multinomial in relative entropy," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6297–6302, Oct. 2020.
[25] F. R. Guo and T. S. Richardson, "Chernoff-type concentration of empirical probabilities in relative entropy," *IEEE Trans. Inf. Theory*, vol. 67, no. 1, pp. 549–558, Jan. 2021.
[26] K. Tatwawadi, J. Jiao, and T. Weissman, "Minimax redundancy for Markov chains with large state space," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 216–220.

[27] L. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 2, pp. 211–215, Mar. 1983.
[28] E. Parzen, *Stochastic Processes*. Toronto, ON, Canada: Holden Day, 1962.
[29] R. Montenegro and P. Tetali, "Mathematical aspects of mixing times in Markov chains," *Found. Trends Theor. Comput. Sci.*, vol. 1, no. 3, pp. 237–354, 2006.
[30] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.
[31] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.
[32] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems Inf. Transmiss.*, vol. 23, no. 3, pp. 175–186, 1987.
[33] B. Ryabko, "Prediction of random sequences and universal coding," *Prob. Pered. Inf.*, vol. 24, no. 2, pp. 87–96, 1988.
[34] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2104–2109, Sep. 1999.
[35] P. Jacquet, and W. Szpankowski, "A combinatorial problem arising in information theory: Precise minimax redundancy for Markov sources," in *Mathematics and Computer Science II: Algorithms, Trees, Combinatorics and Probabilities*. Basel, Switzerland: Birkhäuser, 2002, pp. 311–328.
[36] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *Ann. Math. Statist.*, vol. 35, no. 2, pp. 876–879, Jun. 1964.
[37] M. Obremski and M. Skorski, "Complexity of estimating Rényi entropy of Markov chains," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2264–2269.
[38] A. Ben-Hamou, R. I. Oliveira, and Y. Peres, "Estimating graph parameters via random walks with restarts," in *Proc. 29th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, 2018, pp. 1702–1714.
[39] C. Daskalakis, N. Dikkala, and N. Gravin, "Testing symmetric Markov chains from a single trajectory," in *Proc. Conf. Learn. Theory*, 2018, pp. 385–409.
[40] Y. Cherapanamjeri and P. L. Bartlett, "Testing symmetric Markov chains without hitting," in *Proc. Conf. Learn. Theory*, 2019, pp. 758–785.
[41] G. Wolfer and A. Kontorovich, "Minimax testing of identity to a reference ergodic Markov chain," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 191–201.
[42] S. Fried and G. Wolfer, "Identity testing of reversible Markov chains," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 798–817.
[43] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY, USA: Academic Press, 1982.
[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2006.
[45] J. H. B. Kemperman, "On the Shannon capacity of an arbitrary channel," in *Proc. Indagationes Mathematicae*, vol. 77, no. 2, 1974, pp. 101–115.
[46] V. K. Trofimov, "Redundancy of universal coding of arbitrary Markov sources," *Prob. Pered. Inf.*, vol. 10, no. 4, pp. 16–24, 1974.
[47] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, Mar. 1997.
[48] W. Szpankowski and M. J. Weinberger, "Minimax pointwise redundancy for memoryless models over large alphabets," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4094–4104, Jul. 2012.
[49] F. Liang and A. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2708–2726, Nov. 2004.
[50] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[51] S. Janson, "On concentration of probability," *Contemp. Combinatorics*, vol. 10, no. 3, pp. 1–9, May 2002.
[52] T. D. Ahle, "Sharp and simple bounds for the raw moments of the binomial and Poisson distributions," *Statist. Probab. Lett.*, vol. 182, Mar. 2022, Art. no. 109306.
[53] R. Latała, "Estimation of moments of sums of independent real random variables," *Ann. Probab.*, vol. 25, no. 3, pp. 1502–1513, Jul. 1997.
[54] A. Hoffman, "Three observations on nonnegative matrices," *J. Res. Nat. Bur. Standards B*, vol. 71, no. 1, pp. 39–41, 1967.
[55] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci. USA*, vol. 39, no. 1, pp. 42–47, 1953.

Yanjun Han (Member, IEEE) received the B.Eng. degree (Hons.) in electronic engineering from Tsinghua University, Beijing, China, in 2015, and the M.S. and Ph.D. degrees from Stanford University, in 2017 and 2021, respectively. He is currently a Norbert Wiener Post-Doctoral Associate at the Institute for Data, Systems, and Society, Massachusetts Institute of Technology. His research interests include statistical machine learning, high-dimensional and nonparametric statistics, information theory, online learning and bandits, and their applications.

Soham Jana received the Bachelor of Statistics and Master of Statistics degrees (Hons.) with a specialization in theoretical statistics from the Indian Statistical Institute, Kolkata, India, in 2015 and 2017, respectively, and the Ph.D. degree in statistics and data science from Yale University, in 2022. He is a Post-Doctoral Research Associate at the Department of Operations Research and Financial Engineering, Princeton University. His research interests include statistical machine learning, high-dimensional and non-parametric statistics, online learning, mixture models, and robust estimation.

Yihong Wu received the B.E. degree from Tsinghua University in 2006 and the Ph.D. degree from Princeton University in 2011. He is a Professor at the Department of Statistics and Data Science, Yale University. His research interests include theoretical and algorithmic aspects of high-dimensional statistics, information theory, and optimization. He was a recipient of the NSF CAREER Award in 2017 and the Sloan Research Fellowship in mathematics in 2018.