Spring 2022

Learning Non-Parametric and High-Dimensional Distributions via Information-Theoretic Methods

Soham Jana
Yale University Graduate School of Arts and Sciences, soham.jana@yale.edu

Abstract

Learning Non-parametric and High-dimensional Distributions via Information-theoretic Methods

Soham Jana

2022

Learning distributions that govern the generation of data and estimating related functionals are at the foundations of many classical statistical problems. In this dissertation we intend to investigate such topics when either the hypothesized model is non-parametric or the number of free parameters in the model grows along with the sample size. Using techniques based on information-theoretic divergences and related mutual-information based methods, we study the following class of problems with the goal of obtaining minimax rate-optimal methods for learning the target distributions.

(i) Estimation in compound decision and empirical Bayes settings: To estimate the data-generating distribution, one often takes the following two-step approach. In the first step the statistician estimates the distribution of the parameters, either the empirical distribution or the postulated prior, and then in the second step plugs in the estimate to approximate the target of interest. In the literature, the estimation of empirical distribution is known as the compound decision problem and the estimation of prior is known as the problem of empirical Bayes. In our work we use the method of minimum-distance estimation for approximating these distributions. Considering certain discrete data setups, we show that the minimum-distance based method provides theoretically and practically sound choices for estimation. The computational and algorithmic aspects of the estimators are also analyzed.

(ii) Prediction with Markov chains: Given observations from a Markov chain with unknown statistics, we study the problem of predicting the next entry in the trajectory. Existing analysis for such a setup involving dependent data usually centers around concentration inequalities that uses various extraneous conditions on the mixing properties. This makes it difficult to achieve results independent of such restrictions. We

introduce information-theoretic techniques to bypass such issues and obtain fundamental limits for the related minimax problems. We also analyze conditions on the mixing properties that produce a parametric rate of prediction errors.

Learning Non-parametric and High-dimensional Distributions via Information-theoretic Methods

A Dissertation

Presented to the Faculty of the Graduate School

Of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

By

Soham Jana

Dissertation Director: Yihong Wu

May 2022

Dedication

This thesis is dedicated to my parents Ramen and Shipra, and my sister Shreyasi, whose endless love and support have made this work possible.

Contents

Acknowledgments

I am greatly indebted to my advisor, Prof. Yihong Wu, without whose guidance this work would not have been possible. As a young Padawan, I have been fortunate to see the master at work, and I hope to draw inspiration from this experience in the days to come. It is always a blessing to be in the presence of hard-working, knowledgeable, and passionate researchers with interests beyond academics, and I consider myself lucky to find all of it in my advisor. I have immensely benefited from his display of a deep understanding of the interplay between information theory and statistics. It was from him that I learned the attraction of explaining complex statistical results in layman's terms. While Yihong most certainly gave me the freedom to explore my strengths and weaknesses, whenever it was necessary, he always worked alongside me on the problems. I will cherish the late-night email exchanges about research and solving problems together during our afternoon meetings. From the early stages of academia to choosing proper career paths, Yihong's sincere support and countless life advice always instilled the belief in me to overcome challenges whenever I faced them.

Next, I would like to thank Prof. David Pollard, Prof. Andrew Barron, Prof. Harrison Zhou, Prof. John Emerson, Prof. Zhou Fan, and many other faculties at Yale who were part of the journey. I still miss the regular meetings with David on Saturdays during the pre-COVID times. I will never forget our discussions on topics ranging from Stochastic Calculus to Markov Chains and the occasional fun ones on how he was keeping the deers away from his squash farm. My memories with Andrew are almost as old as I have known the statistics department at Yale. My acceptance mail at the department came from Andrew. He was also the first person I met when I went to the department for the first time. I am incredibly grateful for all the insightful discussions on information theory and neural networks, and his

suggestions and help during my postdoctoral applications. Discussions and advising sessions with Harry, Jay, and Zhou during different classes as both student and teaching fellow were fundamental parts of my doctoral career.

I am also extremely grateful to my Master's advisor Prof. Ayanendranath Basu at the Indian Statistical Institute, Kolkata, whose mentoring and guidance laid the groundwork of my doctoral studies. Prof. Basu was the first to introduce me to the minimum-distance methods in statistics and teach me the intricacies of many related statistical techniques. My first publication was in collaboration with him. His support has been fundamental in shaping my graduate career from the beginning till the end, and I will be forever obliged for that.

The Yale administrative staff, Joann DelVecchio, Karen Kavanaugh, Elizabeth Torres and Dawn Hemstock have also been the most helpful. Their constant presence and support in every aspect of departmental work have made it much easier to focus on my studies.

I would also like to thank my friends who have been part of my life in the USA for the last five years. Thanks to Pushkar da and Subha da for being the first housing mates I ever had at Yale. I will dearly miss our late evening dinner at Tomatillo, brunch at the colleges, regular cricket practice, hanging out at the HGS courtyard till midnight with Dhruba da and Uddipan. Thanks to Raktim da for teaching me chess, Gourab da, Ipsita di, Titas di, Deepto da, Subhashish da for all the card games and chitchat. Thanks to Maruf da for all the deep discussions we had about life. Thanks to Anindita for all the movies we watched together, the jokes and puns we shared and being there whenever I was stressed out with studies and more. Thanks to Aritra, Sayoni, and Uddipan for the uncountable parties and dumb charades. Thanks to Mansa for being a great housemate for the last two years. Thanks to Milind, Varun, Sateja, Vidul, Ariktha for all the SAGA events, birthday parties, trips to East Rock, and zoom hangouts. Thanks a lot to my departmental buddies Colleen and Brandon, who were there through thick and thin and helped me get accustomed to the American culture. Thanks to Byungmin and Ganlin for being part of my reading groups on inequalities and random graphs. Thanks to Vittorio, Rifaat, and Yunus for all the poker games and solving problem-sets together. Thanks to all my friends from ISI, Rudra, Sohom, Sagnik, Dhrubajyoti, Rohan, Debangan, Samriddha and everyone else who helped

me along the way. Thanks to my violin teachers Ariel and Jinie who introduced me to western classical music and friend Royce for the numerous practice sessions together in the HGS basement.

Finally, I would like to thank my parents, grandparents, and other family members for their love and nurture. Their continuous care and good wishes has made it a delightful journey altogether. I still look forward to the days I get back to India and meet them.

List of Figures

List of Tables

Chapter 1

Introduction

Data analyses in many statistical procedures involve estimating the underlying probability law and its functionals. Given a family of model distributions and a loss function, the statistician's job is to search for a member of that family that produces the least approximation error. For our work, we consider the loss functions over the space of distributions, commonly known as the *divergences*. These measures of distances need not be metrics; more specifically, they need not be symmetric in the arguments or satisfy the triangle inequality. Notable instances of statistical divergences include:

- Squared Hellinger distance (H^2), known for its robustness properties, see Basu et al. (2011) for a detailed exposition.
- Total variation (TV) distance, related to the LeCam's two points method (Yu, 1997), for analyzing mixing of Markov chains Levin and Peres (2017b),
- Chi-square (χ^2) divergence, used for the goodness of fit test,
- Kulback Leibler (KL) divergence, related to mutual information (Rényi, 1961). [1]

Statistical analysis based on such divergences dates back to as early as the 1900s. For example, in his revolutionary paper (Pearson, 1900) Karl Pearson discussed the uses of the Chi-square divergence to describe the goodness of fit of a given probability distribution to

[1] Squared difference between the empirical and estimated distribution functions is often considered a useful distance measure in the statistical literature. For example it is used in the Cramér-von Mises criterion (Cramér, 1928; Doob, 1937). However, it is not an f-divergence and it does not play a role in this thesis.

the data. Systemic studies of the problem of distribution estimation based on similar classes of divergences were first carried out in a series of papers by the renowned Jacob Wolfowitz (Wolfowitz, 1953, 1954, 1957).

The divergences we mainly use in our work are based on the f-divergence family from information theory. Introduced formally in (Rényi, 1961) and later extended by Csiszár (1964, 1972), these divergences play important roles in quantifying entropy and information. Let P and Q be two probability distributions over a space $\mathcal{X}$ such that P is absolutely continuous with respect to Q. Then, for a convex function f such that $f(1) = 0$, the f-divergence between P, Q is defined as

$$D_f(P\|Q) = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ.$$

For the divergences KL, H^2, TVand χ^2 the function $f(t)$ is given by $t\log t, (\sqrt{t}-1)^2, \frac{1}{2}|t-1|$ and $(t-1)^2$ respectively. In the following thesis, we try to learn about unknown data generating distributions based on such information-theoretic divergences and associated mutual information based methods. The models we use for this purpose are either non-parametric or high-dimensional in nature. Brief descriptions of the problems are provided below.

1.1 Estimating distributions of parameters

To estimate population characteristics, it is a standard procedure to use some cost-effective sampling scheme even when the population size is finite. In many practical scenarios, the sampling process is likely to survey only a vanishing portion of the entire population. In such situations, using information-theoretic arguments, one can show that the consistent estimation of the whole population distribution is impossible even when the population size increases to infinity. Nonetheless, in practice, the quantity of interest often is some function of the parameters for which consistency might be ensured even when the sample size is far less than the dimension of the parameters. It is interesting to ask what type of inferences can be drawn for parameter estimation problems in such scenarios.

As an extension of the problem with large number of parameters, one can consider models

that assume non-parametric prior distributions on the parameters. The resulting structure of the data generating distribution is then given by a mixture of distributions. Some important examples of models in this context are mixtures of Gaussian distributions in the location and scale parameters or mixtures of Poisson distributions in the mean parameter. A celebrated strategy of estimating the prior in this setup uses the Non-Parametric Maximum Likelihood Estimator (NPMLE) (Lindsay, 1983a). The above estimator produces fascinating results in many applications. However, the theoretical analysis of the estimator is challenging as it rarely has closed-form expression. Consequently, the question arises: How would one compute the NPMLE? What alternatives are there to the NPMLE with similar practical benefits? Also, given such an estimator, what are corresponding statistical guarantees?

In our work, we intend to obtain a general solution for these problems in the discrete setup. Suppose that we have a parametric family of distributions $\{P_\theta : \theta \in \Theta\}$ and we observe independent samples $X_j \sim P_{\theta_j}, j = 1 \ldots, n$. In the frequentist setting, we assume that θ_j's are discrete points in Θ often want to learn about $\{\theta_j\}_{j=1}^n$. This is known as the compound statistical decision problem Robbins (1951). In our work we try to estimate the empirical distribution of $\{\theta_j\}_{j=1}^n$, also known as the profile of the population, given by $\pi = \frac{1}{n}\sum_{j=1}^n \delta_{\theta_j}$. Important population characteristics such as the number of distinct types and the entropy of the data-generating distribution are linear functionals of the profile, which signifies the usefulness of the problem. In the Bayesian settings, we assume that the parameters $\{\theta_j\}_{j=1}^n$ are independently distributed according to some prior π on Θ and want to estimate the prior. The problem of estimating the prior is known as the empirical Bayes problem. To estimate complex functionals of π, such as higher-order moments, medians, etc., an easy approach is to estimate $\widehat{\pi}$ of π and then compute the corresponding functional of $\widehat{\pi}$. Our focus for this dissertation is the method of minimum-distance estimation for estimating $\widehat{\pi}$, introduced by Kiefer and Wolfowitz (1956). Let that $\nu(\pi)$ denote a population parameter for which a sample estimate $\widehat{\nu}$ is available. Suppose that d denotes a statistical divergence on the space that includes both $\nu(\pi)$ and $\widehat{\nu}$. Then we define the minimum d-distance estimator

of π over the constraint set Π_k as

$$\widehat{\pi} = \operatorname*{argmin}_{\pi' \in \Pi_k} d(\widehat{\nu}, \nu(\pi)).$$

For our work, we will choose $\widehat{\nu}$ to be the sample empirical distribution and $\nu(\pi)$ to be a suitable population-representative known up to π. We use a class of such minimum-distance estimators to develop a unifying technique to solve the related learning problems.

In Chapter 2 we demonstrate the use of minimum distance estimators for estimating the profile. Consider a population consisting of k individuals, each belonging to one of k types (some types can be empty). Without any structural restrictions, it is impossible to learn the composition of the full population, having observed only a small (random) subsample of size $m = o(k)$. Nevertheless, we show that in the sublinear regime of $m = \omega(k/\log k)$, it is possible to consistently estimate the profile of the population in total variation distance. We also prove that in the linear regime of $m = ck$ for any constant c the optimal rate is $\Theta(1/\log k)$. Our estimator is based on Wolfowitz's minimum distance method, which entails solving a linear program (LP) of size k. We show that there is a single infinite-dimensional LP whose value simultaneously characterizes the risk of the minimum distance estimator and certifies its minimax optimality. The sharp convergence rate is obtained by evaluating this LP using complex-analytic techniques. This chapter is the reproduction of the work Jana et al. (2020).

The Chapter 3 of the thesis ventures into estimation of the prior π in the empirical Bayes settings. In our work, we study the problem for the Poisson mixtures, i.e., we assume unknown prior on the mean parameter of Poisson. We propose a family of estimators based on minimum-distance methodology, including and generalizing NPMLE, that is minimax optimal for estimating data generating distributions in expected squared Hellinger distance for priors that are assumed to have either bounded support or subexponential tails. We also consider the related problem of estimating the mean parameter of the Poisson random variable based on training samples. In Bayes settings, the metric of choice for measuring the performance of any estimator in squared error risk is its excess error over the Bayes estimator (which is unknown without information about prior). The Robbins method is the most

iconic and classical procedure in the empirical Bayes literature. When the data is generated according to the Poisson distribution, this method has been recently proven to produce a minimax rate optimal worst-case excess squared error loss over the Bayes risk (regret) in different prior classes (Brown et al., 2013; Polyanskiy and Wu, 2021). However, in practice, it can be precarious, and is well-recognized to be less smooth and destabilized by small sample counts in all finite-sample scenarios. In this work, we propose a spectrum of minimum distance empirical Bayes estimators based on previous density estimates that achieve the optimal regret, based on a training sample of size n, for bounded support $\left(\Theta(\frac{(\log n)^2}{n(\log\log n)^2})\right)$ and subexponential tails $\left(\Theta(\frac{(\log n)^3}{n})\right)$, and significantly outperform Robbins in practice. Our estimators also provide much more interpretable results due to their Bayesian form.

1.2 Predicting trajectories from Markov sources

The independence assumption on the sample observations is incompatible with and fails to capture the inherent dependency structure in many real-life situations. Besides estimating the model parameters in such a setting, there is significant importance in intelligently predicting future observations. With recent advancements in machine learning and AI research, such sequential decision-making has garnered widespread interest. Examples of possible applications include predicting future household expenditure based on the trajectory of daily/monthly expenses, advising patients based on evolving medical conditions, improving auto-complete features of search engines. Despite the numerous advances in the application sector, the theoretical aspects of such dependent data modeling are comparatively much less developed. For most structures, the problem of estimation/prediction is challenging as even the basic techniques such as the Central Limit Theorem, Law of Large Numbers are not directly applicable. Additionally, the sample size might be limited to allow such asymptotic methods. For example, consider a trajectory coming from an unknown Markov source of finite memory. The usual analysis centers around concentration inequalities (Lezaud, 1998; Paulin, 2015) and existing results (Han et al., 2018a; Hao et al., 2018; Hsu et al., 2019) require various extraneous assumptions such as fast mixing of the underlying Markov chain and stationary distribution that is bounded away from zero. One can ask: Is it impossible

to have meaningful results in an assumption-free framework? We have tried to address such questions for the prediction problem based on Markov chains, assuming only stationarity and otherwise allowing the chain to mix arbitrarily slowly. The principal technique we use in the analysis is based on properties of mutual information and their connection to the prediction errors.

Denote $\{x_\ell, x_{\ell+1}, \ldots, x_t\}$ by x_ℓ^t and x_1^t by x^t. Suppose that we observe a sample trajectory X^n from a stationary Markov chain with transition matrix of order m (i.e., the law of X_k given X^{k-1} is same as the conditional law of X_k given X_{k-m+1}^k) and stationary distribution π. We consider the problem of predicting the next entry on the trajectory X_{n+1}. In the context of m^{th}-order Markov chains this is equivalent to predicting $P_{X_{n-m+1}^n}(X_{n+1})$, the distribution of X_{n+1} given the last m observations. Denote by $D(P\|Q)$ the Kullback-Leibler (KL) divergence between discrete distributions P, Q. In our work we measure the error of prediction in terms of the expected KL divergence and study the minimax objective

$$\mathsf{Risk}_{k,n,m} \triangleq \inf_{\widehat{P}} \sup_{P} \mathbb{E}[D(P_{X_{n-m+1}^n} \| \widehat{P}_{X_{n-m+1}^n})]$$

where the suppremum is taken over all stationary m^{th}-order Markov chains P and the infimum is taken over all valid probability distribution estimates $\widehat{P}$ based on the sample X^n. The problem has been discussed in Falahatgar et al. (2016) for the case $k = 2, m = 1$, where they show $\mathsf{Risk}_{2,n,1} = \Theta(\frac{\log\log n}{n})$. Their analysis for the upper bound uses expansion of the related model probabilities based on binomial distributions and using Chernoff-type inequalities for the Binomial distribution. Such an explicit calculation is significantly more challenging for larger values of k and m. An attempt to resolve the case of $m = 1$ with general $k \geq 2$ is made by Hao et al. (2018) using concentration inequalities devised for standard Markov chains. Even though this method is much more generalizable than the previous approaches, the bounds become trivial when the spectral gaps of the underlying transition matrix or the probabilities in the stationary distribution is close to zero. As a result the exact characterization of the case of $\mathsf{Risk}_{k,n,1}$ was still unsolved.

In Chapter 4 we analyze $\mathsf{Risk}_{k,n,m}$ for general values of k, n, m. We first consider the case of $m = 1$. It is shown that given trajectory of length n with k-states ($\lesssim \sqrt{n}$) we have

$\mathsf{Risk}_{k,n,1} = \Theta(\frac{k^2}{n}\log\frac{n}{k^2})$. Our proof of the upper bound relies on relating the prediction risk with problem of redundancy (Davisson, 1983; Yang and Barron, 1999). The proof of lower bound follows from representing the Bayes lower bound to the prediction error via mutual information and then maximizing it over different prior distributions. We also studied how the prediction error behaves when we condition on different subsets of parameters that determine dependency. Consider the absolute spectral gap γ_* (defined for reversible irreducible chains), which takes values in [0,1] and note that $\frac{1}{\gamma_*}$ serves as a measure of the dependency among samples. We study the restricted minimax risk

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

where $\mathcal{M}_k(\gamma_0)$ is the set of transition matrices corresponding to irreducible and reversible chains whose absolute spectral gap exceeds γ_0. We have shown that for $k = \Theta(1)$ states, the minimax risk is $\Theta_k(\frac{1}{n})$ if and only if $\gamma_* = \Theta(1)$. We also show that when $k = 2$, for each $\gamma_0 \in (0,1)$ the minimax rate over chains with $\gamma_* \geq \gamma_0$ is given by $\Theta(\frac{1}{n}\{\max\{1, \log\log(\min\{n, \frac{1}{\gamma_0}\})\}\})$, implying that for 2 state chains we completely determine the minimax rate as a function of sample size and γ_* upon observing a single trajectory. This and the previous parts of the chapter are from the reproduction of the work Han et al. (2021).

Finally, we studied the case $m \geq 2$. We show that when the sample size n is moderately large (more specifically when $2 \leq k \leq \sqrt[m+1]{n}/C_m$ for appropriate constant C_m only depending on m) we achieve the minimax rate $\frac{k^{m+1}}{n}\log\frac{n}{k^{m+1}}$. In this regime, the structural properties of Markov transition kernels are comparatively less understood, making it difficult to extend the previous proof techniques based on spectral gap from the case of $m = 1$. Instead, we still apply information-theoretic techniques by relating the risk to redundancy, which in turn can be bounded by mutual information. Notably, the lower bound relies on a careful construction of a high-order Markov chain whose pseudo spectral gap can be bounded.

Chapter 2

Extrapolating the profile of a finite population

(This is a joint work with Prof. Yury Polyanskiy and Prof. Yihong Wu)

2.1 Introduction

Consider a finite population, say, an urn of at most k colored balls, with colors indexed by, without loss of generality, $[k] \triangleq \{1, \ldots, k\}$. Let θ_j denote the the number of balls of color $j \in [k]$ present in the urn. We observe a subsample, obtained by revealing each ball independently with probability p. This sampling scheme is referred to as the Bernoulli sampling model (Bunge and Fitzpatrick, 1993), a specific form of sampling without replacements. We will be interested in both the *linear* and the *sublinear* regime, in which the sampling probability p is a small constant or vanishing as k grows, respectively.

It is not hard to show (see Appendix 2.6.1) that unless all but a vanishing fraction of the urn is observed, it is impossible to consistently estimate the empirical distribution of the colors, which aligns with the conventional wisdom that the sample size needs to exceed the number of parameters. Fortunately, many interesting properties about the population (such as entropy, number of distinct elements) are label-invariant and hence learnable through the *profile* of the population (Orlitsky et al., 2005), defined as the empirical distribution of

$\theta = (\theta_1, \ldots, \theta_k)$:

$$\pi = \frac{1}{k} \sum_{j=1}^{k} \delta_{\theta_j}.$$

where δ_m denotes the Dirac measure (point mass) at m, Note that π is supported on $\{0, \ldots, k\}$ with mean at most one and probability mass function given by $\pi_m = \frac{1}{k} \sum_{j=1}^{k} \mathbf{1}_{\{\theta_j = m\}}$ for $m = 0, \ldots, k$. The profile provides information about the diversity of a population. For example, $\pi = (1 - \frac{1}{k})\delta_0 + \frac{1}{k}\delta_k$ and $\pi = \delta_1$ correspond to the two extremes of all balls having the same color and different colors, respectively. Furthermore, π_0 encodes the total number c of distinct colors in urn, since $\pi_0 = 1 - c/k$.

Based on the subsampled population, our goal is to reconstruct the profile π of the full population. Since many symmetric properties can be expressed as its linear functionals, estimating π under the total variation (TV) distance allows simultaneous estimation of all such bounded properties. Our main result is that the profile can be estimated consistently even in the sublinear regime.

Let $X_j \sim \mathrm{Binom}(\theta_j, p)$ be the number of observed balls of color j. The minimax TV risk of estimating π is defined as

$$R(k) = \inf \sup \mathbb{E}[\|\pi - \widehat{\pi}\|_{\mathrm{TV}}].$$

where $\|\pi - \widehat{\pi}\|_{\mathrm{TV}} \triangleq \frac{1}{2} \sum_{m \geq 0} |\pi_m - \widehat{\pi}_m|$, the supremum is over all urns of at most k balls, and the infimum is over all estimators $\widehat{\pi}$ as a function of $X = (X_1, \ldots, X_k)$. Our main result is the following.

Theorem 1. *There exist absolute constants c, C, d_0, such that if $\log k \geq \frac{d_0}{\bar{p}}$, then*

$$\min\left\{\frac{\bar{p}}{p}, \sqrt{\log k}\right\} \frac{c}{\log k} \leq R(k) \leq \min\left\{\frac{C}{p \log k}, 1\right\},$$

where $\bar{p} = 1 - p$. Furthermore, the upper bound in fact holds for all $p \in (0, 1)$, achieved by a minimum-distance estimator computable in polynomial time.

In the linear regime, Theorem 1 shows that the optimal TV rate is $\Theta(\frac{1}{\log k})$ for any

constant sampling probability p. This should be contrasted with the estimation of π_0, known as the distinct elements problem, which has been extensive studied in the literature (Bunge and Fitzpatrick, 1993; Charikar et al., 2000; Raskhodnikova et al., 2009; Valiant and Valiant, 2011; Wu and Yang, 2018). The precise behavior of the minimax risk of estimating π_0 was determined in Wu and Yang (2018). In particular, if $\frac{1}{\log k} \lesssim p \lesssim 1$, the optimal rate of π_0 is $k^{-\Theta(p)}$, much faster than estimating π itself. Our result refines this observation and reveals the following dichotomy: the polynomial rate $k^{-\Theta(p)}$ holds not just for estimating π_0 but for all π_m with $m = o(\log k)$; however, for $m = \Theta(\log k)$, π_m is much harder to estimate and the rate is no faster than $\Omega(\frac{1}{(\log k)^2})$. This explains the overall TV risk $\Omega(\frac{1}{\log k})$ for estimating the full distribution π.

In the sublinear regime, Theorem 1 shows that consistent estimation is possible if $p = \omega(\frac{1}{\log k})$. Although our current lower bound does not conclude its optimality, it is indeed the case based on existing impossibility results of the distinct element problem that shows π_0 cannot be estimated with vanishing error if $p = O(\frac{1}{\log k})$ (Valiant, 2012; Wu and Yang, 2018).

For simplicity, we focus on the Bernoulli sampling model in this paper. The results can be extended to models such as iid sampling or Poisson sampling by the usual simulation or reduction argument (cf. (Wu and Yang, 2018, Appendix A)).

2.1.1 Related work

While the precise question we are considering here was not studied before, there is a long history of related work. First we observe that the goal of estimating functionals of $\theta = (\theta_1, \ldots, \theta_k)$ is a "compound statistical decision problem", in the language of Robbins (1951). Instead of studying minimax risks of estimating θ or its functionals, (Robbins, 1951) proposed an alternative goal ("submminimaxity"), which in our case can be rephrased as follows: construct an estimator which has vanishing excess risk (regret) over that of the oracle estimator $\widehat{k}_j(X_j, \pi)$ having access to empirical distribution π of θ. The general recipe proposed in Robbins (1951) (and later promulgated by Robbins (1956) under the name of "empirical Bayes"), may roughly be described as a two-step procedure: first, one produces an estimate $\widehat{\pi}$ of π, and then, second, substitutes it into the oracle estimator obtaining

$\widehat{k}_j(X_j, \widehat{\pi})$. Thus, Robbins (Robbins, 1951, p. 146) asked (his Problem I) how well can the first step be done? Our work addresses this question.

The main part of our theorem characterizes how well the "prior" π can be estimated. We mention that while empirical Bayes method is sometimes understood only as a way to derive estimates of a particular functional of the prior, as, for example, in the Good-Turing estimator for the number of unseen species, the idea of estimating the prior itself has also been proposed in Edelman (1988); Robbins (1956). Furthermore, the solution advocated therein, Wolfowitz's *minimum distance estimator* (Wolfowitz, 1957), is the one we employ in the proof of our result. In this regard, one of the main contributions of the paper is showing that performance of the minimum distance estimators is characterized by means of a certain function $\delta_{\mathrm{TV}}(t)$, defined as the value of an infinite-dimensional linear program, which simultaneously can also be used to produce a matching *lower bound*. This duality between the upper and the lower bound has previously been observed and operationalized in the context of estimating *a single linear functional* in Juditsky and Nemirovski (2009); Polyanskiy et al. (2017); Polyanskiy and Wu (2019). Here we extend this program to estimating the *full distribution*, and evaluate the relevant δ_{TV} function using complex-analytic techniques.

Arguably, the counterintuitive part of our result is the possibility of estimating the profile π consistently in TV, despite the absence of structural assumptions on the urn configuration and despite p possibly vanishing. In fact, this is a manifestation of the fascinating effect originally discovered by Orlitsky et al. (2005) and further developed in Han et al. (2018b); Valiant and Valiant (2013), namely, although there exists no consistent estimator of the empirical color distribution, its sorted version can be estimated consistently. Nevertheless, the best upper bound that can be extracted (see Section 2.5.1 for details) from existing results is $O(\frac{1}{\sqrt{\log k}})$ in the linear regime and there is no applicable lower bound. Theorem 1 shows that this rate is suboptimal by a square root factor, potentially due to the fact that these previous work did not exploit the finiteness of the population.

In terms of techniques, while the approach of Wu and Yang (2018) to the distinct elements problem relies on polynomial interpolation and approximation, both the scheme (minimum distance estimator) and the lower bound in the present paper involve linear programming

(LP), which is more akin in spirit to the work of Polyanskiy and Wu (2019); Valiant and Valiant (2011). The technical novelty here is that we use tools from complex analysis to analyze the behavior of the LP.

Finally, we mention that a different line of research tracing back to Lord (1969) studies the "mirror image" of our problem: estimating the empirical distribution of parameters $p_1, \dots, p_k$ from samples $X_j \sim \mathrm{Binom}(\theta, p_j)$. The recent work of Tian et al. (2017) uses the method of moments to obtain the optimal rate for $\theta = o(\log k)$. This is further improved in Vinayak et al. (2019) by analyzing the nonparametric maximum likelihood. Alas, in this model, even for large population it is not possible to achieve consistent estimation without $\theta \to \infty$.

The rest of the paper is organized as follows. Section 2.2 introduces the minimum distance estimator and a general characterization of its risk by a linear program. Sections 2.3 and 2.4 are devoted to analyzing the behavior of this LP using complex-analytic techniques and Laguerre polynomials, completing the proof of Theorem 1. Section 2.5 contains a detailed discussion on related technical results and a list of open problems. Omitted proofs are contained in the rest of the appendices.

2.2 Minimum distance estimator and statistical guarantees

As mentioned in the last section, estimation of the profile revolves around the idea of minimum distance method, which fits a statistical model that is closest to the sample distribution with respect to some meaningful statistical distance. Examples of minimum distance estimators can be traced back to as early as Pearson (1900), which led to the discovery of the famous minimum chi-square method. In the 1950's, Wolfowitz studied minimum distance methods for the first time as a class, for obtaining strongly (almost surely) consistent estimators (Wolfowitz, 1957). The pioneering work of Beran (1977) demonstrates how minimum-Hellinger method can improve upon classical estimators such as the maximum likelihood in the presence of outliers. For a comprehensive account and more recent development we refer the readers to the monograph Basu et al. (2011).

To describe the paradigm of the minimum distance estimators we first introduce the

general setting of Robbins' Problem I mentioned in Section 2.1.1. Consider a parametric family of distributions $\{P_\theta : \theta \in \Theta\}$ on some measurable space $\mathcal{X}$, viewed also as a Markov transition kernel P from Θ to $\mathcal{X}$. Let d be a distance on the space of priors $\mathcal{P}(\Theta)$. Select $\theta_1, \ldots, \theta_k$ from Θ such that $\frac{1}{k}\sum_{j=1}^k c(\theta_j) \le 1$, where $c : \Theta \to \mathbb{R}$ is some cost function (could be zero), resulting in the empirical distribution $\pi \triangleq \frac{1}{k}\sum_{j=1}^k \delta_{\theta_j}$. Given observations $X_j \overset{iid}{\sim} P_{\theta_j}$, an estimate $\widehat{\pi}(X_1, \ldots, X_k)$ is produced with the goal of minimizing $\mathbb{E}[d(\widehat{\pi}, \pi)]$. The minimax risk is defined as

$$R(k) = \inf_{\widehat{\pi}} \sup_{\theta_1,\ldots,\theta_k} \mathbb{E}[d(\widehat{\pi}, \pi)]\,.$$

Remark 1. Note that Robbins also defined a related Problem II (Robbins, 1951, p. 147) in which $\theta_j \overset{iid}{\sim} G$ with $\mathbb{E}_G[c(\theta)] \le 1$ and the goal is to estimate the prior G instead of the (now random) empirical distribution π. The minimax risk $R_2(k)$ is similarly defined as the supremum over all such G. We argue that in many cases the difference between $R(k)$ and $R_2(k)$ is insignificant.

Indeed, let $\tau_k = \sup_G \mathbb{E}[d(G, \pi)]$, which due to concentration we assume is $o(R(k))$. The comparison $R_2(k) \le R(k) + \tau_k$ is by conditioning on π. In the opposite direction, if, for example, $d(\cdot, \cdot) \le 1$, then $R(k) \le R_2(m) + \frac{m^2}{2k}$ since by sampling m times from $(X_1, \ldots, X_k)$ with replacement we get m samples from Problem 2's setting with $G = \pi$ (except for a set of realizations of probability $\frac{m^2}{2k}$ on which we drew some X_j multiple times). Applying Problem 2's estimator for m samples we get the inequality. In interesting cases, $R_2(k) \asymp R_2(k^\alpha) \ll k^{-\beta}$ for any $\alpha, \beta > 0$, and thus we get $R_1(k) \asymp R_2(k)$.

To solve this problem we proceed by choosing an auxiliary metric ρ on $\mathcal{P}(\Theta)$, the set of probability measures on Θ. Let $\widehat{\nu} = \frac{1}{k}\sum_{j=1}^k \delta_{X_j}$ be the empirical distribution of the sample. Note that in expectation we have, for all $\theta_1, \ldots, \theta_k$,

$$\mathbb{E}[\widehat{\nu}] = \pi P.$$

where $\pi P = \int P_\theta \pi(d\theta) = \frac{1}{k}\sum_{j=1}^k P_{\theta_j}$. This motivates the following minimum-distance

estimator (putting existence of minimum aside):

$$\widehat{\pi} = \operatorname*{argmin}_{\pi'} \left\{ \rho(\widehat{\nu}, \pi' P) : \mathbb{E}_{\pi'}[c(\theta)] \le 1 \right\} .$$

To analyze this estimator, suppose, in addition to (2.2), we have the high-probability guarantee:

$$\mathbb{P}[\rho(\pi P, \widehat{\nu}) > t_k] \le \epsilon_k$$

for some sequences $t_k, \epsilon_k \to 0$. By the triangle inequality we also have $\mathbb{P}[\rho(\widehat{\pi} P, \pi P) > 2t_k] \le \epsilon_k$. Finally, defining the following *deconvolution function*:

$$\delta(t) \triangleq \sup\{ d(\pi, \pi') : \rho(\pi P, \pi P') \le t, \mathbb{E}_{\pi}[c(\theta)] \le 1, \mathbb{E}_{\pi'}[c(\theta)] \le 1 \} ,$$

where the supremization is over all distributions $\pi, \pi' \in \mathcal{P}(\Theta)$. Then we immediately obtain the high-probability risk bound $\mathbb{P}[d(\widehat{\pi}, \pi) > \delta(2t_k)] \le \epsilon_k$. Using other properties of d and c, we can typically convert this into an upper bound for the average risk like $\mathbb{E}[d(\widehat{\pi}, \pi)] \lesssim \delta(2t_k)$. Selecting different auxiliary metric ρ's results in different estimators. For example, the choice of ρ equal to the Kullback-Leibler divergence results in a the non-parametric maximum-likelihood estimator. As stated this is all well known. *Our key contribution is the following:* While ρ is left arbitrary so far, the choice of ρ being total variation (or Hellinger) distance is special since it comes with an essentially matching lower bound.

> *Meta-principle.* Suppose the loss function d is of seminorm-type, namely $d(\pi, \pi') = \sup_{T \in \mathcal{T}} \langle T, \pi - \pi' \rangle$ for some dual pairing $\langle \cdot, \cdot \rangle$ and a family of linear functionals $\mathcal{T}$ on $\mathcal{P}(\Theta)$. Take $\rho(\cdot, \cdot) = \| \cdot - \cdot \|_{\mathrm{TV}}$. Then under regularity conditions on $(\Theta, \mathcal{X}, c, P, \mathcal{T})$ we have
>
> $$\delta(1/k) \lesssim R(k) \lesssim \delta(t_k) .$$
>
> Thus, when $\delta(1/k) \asymp \delta(t_k)$ we get the sharp rate.

Working out general conditions for the applicability of this program is left for future work. Here we focus on the model discussed in the introduction. Recall $\pi = (\pi_0, \ldots, \pi_k)$ in (2.1) denotes the profile of the urn. In the Bernoulli sampling model, the observed numbers

of balls with color j are independently distributed as

$$X_j \overset{\text{ind.}}{\sim} \text{Binom}(\theta_j, p), \quad j \in [k].$$

Let $\widehat{\nu} = \frac{1}{k}\sum_{j=1}^{k} \delta_{X_j}$ denote the empirical distribution of the X_j's. Then for each $m \geq 0$, we have $\widehat{\nu}_m = \frac{Y_m}{k}$, where

$$Y_m = \sum_{j \in [k]} 1\{X_j = m\}$$

denotes the number of colors that are observed exactly m times.[1] Define the Markov kernel $P : \mathbb{Z}_+ \to \mathbb{Z}_+$ by $P(i, \cdot) = \text{Binom}(i, p)$, whose transition matrix $P = (P_{im})$ is given by

$$P_{im} = \binom{i}{m} p^m (1-p)^{i-m}, \quad i, m \geq 0.$$

Then as in (2.2), we have the unbiased relation $\mathbb{E}[\widehat{\nu}] = \pi P$. Particularizing (2.2) with $\rho = \|\cdot\|_{\text{TV}}$ and $c(\theta) = \theta$, we obtain the following the minimum distance estimator:

$$\widehat{\pi} = \operatorname*{argmin}_{\pi' \in \Pi_k} \|\pi' P - \widehat{\nu}\|_{\text{TV}}$$

where

$$\Pi_k \triangleq \left\{ \pi' \in \mathcal{P}\{0, 1, \ldots, k\} : \sum_{m=0}^{k} m\pi'_m \leq 1 \right\},$$

with $\mathcal{P}\{0, 1, \ldots, k\}$ being the set of all probability mass functions on $\{0, 1, \ldots, k\}$. As mentioned in Section 4.1, the true profile π belongs to Π_k. The estimator (2.2) is an LP with $k+1$ variables and can be solved in time that is polynomial in k. We will show that it attains the minimax upper bound in Theorem 1. As the first step, we relate the minimax risk $R(k)$ to the following LP of modulus of continuity type: for each $0 < t < 1$,

$$\delta_{\text{TV}}(t) \triangleq \sup\{\|\pi - \pi'\|_{\text{TV}} : \|\pi P - \pi' P\|_{\text{TV}} \leq t;\ \pi, \pi' \in \Pi\},$$

where $\Pi \triangleq \Pi_\infty$ as in (2.2), that is, the set of all distributions on $\mathbb{Z}_+$ with mean at most one.

[1] Technically, ν_0 is not directly observed from the sample. Nevertheless, one can compute it by $\widehat{\nu}_0 \triangleq 1 - \sum_{m=1}^{k} \widehat{\nu}_m$.

The following result shows that the value of this LP characterizes the minimax risk.

Theorem 2. *There exist absolute constants C_1, C_2, d_0 such that for all $k \geq d_0$*

$$\frac{1}{72}\delta_{\mathrm{TV}}\left(\frac{1}{6k}\right) - \frac{C_2}{\sqrt{k}} \leq R(k) \leq 2\delta_{\mathrm{TV}}\left(\sqrt{\frac{C_1 \log k}{k}}\right), \tag{2.1}$$

where the upper bound is attained by the minimum distance estimator given in (2.2).

The proof of Theorem 2 is given in Appendix 2.6.2. The main idea is as follows. By virtue of the minimum distance estimator $\widehat{\pi}$ and the triangle inequality, we have:

$$\|\widehat{\pi}P - \pi P\|_{\mathrm{TV}} \leq \|\widehat{\pi}P - \widehat{\nu}\|_{\mathrm{TV}} + \|\pi P - \widehat{\nu}\|_{\mathrm{TV}} \leq 2\|\pi P - \widehat{\nu}\|_{\mathrm{TV}},$$

which implies that $(\pi, \widehat{\pi})$ is a feasible pair for $\delta_{\mathrm{TV}}(t)$ with $t = 2\|\pi P - \widehat{\nu}\|_{\mathrm{TV}}$, and hence the following deterministic bound:

$$\|\widehat{\pi} - \pi\|_{\mathrm{TV}} \leq \delta_{\mathrm{TV}}(2\|\pi P - \widehat{\nu}\|_{\mathrm{TV}})$$

Recall from (2.2) that $\widehat{\nu}$ is an unbiased estimator of πP. Furthermore, by concentration inequality one can show that with high probability that $\|\widehat{\nu} - \pi P\|_{\mathrm{TV}} = O(\sqrt{\frac{\log k}{k}})$, from which the upper bound quickly follows. The lower bound follows from that of estimating linear functionals developed in Polyanskiy and Wu (2019). Roughly speaking, we use the optimal solution (π, π') for $\delta_{\mathrm{TV}}(\Theta(1/k))$ to randomly generate two urns of size $\Theta(k)$ whose sampled version are statistically indistinguishable. With appropriate truncation argument, this can be turned into a valid minimax lower bound via Le Cam's method (Tsybakov, 2004).

Theorem 2 allows us to reduce the statistical problem (2.1) to studying the behavior of $\delta_{\mathrm{TV}}(t)$ for small t. This is characterized by the following lemma:

Lemma 3. *(1) There exists absolute constant $C_3 > 0$ such that for all p, t we have*

$$\delta_{\mathrm{TV}}(t) \leq \min\left\{\frac{C_3}{p\log(1/t)}, 1\right\}. \tag{2.2}$$

(2) There exist absolute constants $C_4, t_0 > 0$ such that for any $p \in (0,1)$, $t \le t_0$,

$$\delta_{\mathrm{TV}}(t) \ge \min\left\{\frac{\bar{p}}{p}, \sqrt{\log(1/t)}\right\} \frac{C_4}{\log(1/t)}. \tag{2.3}$$

Combining Theorem 2 and Lemma 3 yields the main result in Theorem 1. The next two sections are devoted to the proof of Lemma 3.

Remark 2 (Reverse data processing)**.** Note that by the data processing inequality (DPI) of TV distance, we have $\|\pi P - \pi' P\|_{\mathrm{TV}} \le \|\pi - \pi\|_{\mathrm{TV}}$ and hence $\delta_{\mathrm{TV}}(t) \ge t$. Therefore Lemma 3 can be understood as a *reverse DPI* for the binomial kernel P in (2.2). For example, if $p = \Theta(1)$, then (2.2) implies that (which is the best possible in view of (2.3)):

$$\|\pi P - \pi' P\|_{\mathrm{TV}} \ge \exp\left\{-\Theta\left(\frac{1}{\|\pi - \pi\|_{\mathrm{TV}}^2}\right)\right\}.$$

2.3 Upper bound on $\delta_{\mathrm{TV}}(t)$ by H^∞-relaxation

To bound $\delta_{\mathrm{TV}}(t)$ from above, we first relate it to the following LP

$$\delta_*(t) \triangleq \sup_{\Delta}\left\{\sum_{m=0}^{\infty} |\Delta_m| : \|\Delta P\|_1 \le t, \sum_{m=0}^{\infty} m|\Delta_m| \le 1\right\}. \tag{2.4}$$

The next lemma shows how the two LPs (2.2) and (2.4) are related. The proof is straightforward and deferred till Appendix 2.6.3.

Lemma 4. *For all $t \in [0,1]$ we have $\frac{1}{2}(\delta_*(t) - t) \le \delta_{\mathrm{TV}}(t) \le \delta_*(t)$.*

Remark 3. Note that our only goal is to substitute estimates on δ_{TV} into (2.1). Therefore, due to the presence of the (unavoidable) second term in the LHS of (2.1), the slight difference between $\delta_*(t) - t$ and $\delta_*(t)$ in the lower bound in Lemma 4 is completely irrelevant and we can essentially think of δ_{TV} and δ_* as universally within a factor of two of each other.

Proof of upper bound in Lemma 3. We start with recalling a few facts from the complex analysis. Denote the sup-norm of a holomorphic function f over an open set $V \subset \mathbb{C}$ by $\|\cdot\|_{H_\infty(V)}$. Let $D = D_1$ be the open unit disk in $\mathbb{C}$ and denote the horodisks for $0 < p \le 1$

as

$$D_p \triangleq \bar{p} + pD = \{z \in \mathbb{C} : |z - \bar{p}| \le p\}\,.$$

In addition, we also define another norm for functions analytic in the neighborhood of the origin:

$$\|f\|_A \triangleq \sum_{j=0}^{\infty} |a_j|, \qquad f(z) \triangleq \sum_{j\ge 0} a_j z^j\,. \tag{2.5}$$

Since $f(re^{i\omega}) \le \sum_{n\ge 0} r^n |a_n| \le \|f\|_A$, we have

$$\|f\|_{H^\infty(D)} \le \|f\|_A\,.$$

In (Polyanskiy et al., 2017, (39)) by an application of Hadamard's three-lines theorem, it was shown that for any $q \in (0,1)$ and any holomorphic function f

$$\|f\|_{H^\infty(D_{1/2})} \le \|f\|^{\frac{1-2q}{\bar{q}}}_{H^\infty(D)} \|f\|^{\frac{q}{\bar{q}}}_{H^\infty(D_q)}\,. \tag{2.6}$$

Indeed, reparametrizing $f(z) = g(\frac{1+z}{1-z})$, we have

$$\|g\|_{H^\infty(\Re=r)} = \|f\|_{H^\infty(D_{1/(1+r)})}. \tag{2.7}$$

for $r \ge 0$. Then the Hadamard three-lines theorem applied to g shows that $r \mapsto \log \|f\|_{H^\infty(D_{1/(1+r)})}$ is convex, proving (2.6). A straightforward generalization (with a different choice of the middle line in the Hadamard theorem) shows that more generally for any $1 > q_1 > q > 0$ we have

$$\|f\|_{H^\infty(D_{q_1})} \le \|f\|^{1-\frac{q\bar{q}_1}{\bar{q}q_1}}_{H^\infty(D)} \|f\|^{\frac{q\bar{q}_1}{\bar{q}q_1}}_{H^\infty(D_q)}\,. \tag{2.8}$$

Next, for any f holomorphic on λD for $\lambda > 0$ we have the following estimate

$$\frac{1}{\ell!}|f^{(\ell)}(0)| \le \lambda^{-\ell}\|f\|_{H^\infty(\lambda D)}\,. \tag{2.9}$$

which follows by a Cauchy integral formula: $\frac{f^{(\ell)}(0)}{\ell!} = \frac{1}{2\pi i}\oint_{|z|=\lambda} \frac{f(z)}{z^{\ell+1}}\,dz$.

With these preparations we move to the proof of (2.2). Consider any sequence Δ feasible

for $\delta_*(t)$. For each absolutely summable sequence Δ, we consider its z-transform: $f_\Delta(z) \triangleq \sum_{m\geq 0}\Delta_m z^m$, which is a holomorphic function on the open unit disk D. Furthermore, using the definition of P in (2.2) and the binomial identity, it is straightforward to verify that $f_{\Delta P} = Pf_\Delta$, where the Markov kernel P acts on f as a composition operator $(Pf)(z) \triangleq f(pz+\bar{p})$, where $\bar{p} \triangleq 1-p$. Given this observation we see that the definition of $\delta_*(t)$ can also be restated as optimization over all holomorphic functions on the unit disk, cf. (2.5):

$$\delta_*(t) = \sup_f \left\{ \|f\|_A : \|Pf\|_A \leq t, \|f'\|_A \leq 1 \right\} . \tag{2.10}$$

For any feasible f in (2.10) we have that $\|f'\|_{H^\infty(D)} \leq 1$ and $\|f\|_{H^\infty(D_p)} \leq t$. Thus, integrating f' from some point in D_p we obtain that also $\|f\|_{H^\infty(D)} \leq 1+t \leq 2$. Therefore, applying (2.8) to f we get $\|f\|_{H^\infty(D_{3/4})} \leq 2t^{\min(\frac{p}{3\bar{p}},1)}$. Next, since $\frac{1}{2}D \subset D_{3/4}$ we have from (2.9)

$$|\Delta_\ell| = \frac{1}{\ell!}|f^{(\ell)}(0)| \leq 2^\ell t^{\min(\frac{p}{3\bar{p}},1)} \leq 2^\ell t^{p/3} . \tag{2.11}$$

Finally, since for any Δ feasible for $\delta_*(t)$ we have $\sum_m m|\Delta_m| \leq 1$, Markov inequality implies $\sum_{m\geq J}|\Delta_m| \leq \frac{1}{J}$ for any integer $J \geq 1$. Together with (2.11) we conclude that for any feasible Δ-sequence

$$\sum_m |\Delta_m| \leq J2^J t^{\frac{p}{3}} + \frac{1}{J} \leq \frac{1}{J}\left(1+6^J t^{p/3}\right) , \tag{2.12}$$

where in the last step we used $J^2 \leq 3^J$. Hence, whenever $J \leq \left\lfloor \frac{p\log\frac{1}{t}}{3\log 6} \right\rfloor$, the right-hand side of (2.12) can be upper-bounded by $\frac{2}{J}$. This, in view of Lemma 4 completes the proof of (2.2) since by definition $\delta_{\mathrm{TV}} \leq 1$. □

Remark 4. Note that functions that saturate (2.6) are $f(z) = e^{-m\frac{1+z}{1-z}}$ where $m \sim \log\frac{1}{t}$. Computing Taylor coefficients $[z^\ell]f(z)$ of $f(z)$ for $\ell = \Theta(m)$ can be done by applying the saddle-point method to the integral

$$[z^\ell]f(z) = \frac{1}{2\pi i}\oint e^{-m\frac{1+z}{1-z}-(\ell+1)\log z}dz .$$

It turns out that these coefficients behave in the following way, when $\ell/m = \Theta(1)$:

$$[z^\ell]f(z) = \begin{cases} e^{-\Theta(m)}, & \ell/m < 1/2 \\ \Theta\left(\frac{1}{\sqrt{m}}\right), & \ell/m > 1/2 \end{cases}$$

This dichotomy corresponds to critical points of the function $\frac{1+z}{1-z} - \frac{\ell}{m}\log z$ leaving the unit circle when $\ell/m < 1/2$. This shows that the estimate in (2.12) is qualitatively tight. This effect of sudden jump in the magnitude of coefficients will be the basis of the lower bound in the next section.

2.4 Lower bound on $\delta_{\mathrm{TV}}(t)$

In view of Lemma 4 it suffices to consider $\delta_*(t)$ in (2.4). Given the equivalent definition (2.10), as a warm-up, let us naively replace all $\|\cdot\|_A$ norms with $\|\cdot\|_{H^\infty(D)}$. We then get the following optimization problem:

$$\delta_{H^\infty}(t) \triangleq \sup\{\|f\|_{H^\infty(D)} : \|f'\|_{H^\infty(D)} \le 1, \|f\|_{H^\infty(\bar{p}+pD)} \le t\}$$

Note that even though the objective function of (2.4) is smaller than that of $\delta_*(t)$, the feasible set is also a relaxation. Thus $\delta_{H^\infty}(t)$ does not constitute a valid lower bound to $\delta_*(t)$; nevertheless its solution, given in the following lemma, provides important insight on constructing a near-optimal solution for $\delta_*(t)$.

Lemma 5. $\delta_{H^\infty}(t) = \Theta_p\left(\frac{1}{\log(1/t)}\right)$.

Proof. For the upper bound, as before we reparameterize $f(z) = g(w)$ with $w = \frac{1+z}{1-z}$. Then (2.7) with $r = 1/p - 1$ implies that $\|g\|_{H^\infty(\Re > \bar{p}/p)} = \|f\|_{H^\infty(\bar{p}+pD)} \le t$. By Cauchy's integral formula, we conclude that for some constant C_p (here and below possibly different on each line) we have $\|g'\|_{H^\infty(\Re > 2\bar{p}/p)} \le C_p t$.

Note that $g'(w) = \frac{2}{(1+w)^2} f'(\frac{w-1}{w+1})$. Applying (2.7) again with $r = 0$ yields $\|g'\|_{H^\infty(\Re>0)} \le 2$. Thus from Hadamard's three lines theorem we conclude for any $\epsilon \in (0, \bar{p}/p)$, $\|g'\|_{H^\infty(\Re=\epsilon)} \le C_p t^{\min\{\epsilon p/(2\bar{p}),1\}}$.

Finally, for any $\omega \in \mathbb{R}$, integrating the derivative horizontally yields:

$$|g(i\omega) - g(i\omega + \bar{p}/p)| \le C_p \int_0^{\bar{p}/p} t^{\epsilon p/(2\bar{p})} d\epsilon \le C_p \frac{1}{\log \frac{1}{t}}$$

Since $|g(i\omega+\bar{p}/p)| \le \|g\|_{H^\infty(\Re=\bar{p}/p)} \le t$, we conclude that on $\{\Re = 0\}$ we have $\|g\|_{H^\infty(\Re=0)} = \|f\|_{H^\infty(D)} \le C_p \frac{1}{\log \frac{1}{t}}$, proving the upper bound part.

For the lower bound, consider the following function

$$f(z) = \frac{c_p}{\log(1/t)} (1-z)^2 t^{\frac{p}{\bar{p}} \frac{1+z}{1-z}} \tag{2.13}$$

for some constant $c_p > 0$. Then using (2.7) we have $\|f\|_{H^\infty(\bar{p}+pD)} \le \frac{4c_p}{\log(1/t)} \sup_{z \in \bar{p}+pD} |t^{\frac{p}{\bar{p}} \frac{1+z}{1-z}}| = \frac{4c_p t}{\log(1/t)}$, and

$$\begin{aligned}
\|f'\|_{H^\infty(D)} &= c_p \left\| -\frac{2}{\log(1/t)} (1-z) t^{\frac{p}{\bar{p}} \frac{1+z}{1-z}} - \frac{2p}{\bar{p}} t^{\frac{p}{\bar{p}} \frac{1+z}{1-z}} \right\|_{H^\infty(D)} \\
&\le c_p \left(\frac{4}{\log(1/t)} + \frac{2p}{\bar{p}} \right) \left\| t^{\frac{p}{\bar{p}} \frac{1+z}{1-z}} \right\|_{H^\infty(D)} \overset{(2.7)}{=} c_p \left(\frac{4}{\log(1/t)} + \frac{2p}{\bar{p}} \right) \le \frac{2c_p(1+\bar{p})}{\bar{p}}
\end{aligned}$$

where the last inequality follows from $\log(1/t) \ge 1$ for all small t. This shows f is feasible for $\delta_{H^\infty}(t)$ for small c_p. Finally noticing that $\|f\|_{H^\infty(D)} \ge |f(-1)| = \frac{c_p}{\log(1/t)}$ concludes the proof. □

Next we modify (2.13) to produce a feasible solution for $\delta_*(t)$ leading to the following lower bound, which, in view of Lemma 4, provides the required bound in (2.3) on $\delta_{\mathrm{TV}}(t)$.

Lemma 6. *There exist absolute constants $C > 0$ and $\widetilde{\beta}_0 > 0$ such that for all $t > 0$ and $p \in [0, 1)$,*

$$\delta_*(t) \ge \frac{C}{\widetilde{\beta}}, \quad \widetilde{\beta} \triangleq \max\left(\frac{p}{1-p} \log \frac{1}{t}, \sqrt{\frac{\log \frac{1}{t}}{1-p}} \right) \tag{2.14}$$

provided that $\widetilde{\beta} \ge \widetilde{\beta}_0$.

Proof. Fix $p, t \in (0, 1)$. Considering (2.10) our goal is to find a feasible function and bound its $\|\cdot\|_A$ norm from below. Our main tool for converting between the $\|\cdot\|_A$ norms in

the definition (2.10) and the more convenient H^∞ norms is the following general result complementing (2.3): For any $r > 1$,

$$\|f\|_A \le \frac{1}{\sqrt{1-r^{-2}}} \|f\|_{H^\infty(rD)} \,. \tag{2.15}$$

Indeed, let $f(z) = \sum_{n\ge 0} a_n z^n$ and let $\widetilde{f}(z) = \sum_{n\ge 0} \widetilde{a}_n z^n$ with $\widetilde{a}_n = a_n r^n$ and thus $\widetilde{f}(z) = f(rz)$. From the Plancherel identity we have

$$\sum_n |\widetilde{a}_n|^2 = \frac{1}{2\pi} \int_0^{2\pi} |\widetilde{f}(e^{i\omega})|^2 d\omega \le \|\widetilde{f}\|^2_{H^\infty(D)} = \|f\|^2_{H^\infty(rD)} \,.$$

Thus, (2.15) follows from an application of Cauchy-Schwarz inequality:

$$\sum_n |a_n| = \sum_n r^{-n} |\widetilde{a}_n| \le \sqrt{\sum_{n\ge 0} r^{-2n}} \|f\|_{H^\infty(rD)} = \frac{1}{\sqrt{1-r^{-2}}} \|f\|_{H^\infty(rD)} .$$

Next, fix some $\beta \ge \beta_0$ and $\tau \in (0,1)$, where $\beta_0 \ge 1$ is a numeric constant to be specified later, and let $\alpha = 1 - \tau \in (0,1)$. Consider the function, a modified version of (2.13), given by

$$h(z) = \widetilde{h}(\alpha z), \qquad \widetilde{h}(z) = \exp\left(-\beta \frac{1+z}{1-z}\right).$$

Using (2.7), we can explicitly calculate that for any $0 < q \le 1$:

$$\|\widetilde{h}\|_{H^\infty(1-q+qD)} = e^{-\beta \frac{1-q}{q}} \,. \tag{2.16}$$

We will show below the following estimates (all positive numerical constants below, i.e. those that are independent of parameters p, t, β, are denoted by a common symbol C):

$$\|h\|_A \ge C\sqrt{\beta}(1-\tau)^{\frac{3\beta}{2}} \tag{2.17}$$

$$\|h(p\cdot + \bar{p})\|_A \le \tau^{-\frac{1}{2}} e^{-\beta E}, \qquad E \triangleq \frac{\bar{\tau}\bar{p}}{p + \bar{p}\tau} \tag{2.18}$$

$$\|h'\|_A \le 2\tau^{-\frac{3}{2}} \,. \tag{2.19}$$

Thus, taking $f(z) = \frac{1}{2}\tau^{\frac{3}{2}} h(z)$ in (2.10) proves that for all $\beta > \beta_0$ we have

$$\delta_* \left(\frac{\tau}{2} e^{-\beta E} \right) \geq C \sqrt{\beta \tau^3} (1-\tau)^{\frac{3\beta}{2}} \tag{2.20}$$

To show that (2.20) implies (2.14) we set $\tau = \frac{1}{\beta}$ and thus the last term in (2.20) can be lower bounded by $(1 - 1/\beta_0)^{3\beta_0/2}$ and be absorbed into C. Notice also that if $\beta \geq 2$ then $\bar{\tau} \geq 1/2$ and thus $E \geq \frac{\bar{p}}{2} \frac{1}{\frac{1}{\beta}+p}$. Since $\tau \leq 1$ and δ_* is monotone in its argument we can simplify

$$\delta_* \left(\exp\left\{ -\frac{\beta}{\frac{1}{\beta}+p} \frac{\bar{p}}{2} \right\} \right) \geq \frac{C}{\beta} \tag{2.21}$$

Note next that for any $\mu, p > 0$, taking $x = \max(\mu p, \sqrt{\mu})$ implies $\frac{x}{\frac{1}{x}+p} \geq \frac{\mu}{2}$, which is verified by considering the two cases $\mu p \lessgtr \sqrt{\mu}$ separately. Then, defining $\mu \triangleq \frac{4}{\bar{p}} \log \frac{1}{t}$ and taking $\beta = \max(\mu p, \sqrt{\mu})$ ensures the argument of δ_* in (2.21) is at most t. In summary, we obtain the bound (2.14) for all $t \leq t_0$.

We proceed to proving (2.17)-(2.19). For (2.18) we set $r = \frac{1-\alpha\bar{p}}{\alpha p}$ in (2.15) and get

$$\|h(p\cdot + \bar{p})\|_A \leq c \|h(p\cdot + \bar{p})\|_{H^\infty(rD)} = c\|h\|_{H^\infty(\bar{p}+prD)} = c e^{-\beta \frac{\alpha\bar{p}}{1-\alpha\bar{p}}},$$

where we denoted $c = \sqrt{\frac{1}{1-r^{-2}}}$ and also applied (2.16) with $q = \alpha p r = 1 - \alpha\bar{p}$. We next bound $c \leq \left(1 - r^{-1}\right)^{-1/2} = (1-\alpha\bar{p})^{1/2} (1-\alpha)^{-1/2} \leq (1-\alpha)^{-1/2}$.

For (2.19) we first notice that for any function f holomorphic on $r_2 D$ we can estimate its derivative on $r_1 D$, where $r_1 < r_2$ via Cauchy integral formula as $\|f'\|_{H^\infty(r_1 D)} \leq (r_2 - r_1)^{-1} \|f\|_{H^\infty(r_2 D)}$. Applying this with $f = h$, $r_1 = \frac{1+r_2}{2}$ and $r_2 = \frac{1}{\alpha}$ we get

$$\|h'\|_{H^\infty(r_2 D)} \leq \sqrt{2} \left(\alpha^{-1} - 1\right)^{-1/2} \|h\|_{H^\infty(D/\alpha)} = \sqrt{2} \left(\alpha^{-1} - 1\right)^{-1/2},$$

last step being again via (2.16) with $q = 1$. Applying now (2.15) with $r = r_2$ we obtain overall

$$\|h'\|_A \leq \frac{2\alpha}{(1-\alpha)\sqrt{1-\alpha^2}} \leq \frac{2}{(1-\alpha)^{3/2}}.$$

To show (2.17), we need to analyze the Taylor coefficients of h explicitly as the H^∞-norm

bound is too weak. A natural and straightforward way is to apply the saddle-point method to study these coefficients. However, due to the special nature of h its coefficients have already been well understood. Indeed, in (Szegő, 1939, 5.1.9)) it shown that for each $x \in \mathbb{C}$ and $|v| < 1$

$$e^{-x\frac{v}{1-v}} = \sum_{n=0}^{\infty} v^n L_n^{(-1)}(x)\,, \tag{2.22}$$

where $L_n^{(-1)}(x)$ are *generalized Laguerre polynomial* of degree n. We will not need explicit formulae of these polynomials and only rely on their asymptotics (of Plancherel-Rotach type), cf. (Szegő, 1939, 8.22.9): For each $\epsilon > 0$ there exists a $C_\epsilon > 0$ such that for any $n \geq 0$, any $\epsilon \leq \phi \leq \pi/2 - \epsilon n^{-1}$, we have

$$L_n^{(-1)}(x) = e^{\frac{x}{2}}(-1)^n(\pi \sin\phi)^{-\frac{1}{2}} x^{\frac{1}{4}} n^{-\frac{3}{4}} \left\{ \sin\left[n(\sin(2\phi) - 2\phi) + \frac{3\pi}{4} \right] + (nx)^{-\frac{1}{2}} O_\epsilon(1) \right\} \tag{2.23}$$

where $x = 4n\cos^2\phi$ and the $O_\epsilon(1)$ is uniformly bounded by C_ϵ for all n and ϕ.

Comparing (2.22) with the definition of h we get $h(z) = e^{-\beta}\sum_{m\geq 0} L_m^{-1}(2\beta) z^m \alpha^m$. In other words, if we denote the m-th coefficient of $h(z)$ by Δ_m, then

$$\Delta_m = e^{-\beta}\alpha^m L_m^{-1}(2\beta)\,. \tag{2.24}$$

Due to the oscillatory nature of the Laguerre polynomial, it is not possible to bound $|\Delta_m|$ away from zero. Nevertheless, the following lemma shows that two consecutive terms cannot be simultaneously small:

Lemma 7. *For all $m \in (\beta, 3\beta/2)$ and for sufficiently large β,*

$$|\Delta_m| + |\Delta_{m+1}| \geq \alpha^{3\beta/2}\beta^{-1/2}\frac{\sqrt{2}}{6}\,. \tag{2.25}$$

From here (2.17) follows simply by $\|h\|_A \geq \sum_{\beta\leq m\leq 3\beta/2} |\Delta_m| \geq \alpha^{3\beta/2}\frac{\sqrt{2\beta}}{24}$. We note that the estimate (2.17) is tight. Indeed, applying (2.15) with $r = \frac{1}{\alpha}$ yields $\|h\|_A \leq \frac{1}{\sqrt{1-\alpha^2}} \leq 1/\sqrt{\tau}$, where we also used $\|h\|_{H^\infty(D/\alpha)} = \|\widetilde{h}\|_{H^\infty(D)} = 1$ via (2.16) with $q = 1$. □

2.5 Discussions

2.5.1 Comparison with previous results

In this section we review previous results on estimating sorted distribution or profile under different loss function and different sampling model. To this end, let us consider an urn with exactly k balls. Then its composition can be described by the distribution μ on $[k]$ with $\mu(x) = \theta_j/k$. When we go from μ to π we erase the "color labels" (i.e., if the balls in the urn are arranged as piles of distinct colors, going from μ to π is analogous to turning off the lights so that only the heights of each pile, but not their colors, are shown). This could have been done in a different way by sorting μ. Namely, let us define

$$\mu_i^{\downarrow} = i\text{-th largest atom of } \mu.$$

Note that π and $\mu^{\downarrow}$ can be expressed in terms of one another. In fact we have

$$\|\pi^1 - \pi^2\|_{\mathrm{TV}} \le 2\|\mu^{1\downarrow} - \mu^{2\downarrow}\|_{\mathrm{TV}} \le 2\|\mu^1 - \mu^2\|_{\mathrm{TV}} \tag{2.26}$$

Indeed, the second inequality follows from the fact that decreasing rearrangement minimizes the ℓ_1-distance. To prove the first inequality, note that

$$2\|\mu^{1\downarrow} - \mu^{2\downarrow}\|_{\mathrm{TV}} = \sum_j \left| \sum_{i \ge j} \pi_i^1 - \pi_i^2 \right| = W_1(\pi^1, \pi^2)\,. \tag{2.27}$$

where W_1 denotes the 1-Wasserstein distance between probability distributions and, in one dimension, coincides with the L_1-distance between the cumulative distribution functions (CDFs). Since π^1, π^2 are supported on $\mathbb{Z}$, the indicator function 1_E is 1-Lipschitz for any $E \subset \mathbb{Z}$ and thus $W_1(\pi^1, \pi^2) \ge \|\pi^1 - \pi^2\|_{\mathrm{TV}}$.

Can one estimate $\mu^{\downarrow}$ from the sample X? The answer is yes, in both ℓ_∞ and ℓ_1 (TV), as well as other metrics. However, to discuss these results let us move to the setting of Robbins Problem II. Namely, suppose we have $Z^M = (Z_1, \ldots, Z_M) \overset{iid}{\sim} \mu$ with μ some arbitrary distribution on $[k]$. The relevance to the Bernoulli sampling model comes from the following simple reduction: if μ is in fact the empirical distribution of colors, then given

$\mathcal{N}$, which corresponds a sample of size $M' \sim \mathrm{Binom}(k, p)$ from μ without replacement, one can simulate an iid sample $Z_1, \ldots, Z_M$ with $M \approx (1 - e^{-\bar{p}})k$. Hence, any result regarding estimating $\mu^{\downarrow}$ from Z^M with $M = \Theta(k)$ implies a similar result about estimating $\mu^{\downarrow}$ from $\mathcal{N}$ with $p = \Theta(1)$.

We review several results regarding estimating $\mu^{\downarrow}$ from Z^M when μ is general. The pioneering result Orlitsky et al. (2005) only showed consistency, i.e. existence of estimator $\widehat{\mu^{\downarrow}}$ such that

$$\mathbb{E}\|\widehat{\mu^{\downarrow}} - \mu^{\downarrow}\|_{\mathrm{TV}} \to 0$$

without convergence rate. In a later draft Orlitsky et al. (2008) (see also (Anevski et al., 2017, Lemma 3) for a short proof) it was shown that simply estimating $\mu^{\downarrow}$ by a sorted empirical distribution achieves

$$\mathbb{E}[\|\widehat{\mu^{\downarrow}} - \mu^{\downarrow}\|_{\infty}] = O(k^{-\frac{1}{2}} \log k)\,.$$

A much more relevant result to us, however, is the one in Valiant and Valiant (2013). For any two π^1, π^2 they defined yet another distance:

$$D(\pi^1, \pi^2) = \inf_{\nu} \mathbb{E}[|\ln X_1 - \ln X_2|]\,, \tag{2.28}$$

where the infimum is over all couplings of X_1 and X_2 distributed on $\mathbb{Z}_+$ as $\mathbb{P}[X_i = j] = j\pi^i_j$ for $i \in \{1, 2\}, j \in [k]$. They have shown that when $M = a\frac{k}{\log k}$ one can get

$$\mathbb{E}[D(\widehat{\pi}, \pi)] \le O\left(\frac{1}{\sqrt{a}}\right),$$

which, per Valiant (2019), also holds for $a = \Theta(\log k)$. In addition (Valiant and Valiant, 2016, Appendix B) shows $W_1(\pi^1, \pi^2) \le 2D(\pi^1, \pi^2)$. Indeed, let $\boldsymbol{\nu}(\cdot, \cdot)$ be the optimal coupling

in (2.28). Then define a coupling of π^1 to π^2 via

$$\widetilde{\boldsymbol{\nu}}(j_1, j_2) = \begin{cases} \frac{1}{\max(j_1, j_2)} \boldsymbol{\nu}(j_1, j_2), & j_1 \neq 0, j_2 \neq 0 \\ \sum_{j \geq j_1} \left(\frac{1}{j_1} - \frac{1}{j} \right) \boldsymbol{\nu}(j_1, j), & j_2 = 0, j_1 > 0 \\ \sum_{j \geq j_2} \left(\frac{1}{j_2} - \frac{1}{j} \right) \boldsymbol{\nu}(j, j_2), & j_1 = 0, j_2 > 0 \end{cases}$$

and completing $j_1 = j_2 = 0$ as required. Letting $(X_1, X_2) \sim \boldsymbol{\nu}$ and $(\widetilde{X}_1, \widetilde{X}_2) \sim \widetilde{\boldsymbol{\nu}}$ we have that

$$\mathbb{E}[|\widetilde{X}_1 - \widetilde{X}_2|] = 2\mathbb{E}[|\widetilde{X}_1 - \widetilde{X}_2|_+] = 2\mathbb{E}\left[\frac{|X_1 - X_2|}{\max(X_1, X_2)} \right] \leq 2\mathbb{E}[|\ln X_1 - \ln X_2|] = 2D(\pi^1, \pi^2) \,.$$

In all, putting everything together we have that Valiant and Valiant showed that there exists an estimator of $\mu^{\downarrow}$ from $M = \Theta(k)$ samples such that

$$\mathbb{E}[\|\widehat{\mu^{\downarrow}} - \mu^{\downarrow}\|_{\mathrm{TV}}] = O\left(\frac{1}{\sqrt{\log k}} \right) . \tag{2.29}$$

In Han et al. (2018b) it was shown that this rate is minimax optimal over all distributions supported on $[k]$. Note, however, that since the lower bound in Han et al. (2018b) does not produce valid distributions on finite population (namely, μ with rational entries in $\frac{1}{k}\mathbb{Z}$), it does imply that the rate of estimating π in W_1 is $\frac{1}{\sqrt{\log k}}$, cf. (2.27), is sharp.

In all, we see that following the trailblazing work Orlitsky et al. (2005) a number of works have established uniform convergence guarantees in various metrics. Relevant to us is that the best result available is $\|\widehat{\pi} - \pi\|_{\mathrm{TV}} \leq O\left(\frac{1}{\sqrt{\log k}} \right)$, which can obtained by first simulating samples drawn with replacement based on those without replacements, then combining (2.29) with (2.26). We show that this rate is suboptimal by a square root factor.

2.5.2 Open problems

For $1 \le q \le \infty$, let us define by $R_q(k)$ to be the minimax risk of estimating π in the ℓ_q-norm $(\sum_m |\pi_m - \widehat{\pi}_m|^q)^{\frac{1}{q}}$. Then in the linear regime of $p = \Theta(1)$, Theorem 1 shows that

$$\left(\frac{1}{\log k}\right)^{2-\frac{1}{q}} \lesssim R_q(k) \lesssim \frac{1}{\log k}\,,$$

which is only tight for $q = 1$. Our complex-analytic methods seem to be especially well suited for studying the case of $q = 2$ and $q = \infty$, but we were not able to close the gap. The case of ℓ_∞ is of particularly interest as it concerns which individual profile is the hardest to estimate. Our result shows that for those colors that occur $m = \Theta(\log k)$ times, the corresponding π_m is particularly difficult and cannot be estimated better than $\Omega(\frac{1}{(\log k)^2})$. It is unclear if this is the hardest case.

Let us define by $R_{W_1}(k)$ to be the minimax risk of estimating π in the 1-Wasserstein distance $W_1(\pi, \widehat{\pi})$. Given the equivalence (2.27), estimate (2.29) and lower bound $W_1(\pi, \widehat{\pi}) \ge \|\pi - \widehat{\pi}\|_{\mathrm{TV}}$ we get

$$\frac{1}{\log k} \lesssim R_{W_1}(k) \lesssim \frac{1}{\sqrt{\log k}}\,.$$

Due to W_1 being the L_1-distance between the CDFs, the minimax W_1 risk are also amenable to complex-analytic techniques, but so far resisted our attempts. An alternative approach is to generalize the W_1-lower bound construction of Han et al. (2018b); however, as observed in previous work in the distinct elements problem (Valiant, 2012; Wu and Yang, 2018) such moment-based construct is difficult to extend to finite population.

2.6 Appendix

2.6.1 Impossibility of learning the empirical distribution

In this section we show that unless we observe all but a vanishing fraction of the urn, it is impossible to estimate the empirical distribution of the colors consistently. To this end, consider a k-ball urn and let μ denote the empirical distribution of the colors, with $\mu(j) = \frac{\theta_j}{k}, j \in [k]$. Compared to the profile π which is a distribution on $\mathbb{Z}_+$, here μ is a

probability measure on the set of colors $[k]$. Similar to (2.1), we define the minimax TV risk for estimating μ:

$$\widetilde{R}(k) = \inf_{\widehat{\mu}} \sup_{\mu} \mathbb{E}[\|\mu - \widehat{\mu}\|_{\mathrm{TV}}].$$

The following theorem shows that whenever the sampling ratio p is bounded away from one, it is impossible to estimate μ consistently. This observation agrees with the typical behavior in high-dimensional estimation that, absence any structural assumptions, the sample size need to exceed the number of parameters to achieve consistency.

Theorem 8.

$$\widetilde{R}(k) \geq \frac{k-1}{4k} h^{-1}\left(1 - p - \frac{\log_2(k+1)}{k-1}\right)$$

where $h : [0,1] \to [0,1]$ *given by* $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ *is the binary entropy function, and* h^{-1} *is its inverse on* $[0, \frac{1}{2}]$. *Consequently, for any fixed* $p < 1$, $\widetilde{R}(k) = \Omega(1)$.

Proof. The proof follows the mutual information method that compares the amount of information data provides and the minimum amount of information needed to reconstruct the parameters up to a certain accuracy. Consider the following Bayesian setting of a k-ball urn, where $\theta_j \overset{\text{i.i.d.}}{\sim} Ber(1/2)$ for $j = 1, \ldots, k-1$ and $\theta_k = k - \sum_{j<k} \theta_j$. In other words, each of the first $k-1$ colors either is absent or appear exactly once with equal probability. Then for $j \in [k-1]$, the observed X_j is simply the erased version of θ_j with erasure probability $\bar{p}$. Thus the mutual information (in bits) between the parameters $\theta = (\theta_j : j \in [k-1])$ and the observations $X = (X_j : j \in [k])$ can be upper bounded as follows:

$$I(\theta; X) = \underbrace{I(\theta; X_1, \ldots, X_{k-1})}_{=(k-1)p} + \underbrace{I(\theta; X_k | X_1, \ldots, X_{k-1})}_{\leq H(X_k) \leq \log_2(1+k)}$$

where the inequality follows from the fact that X_k takes at most k values. On the other hand, suppose there exists $\widehat{\mu} = \widehat{\mu}(X)$, such that $\mathbb{E}[\|\mu - \widehat{\mu}\|_{\mathrm{TV}}] \leq \epsilon$. Define $\widehat{\theta}_j = \mathbf{1}_{\{\widehat{\mu}_j > \frac{1}{2k}\}}$ for $j \in [k-1]$. Then $2\|\mu - \widehat{\mu}\| \geq \sum_{i=1}^{k-1} \|\mu_j - \widehat{\mu}_j\| \geq \frac{1}{2k} \sum_{i=1}^{k-1} \mathbf{1}_{\{\theta_j \neq \widehat{\theta}_j\}}\|$. Thus $\widehat{\theta}$ are close to θ in Hamming distance: $\sum_{i=1}^{k-1} \mathbb{P}[\theta_j \neq \widehat{\theta}_j] \leq 4\epsilon k$. By the rate-distortion function of Bernoulli distribution (Cover and Thomas, 2006, Chap. 10), their mutual information must be lower

bounded by

$$I(\theta;\widehat{\theta}) \geq (k-1)\left(1 - h\left(\frac{4\epsilon k}{k-1}\right)\right).$$

Combined with the data processing inequality $I(\theta;X) \geq I(\theta;\widehat{\theta})$, the last two displays imply that $\epsilon \geq \frac{k-1}{4k} h^{-1}(\bar{p} - \frac{\log_2(k+1)}{k-1})$ which concludes the proof. □

2.6.2 Proof of Theorem 2

Proof. We first prove the upper bound by analyzing the minimum distance estimator (2.2). Let $\pi \in \Pi_k \subset \Pi$ denote the true profile. Denote the distribution $\nu \triangleq \pi P$. As outlined in Section 2.2 and in view of (2.2), the key step is to show that $\widehat{\nu}$ is concentrated around ν in terms of total variation. To this end, observe that for $m \geq 1$, we have $\mathbb{E}[\widehat{\nu}_m] = \nu_m$ from (2.2). Furthermore,

$$k \cdot \mathrm{Var}[\widehat{\nu}_m] = \frac{1}{k}\mathrm{Var}[Y_m] = \frac{1}{k}\sum_{j\in\mathcal{X}} \mathrm{Var}[1\{X_j = m\}] \leq \frac{1}{k}\sum_{j\in\mathcal{X}} \mathbb{P}[X_j = m] = (\pi P)_m = \nu_m\,.$$

Thus $\mathbb{E}\left[|\widehat{\nu}_m - \nu_m|\right] \leq \sqrt{\nu_m/k}$. Summing over m we get

$$\begin{aligned}
\mathbb{E}[\|\widehat{\nu} - \nu\|_{\mathrm{TV}}] \leq \mathbb{E}\left[\sum_{m=1}^{k} |\widehat{\nu}_m - \nu_m|\right] &\leq \frac{1}{\sqrt{k}} \sum_{m=1}^{k} \sqrt{\nu_m} \\
&\overset{\text{(a)}}{\leq} \frac{1}{\sqrt{k}} \left(\sum_{m=1}^{k} m\nu_m\right)^{1/2} \left(\sum_{m=1}^{k} \frac{1}{m}\right)^{1/2} \\
&\overset{\text{(b)}}{\leq} O\left(\sqrt{\frac{\log k}{k}}\right),
\end{aligned} \tag{2.30}$$

where (a) follows from Cauchy-Schwarz; (b) follows as follows: if we denote $U_1 \sim \pi$ and $U_2|U_1 \sim \mathrm{Binom}(U_1, p)$, then $U_2 \sim \nu$ and hence $\mathbb{E}[U_2] = p\mathbb{E}[U_1] \leq p$ thanks to the mean constraint on $\pi \in \Pi$. Next we show that

$$\mathbb{P}\left[\left|\|\nu - \widehat{\nu}\|_{\mathrm{TV}} - \mathbb{E}\|\nu - \widehat{\nu}\|_{\mathrm{TV}}\right| \geq \epsilon\right] \leq \exp(-C_0 k\epsilon^2) \tag{2.31}$$

for some absolute constant C_0, all $\epsilon > 0$ and k large. For that we aim to show that $\|\nu - \widehat{\nu}\|_{\mathrm{TV}}$ satisfies the bounded difference property and then apply McDiarmid's inequality.

Let $x_1, \ldots, x_{\widetilde{k}}$ be the distinct colors present in the urn with $\widetilde{k} \le k$. Denote $\|\nu - \widehat{\nu}\|_{\mathrm{TV}} = d(N_{x_1}, \ldots, N_{x_{\widetilde{k}}})$ for some function d. Then d satisfies the following: for any $i \in [\widetilde{k}]$ and any $n_1, \ldots, n_{\widetilde{k}}$ with $n_i' \neq n_i$, we have

$$\begin{aligned}
&\left|d(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_{\widetilde{k}}) - d(n_1, \ldots, n_{i-1}, n_i', n_{i+1}, \ldots, n_{\widetilde{k}})\right| \\
\le &\frac{1}{2}\left||\nu_{n_i} - \widehat{\nu}_{n_i}| + |\nu_{n_i'} - \widehat{\nu}_{n_i'}| - \left|\nu_{n_i} - \left(\widehat{\nu}_{n_i} - \frac{1}{k}\right)\right| - \left|\nu_{n_i'} - \left(\widehat{\nu}_{n_i'} + \frac{1}{k}\right)\right|\right| \\
\le &\frac{1}{k}.
\end{aligned}$$

Furthermore, $(N_{x_1}, \ldots, N_{x_{\widetilde{k}}})$ are independent. Then the desired exponential bound in (2.31) follows from McDiarmid's inequality.

Combining (2.30) and (2.31) we get

$$\mathbb{P}\left[\|\nu - \widehat{\nu}\|_{\mathrm{TV}} \ge \sqrt{\frac{C_1 \log k}{k}}\right] \le k^{-1}$$

for some absolute constant C_1. Then taking expectations on both sides of (2.2), for sufficiently large k we get

$$\begin{aligned}
\mathbb{E}\|\widehat{\pi} - \pi\|_{\mathrm{TV}} &\le \mathbb{E}[\delta_{\mathrm{TV}}(2\|\pi P - \widehat{\nu}\|_{\mathrm{TV}})] \\
&\overset{\text{(a)}}{\le} \delta_{\mathrm{TV}}\left(\sqrt{\frac{C_1 \log k}{k}}\right) + k^{-1} \\
&\overset{\text{(b)}}{\le} 2\delta_{\mathrm{TV}}\left(\sqrt{\frac{C_1 \log k}{k}}\right),
\end{aligned}$$

where (a) follows from (2.6.2) and $\delta_{\mathrm{TV}} \le 1$, (b) follows from the universal fact that $\delta_{\mathrm{TV}}(t) \ge t$ (Remark 2) and $\delta_{\mathrm{TV}}(t)$ is increasing in t. This yields the desired upper bound on $R(k)$.

To show the lower bound, consider any bounded function $T : \mathbb{Z}_+ \to [-1, 1]$. Then for distribution π on $\mathbb{Z}_+$, define the linear functional T_π:

$$T_\pi = \sum_m \pi_m T(m).$$

Note that $2\|\widehat{\pi} - \pi\|_{\mathrm{TV}} = \sup_T |T_{\widehat{\pi}} - T_\pi|$ for any estimator $\widehat{\pi}$ of π. Hence the minimax TV

risk of estimating π can be lower bounded by that of estimating T

$$R(k) \geq \frac{1}{2} R_T(k), \quad R_T(k) \triangleq \inf \sup \mathbb{E}\left[|\widehat{T} - T_\pi|\right].$$

where the estimator $\widehat{T}$ depends on $(X_j : j \in \mathcal{X})$ and the supremum is again over all k-ball urns. We are now in position to apply (Polyanskiy and Wu, 2019, Theorem 8) (with $\Theta = \mathbb{Z}_+$, $c(\theta) = \theta$, and $K_V = 1$) to obtain[2]

$$R_T(k) \geq \frac{1}{72} \delta_{\mathrm{TV}}\left(\frac{1}{6k}\right) - \frac{C_2}{\sqrt{k}}$$

where

$$\delta_{\mathrm{TV}}(t, T) = \sup\{|T_{\pi'} - T_\pi| : \mathrm{TV}(\pi' P, \pi P) \leq t, \pi, \pi' \in \Pi\}$$

Finally optimizing over T observing that $\delta_{\mathrm{TV}}(t) = \sup_T \delta_{\mathrm{TV}}(t, T)$ for every $t > 0$ yields the result. □

2.6.3 Proofs of technical lemmas

Lemma 4. We prove the lemma by showing how a feasible solution of one of the programs can be utilized to get a feasible solution of the other one, and vice-versa. Let us start with the second inequality. Given any pair (π,π') feasible for $\delta_{\mathrm{TV}}(t)$, choose $\Delta = (\pi - \pi')/2$. We get

$$\sum_m m|\Delta_m| = \frac{1}{2} \sum_m m|\pi_m - \pi'_m| \leq \frac{1}{2} \sum_m m(\pi_m + \pi'_m) \leq 1.$$

The relation $\|\Delta P\|_1 \leq t$ follows directly from $\|\pi P - \pi' P\|_{\mathrm{TV}} \leq t$. This shows Δ is feasible for $\delta_*(t)$ with $\|\Delta\|_1 = \|\pi - \pi'\|_{\mathrm{TV}}$. This proves the second inequality in Lemma 4.

The first inequality is proven next. Take any non-zero feasible solution $\widetilde{\Delta}$ to $\delta_*(t)$ (which exists because we can always choose $\widetilde{\Delta} = 0$). Next, suppose that $\epsilon \triangleq \sum_m \widetilde{\Delta}_m \neq 0$. Then,

[2] The result of (Polyanskiy and Wu, 2019, Theorem 8) is stated in terms of the χ^2-divergence. The TV version follows by applying (Polyanskiy and Wu, 2019, Proposition 1) to lower bound $\delta_{\chi^2}(t)$ via $\delta_{\mathrm{TV}}(t)$.

let us define $\Delta_j = \widetilde{\Delta}_j$ for $j \geq 1$ and $\Delta_0 = \widetilde{\Delta}_0 - \epsilon$. It is clear that

$$\sum_j \Delta_j = 0 \tag{2.32}$$

Furthermore, since $\langle \widetilde{\Delta} P, \mathbf{1} \rangle = \langle \widetilde{\Delta}, \mathbf{1} \rangle = \epsilon$ we conclude that $|\epsilon| \leq \|\Delta P\|_1 \leq t$. Therefore,

$$\sum_j |\Delta_j| \geq \sum_j |\widetilde{\Delta}_j| - t\,. \tag{2.33}$$

Finally, because $\|rP\|_1 \leq \|r\|_1$ we also have from triangle inequality

$$\|\Delta P\|_1 \leq t + |\epsilon| \leq 2t\,. \tag{2.34}$$

Next we define $\Delta^+ = \max(\Delta, 0)$, $\Delta^- = \max(-\Delta, 0)$, where max is defined coordinate wise. We choose $\{\pi_m\}_{m=0}^{\infty}$ and $\{\pi'_m\}_{m=0}^{\infty}$ as

$$\pi_0 = 1 - \sum_{j=1}^{\infty} \Delta_j^+, \quad \pi'_0 = 1 - \sum_{j=1}^{\infty} \Delta_j^-,$$
$$\pi_m = \Delta_m^+, \quad \pi'_m = \Delta_m^-, \quad m \geq 1$$

Note that under constraints on Δ, we have $\pi, \pi' \in \Pi$. Indeed, $\sum_{m\geq 1} |\Delta_m| \leq \sum_m m|\Delta_m| = \sum_m m|\widetilde{\Delta}_m| \leq 1$ and thus $\pi_0, \pi'_0 \geq 0$. Furthermore, since $|\Delta_m| = \Delta_m^+ + \Delta_m^-$ we have $\sum_m m(\Delta_m^+ + \Delta_m^-) \leq 1$ which implies $\sum_m m(\pi_m + \pi'_m) \leq 1$. This proves $\sum_m m\pi_m \leq 1$ and $\sum_m m\pi'_m \leq 1$. Next, observe that $\pi_0 - \pi'_0 = \Delta_0$ due to (2.32) and thus $\pi - \pi' = \Delta$. From (2.34) we conclude that $\|(\pi - \pi')P\|_{\mathrm{TV}} \leq t$ and hence (π, π') is a feasible pair for $\delta_{\mathrm{TV}}(t)$. And thus via (2.33) we obtain

$$\delta_{\mathrm{TV}}(t) \geq \frac{1}{2}\left(\delta_*(t) - t\right)\,.$$

□

Lemma 7. In view of (2.24) and (2.23) the proof of (2.25) is straightforward but delicate. To simplify analysis we will assume $\beta \to \infty$ and denote by $o(1)$ the terms vanishing with β.

For $m \in \left(\beta, \frac{3\beta}{2}\right)$ we define $\phi_m = \arccos\sqrt{\beta/(2m)}$ and $\theta_m = F(\phi_m)$ where $F(\phi) = \sin(2\phi) - 2\phi$. Here $\phi_m \in (\arccos(1/2), \arccos(1/3))$ and hence is bounded away from both 0 and $\pi/2$ for all m in the above range. Then using (2.23) with $x = 2\beta$, we get that there exist absolute constants β_0, C_7 such that for all $\beta \geq \beta_0$,

$$\begin{aligned}
|L_m^{(-1)}(2\beta)| &\geq \frac{e^\beta}{\sqrt{\pi}\left(1-\frac{1}{3}\right)^{1/4}}(2\beta)^{1/4}m^{-3/4}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| - C_7\beta^{-1}\right\} \\
&\geq \frac{2e^\beta}{\sqrt{\pi}(2/3)^{1/4}3^{3/4}}\beta^{-1/2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| - C_7\beta^{-1}\right\} \\
&\geq \frac{e^\beta\beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| - C_7\beta^{-1}\right\}. \qquad (2.35)
\end{aligned}$$

Now we consider any two consecutive integers m and $m+1$ in $\left(\beta, \frac{3\beta}{2}\right)$. Using (2.35) we get

$$\begin{aligned}
&|L_m^{(-1)}(2\beta)| + |L_{m+1}^{(-1)}(2\beta)| \\
\geq &\frac{e^\beta\beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| + \left|\sin\left((m+1)\,\theta_{m+1} + \frac{3\pi}{4}\right)\right| - 2C_7\beta^{-1}\right\}. \qquad (2.36)
\end{aligned}$$

The phase difference between the two sine terms comes out to be $m(\theta_m - \theta_{m+1}) - \theta_m$. Using the formula $\theta_m = F(\phi_m)$, we get

$$m(\theta_m - \theta_{m+1}) = m(\phi_m - \phi_{m+1})\frac{F(\phi_m) - F(\phi_{m+1})}{\phi_m - \phi_{m+1}}. \qquad (2.37)$$

We will show that the above is bounded away from 0 as m goes to infinity. We first consider the term $m(\phi_m - \phi_{m+1})$. Using $\frac{\mathrm{d}}{\mathrm{d}x}\arccos\sqrt{x} = -\frac{1}{2}\frac{1}{\sqrt{x(1-x)}}$ we deduce that

$$\begin{aligned}
m\left(\phi_m - \phi_{m+1}\right) &= m\left(\arccos\sqrt{\frac{\beta}{2m}} - \arccos\sqrt{\frac{\beta}{2m+2}}\right) \\
&= m\left(\arccos\sqrt{\beta/2m} - \arccos\sqrt{\beta/2m - \frac{\beta/2m}{m+1}}\right) \\
&= \frac{\beta}{2m}\cdot\frac{m}{m+1}\cdot\frac{\arccos\sqrt{\beta/2m} - \arccos\sqrt{\beta/2m - \frac{\beta/2m}{(m+1)}}}{\frac{\beta/2m}{(m+1)}} \\
&= -\frac{1}{2}\sqrt{\frac{\beta/2m}{1-\beta/2m}} + o(1)
\end{aligned}$$

where the $o(1)$ term goes to 0 as m, β tends to infinity with $\frac{\beta}{2m} \in \left(\frac{1}{3}, \frac{1}{2}\right)$. In view of (2.37) using $F'(\phi) = 2\cos(2\phi) - 2$ and $\cos^2(\phi_m) = \frac{\beta}{2m}$ we get

$$\begin{aligned} m(\theta_m - \theta_{m+1}) &= -\frac{1}{2}\sqrt{\frac{\beta/2m}{1-\beta/2m}}F'(\phi_m) + o(1) \\ &= -2\sqrt{\frac{\beta/2m}{1-\beta/2m}}\left(\frac{\beta}{2m} - 1\right) + o(1) \\ &= 2\sqrt{\frac{\beta}{2m}\left(1 - \frac{\beta}{2m}\right)} + o(1) \end{aligned} \tag{2.38}$$

with the same last conditions on m, β. As $\frac{\beta}{2m} \in \left(\frac{1}{3}, \frac{1}{2}\right)$ the above quantity is bounded away from 0. Also (2.38) implies that θ_{m+1} can be approximated as $\theta_m + o(1)$. As we have

$$\begin{aligned} \theta_m &= \sin(2\phi_m) - 2\phi_m \\ &= 2\sin\phi_m \cos\phi_m - 2\phi_m \\ &= 2\sqrt{\frac{\beta}{2m}\left(1 - \frac{\beta}{2m}\right)} - 2\phi_m \end{aligned}$$

continuing (2.36) and using (2.38) we get

$$\begin{aligned} &|L_m^{(-1)}(2\beta)| + |L_{m+1}^{(-1)}(2\beta)| \\ \geq &\frac{e^\beta \beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| + \left|\sin\left(m\theta_m + \frac{3\pi}{4} + \theta_m - 2\sqrt{\frac{\beta}{2m}\left(1 - \frac{\beta}{2m}\right)}\right)\right| + o(1)\right\} \\ = &\frac{e^\beta \beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| + \left|\sin\left(m\theta_m + \frac{3\pi}{4} - 2\phi_m\right)\right| + o(1)\right\}. \end{aligned} \tag{2.39}$$

Now we note that for any real number $a \in (0, \pi)$ the function $s(x) \triangleq |\sin(x)| + |\sin(x-a)|$ has period π and is piecewise concave on the intervals $(0, a)$ and (a, π). As $s(0) = s(a) = s(\pi) = \sin(a)$ we get

$$\inf_j \{|\sin(x)| + |\sin(x-a)|\} = \sin(a).$$

In view of the above, continuing (2.39) we get

$$\begin{aligned}|L_m^{(-1)}(2\beta)| + |L_{m+1}^{(-1)}(2\beta)| &\geq \frac{e^\beta \beta^{-1/2}}{2} \left\{\sin\left(2\phi_m\right) + o(1)\right\} \\ &= \frac{e^\beta \beta^{-1/2}}{2} \left\{2\sqrt{\frac{\beta}{2m}\left(1 - \frac{\beta}{2m}\right)} + o(1)\right\} \\ &\geq \frac{e^\beta \beta^{-1/2}}{2} \left(\frac{2\sqrt{2}}{3} + o(1)\right)\end{aligned}$$

for any $m \in \left(\beta, \frac{3\beta}{2}\right)$. In view of (2.24) this implies (2.25). □

Acknowledgment

The authors thank C. Daskalakis for pointing out Valiant and Valiant (2013) and G. Valiant for communicating Valiant (2019). S. Jana and Y. Wu were supported in part by the NSF Grant CCF-1900507, NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. Y. Polyanskiy's work was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370, and the MIT-IBM Watson AI Lab.

Chapter 3

Poisson empirical Bayes estimation: How to improve upon the (optimal) Robbins estimator

(This is a joint work with Prof. Yury Polyanskiy and Prof. Yihong Wu)

3.1 Introduction

Suppose we have an observation Y coming from the Poisson distribution with an unknown mean θ, which we want to estimate under the squared error loss. Under the Bayesian settings, one assumes that θ is distributed according to some prior G on the positive real line ($\mathbb{R}_+$). It is well established that when G is known, the optimal estimator (termed as the Bayes estimator) is given by the posterior mean $\mathbb{E}[\theta|Y]$. Denote by f_G the marginal distribution of Y, given by the mixture of the Poisson mass function $f_\theta(y) = e^{-\theta}\frac{\theta^y}{y!}, y \in \mathbb{Z}_+ \triangleq \{0, 1, \ldots\}$ and the prior distribution G: $f_G(y) = \int f_\theta(y)G(d\theta)$. Then the Bayes estimator of θ for sample Y is given as

$$\widehat{\theta}_G(Y) = \mathbb{E}_G[\theta|Y] = \frac{\int \theta e^{-\theta}\frac{\theta^Y}{Y!}G(d\theta)}{\int e^{-\theta}\frac{\theta^Y}{Y!}G(d\theta)} = (Y+1)\frac{f_G(Y+1)}{f_G(Y)}. \tag{3.1}$$

Unfortunately, absolute knowledge about the prior is rarely available in practice, and one often resorts to approximating the Bayes estimator. A surprising fact about the Empirical Bayes (EB) theory, introduced first by Robbins (1951, 1956), is that when independent and seemingly unrelated training samples $Y^n \triangleq \{Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} f_G$ are available, it is indeed possible to produce estimate of $\widehat{\theta}_G(Y)$ that has vanishing excess risk over the estimation error of oracle (known as "regret"). Since then, empirical Bayes methodology has been widely used in practice for large-scale data analysis, with notable applications in computational biology especially microarrays (Efron et al., 2001), ecology (Ver Hoef, 1996), sports prediction (Brown, 2008), predicting accidents (Persaud et al., 2010), etc. For a detailed exposition on the theory and methodology on empirical Bayes see Casella (1985); Efron (2014, 2021); Morris (1983); Zhang (2003).

In the literature, the problem of the Poisson EB has two main avenues of solution:

- f-modeling: Construct an approximate Bayes estimator using the empirical counts in the sample. For example, consider the modified Robbins estimator [1] (Polyanskiy and Wu, 2021; Robbins, 1956)

$$\widehat{\theta}_{\text{Robbins}}(Y) \triangleq \widehat{\theta}_{\text{Robbins}}(Y; Y_1, \ldots, Y_n) = (Y+1)\frac{N_n(Y+1)}{N_n(Y)+1} \tag{3.2}$$
$$N_n(y) \triangleq \sum_{i=1}^n \mathbf{1}_{\{Y_i=y\}}, y \in \mathbb{Z}_+.$$

- g-modeling: A more systematic approach that involves obtaining an estimate $\widehat{G}^{(n)}$ of the prior distribution G from Y^n and then computing $\widehat{\theta}_{\widehat{G}^{(n)}}$. Examples of $\widehat{G}^{(n)}$ include the celebrated non-parametric maximum likelihood estimator (NPMLE) (Kiefer and Wolfowitz, 1956; Lindsay, 1983a,b)

$$\widehat{G}^{(n)} = \operatorname*{argmax}_G \prod_{i=1}^n f_G(Y_i)$$

[1] The classical Robbins estimator in Robbins (1956) is constructed for compound estimation settings in which the interest lies in recovering $\theta^n \triangleq \{\theta_i\}_{i=1}^n$ from observations Y^n. The estimate of θ_j is given by $\widehat{\theta}_j = (Y_j+1)\frac{\sum_{i=1}^n \mathbf{1}_{\{Y_i=Y_j+1\}}}{\sum_{i=1}^n \mathbf{1}_{\{Y_i=Y_j\}}}$. This is same as (3.2) when we view $(Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n)$ as training sample for estimating θ_j.

where the maximization is performed over a properly chosen class of prior distributions on $\mathbb{R}_+$.

In recent years, there have been significant developments in analyzing the properties of f-modeling estimators for the Poisson EB, specifically the Robbins method. For studying priors with bounded supports, Brown et al. (2013) considered a certain sampling method which generates a sample of variable size $N \sim \text{Poi}(n)$ (here and below $\text{Poi}(\theta)$ denotes the Poisson distribution with mean θ). There the authors showed that the Robbins estimator achieves a regret rate of $O\left(\frac{(\log n)^2}{n(\log\log n)^2}\right)$. Later, Polyanskiy and Wu (2021) carried out the analysis for a fixed sample size n and showed the estimator is minimax rate-optimal by attaining a similar worst-case lower bound. In the case of priors with a subexponential tail, the same estimator also achieves optimal minimax regret $\Theta\left(\frac{(\log n)^3}{n}\right)$. Albeit its optimality guarantees and simplicity in computation, the Robbins method produces unstable performances in most practical cases. This behavior originates in a finite sample scenario as there will always be $y \in \mathbb{Z}_+$ for which there are fewer sample observations $N(y)$, sometimes as small as 0. When $N(y+1)$ is moderately large this will pull $\widehat{\theta}_{\text{Robbins}}(y)$ estimate towards exceptionally high values. Also if $N(y+1) = 0$ for some y (for example $y \geq \max_{i=1}^n Y_i + 1$) the estimate $\widehat{\theta}_{\text{Robbins}}(y)$ shrinks to 0 irrespective of any existing information about sample observations for y. This contradicts the fact that the Bayes estimator $\widehat{\theta}_G(y)$ is increasing in y for any G (Brown et al., 2013, Section 8). Since the conception of the Robbins method, several articles have drawn criticism to these and similar finite sample issues of the estimator. For historical discussions see (Maritz, 1968, Section 1),(Maritz and Lwin, 2018, Section 1.9). In recent times (Efron and Hastie, 2021, Section 6.1) also pointed out such destabilized behavior with real life examples (e.g., while analyzing insurance claims data from automobile companies). Even though most of these articles have come up with estimators to produce better numerical performances, it is not guaranteed whether they can challenge finite-sample optimality guarantees of the Robbins method. This raises the question of whether it is possible to construct some estimator that removes these problems while maintaining the similar optimality results.

Interestingly all the g-modeling estimators are free from these fluctuating behaviors,

thanks to the aforementioned monotonicity property. More generally, this monotonicity property of g-estimators holds for any mixture model of exponential families (Houwelingen and Stijnen, 1983) and the Poisson model happens to be a specific example. Modeling of the g-type is also convenient for situations where the the statistician has qualitative knowledge about the shape of the prior. Coming to the issue of consistency for g-modeling, (3.1) suggests the possibility of translating point-wise approximation guarantees for the mixture f_G to that of estimating $\widehat{\theta}_G$. The minimum-distance method is one of the existing classical techniques to estimate the data generating mixture distribution. Introduced in the pioneering works Wolfowitz (1953, 1954), these estimators are long-established to produce robust and consistent approximates to the data-generating distributions, see Basu et al. (2011); Beran (1977) for detailed expositions. In particular, for estimating a mixture of discrete distributions, even the NPMLE is of minimum-distance type (see the discussions in Section 3.2.2). As a result, it is natural to ask whether the minimum-distance-based techniques can be utilized for both density estimation and estimation of $\widehat{\theta}_G$. Considering the Poisson mixtures for the class of bounded and subexponential priors, we show a collection of minimum-distance estimators that are optimal for mixture density estimation (in the squared Hellinger distance) the g-estimators based on them also achieve the optimal regret rate. This collection of estimators includes the NPMLE as well.

To motivate our approach further, we provide here some synthetic data examples to compare several g-modeling estimators against the Robbins method. For demonstration purposes let us consider $G = \text{Uniform}[0, 3]$ (similar real and synthetic data experiments for different priors are carried out later in Section 3.5.3). In addition to the NPMLE, we consider the minimum squared Hellinger (H^2) and the minimum Chi-squared (χ^2) divergence estimators of G. Let us denote these divergences respectively as follows

$$\mathsf{H}^2(p, q) = \sum_{y \in \mathcal{Y}} \left(\sqrt{p(y)} - \sqrt{q(y)} \right)^2, \quad \chi^2(p \| q) = \sum_{x \in \mathcal{Y}} \frac{(p(y) - q(y))^2}{q(y)},$$

where p, q are any two probability mass functions on the non-negative integers $\mathbb{Z}_+ \triangleq$

$\{0, 1, 2, \ldots\}$. Denote by p_n^{emp} the empirical mass function of the training samples $\{Y_1, \ldots, Y_n\}$

$$p_n^{\mathsf{emp}}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i = y\}}. \tag{3.3}$$

Then the minimum-H^2 estimator and and minimum-χ^2 divergence estimator are given by

$$\operatorname*{argmin}_Q \mathsf{H}^2(p_n^{\mathsf{emp}}, f_Q), \quad \operatorname*{argmin}_Q \chi^2(p_n^{\mathsf{emp}} \| f_Q),$$

where the minimization is done over the class of all prior distributions on $\mathbb{R}_+$. We fix $n = 50, 100, 225, 500$, simulate $(\theta_1, \ldots, \theta_n)$ independently from G and generate Y_i according to $\mathrm{Poi}(\theta_i), i = 1, \ldots, n$. Then we plot $\widehat{\theta}_{\mathrm{Robbins}}$ and $\widehat{\theta}_{\widehat{G}^{(n)}}$, where $\widehat{G}^{(n)}$ is either the NPMLE or the minimum-H^2 or the minimum-χ^2 divergence estimator, against the true values of θ_i and the Bayes estimates $\widehat{\theta}_G(Y_i)$. As shown in Fig. 3.1 all the three minimum-distance estimators provide a much more consistent approximation of the Bayes estimator compared to the Robbins estimator and can be seen as viable contenders for practical usage.

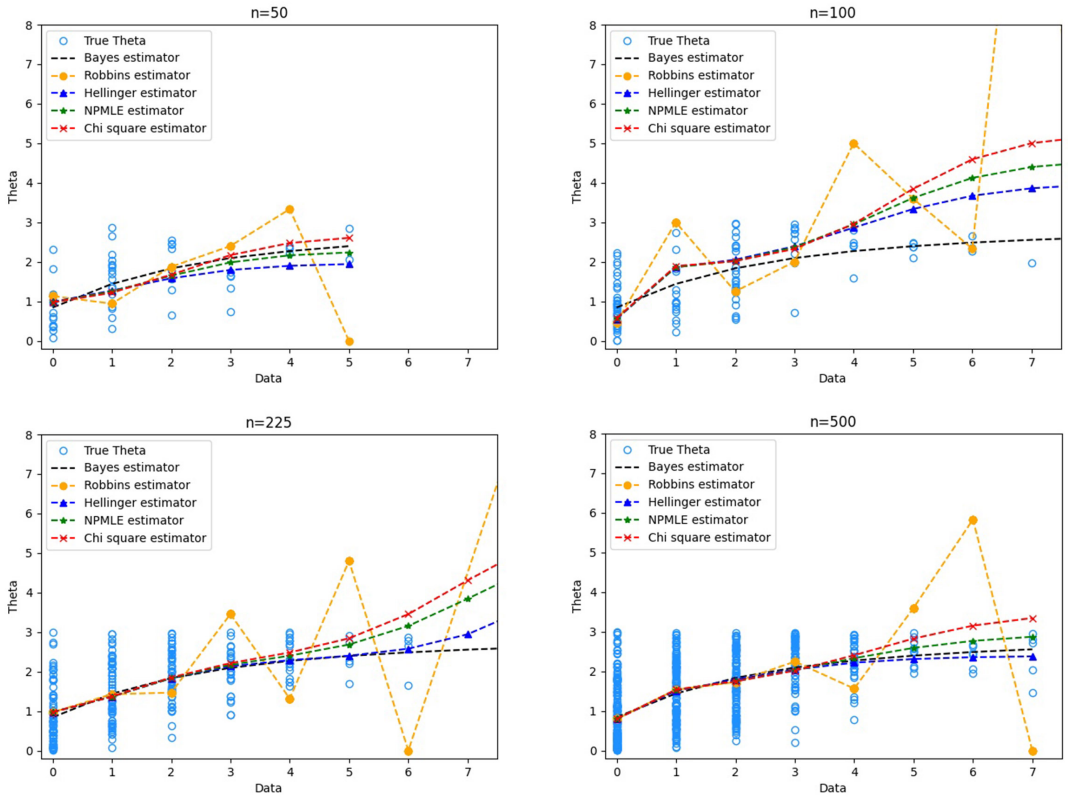

Figure 3.1: Empirical study of prediction for different estimators with Uniform$[0, 3]$ prior

In summary, several g-modeling type estimators promise excellent performances in terms of practical experiments. We show that they also happen to be minimax optimal in various classes of prior distributions.

3.1.1 Related works

Searching for a stable and smooth alternative to the classical Robbins method for the Poisson EB problem has a long history. Maritz (1966) was one of the proponents of using g-modeling estimators to resolve this problem. The author considered modeling the prior using the Gamma distribution and estimated the scale and shape parameters using a χ^2-distance minimization. Maritz (1969) studied the non-decreasing property of the oracle in Poisson EB and used non-decreasing polynomials for approximating the oracle. Lemon and Krutchkoff (1969) considered estimation of the prior using an iterative method, first using the empirical distribution of the training sample Y^n and then using corresponding posterior means of the θ_is. Although, they numerically argue that higher-order iterations might not improve the estimation loss. On a similar line, Bennett and Martz (1972) assumed the existence of density of the prior distribution and used Kernel estimates to approximate the prior. For a detailed exposition on other similar smooth EB methods, see Maritz and Lwin (2018). Our paper also analyzes the regret via such smooth EB procedures, but we use the non-parametric estimation of mixture distributions over different classes of priors.

Analysis of the NPMLE for estimating the mixture distribution is well studied in the literature. Kiefer and Wolfowitz (1956) was one of the preliminary papers to prove the consistency of the NPMLE. Later consistency of the NPMLE for econometric data was analyzed in Heckman and Singer (1984), for exponential and Weibull distributions in Jewell (1982), and for general exponential families in Pfanzagl (1988). For more recent discussions on asymptotic consistency of NPMLE for general mixtures, see Chen (2017). For the Poisson mixture setup, Lambert and Tierney (1984) showed that when the prior has bounded support and is relatively smooth at the origin, then over any finite set, the MLE and the sample proportions assign similar probabilities asymptotically. For more references on the consistency, see Chen (2017). In the current work, we study the Poisson mixture and characterize consistency in a finite-sample, minimax rate optimal way. Early analysis of NPMLE

in the finite-sample context can be attributed to Simar (1976) which showed that for the Poisson mixture case, the support size of NPMLE is bounded from above by the number of distinct entries in the sample and also established its uniqueness. Later Lindsay (1983a,b, 1995) showed that the finite support and uniqueness properties of the NPMLE are closely related to the geometry of the likelihood function. Our paper extends these results for the Poisson mixture and shows that similar properties hold for a spectrum of minimum-distance based estimators that include the NPMLE. For important finite-sample results on mixtures of continuous distributions see Polyanskiy and Wu (2020); Saha and Guntuboyina (2020); Walter and Blum (1984). In a related scenario of estimating heterogeneous mixtures of Poisson distributions, Miao et al. (2021) has explored the application of NPMLE as well.

The initial work on non-parametric estimation for the prior in the context of empirical Bayes regret analysis was carried out in Laird (1982). The author analyzed the NPMLE for the Binomial and the Normal location models (with known but different variances), and the analysis is primarily numerical. The Gaussian location-mixture setup has previously noted a finite-sample theoretical analysis of EB with NPMLE. In the revolutionary paper Zhang (2009), the authors have shown that the NPMLE based mixture density estimates are within $\mathsf{polylog}(n)$ factor of the worst-case estimation guarantees in terms of the expected squared Hellinger distance (Kim, 2014). In the follow-up work, Jiang and Zhang (2009) used this last result on the NPMLE to construct an EB oracle estimator (termed as the GMLEB) and established similar theoretical guarantees for the regret. However, due to the slack in guarantees of mixture estimation, the guarantees on regret presented therein are also suboptimal. Even though our analysis follows along the similar lines of Jiang and Zhang (2009), the striking difference in the results is that we can provide much more accurate theoretical guarantees of the mixture density and EB estimator for the Poisson model. Our improvements can be attributed to accurately estimating the mixture density in the discrete setup of Poisson. Additionally, we generalize the NPMLE based EB estimator to the minimum-distance-based classes of estimators.

3.2 Problem formulation and results

3.2.1 Notations

For any given distribution G let $\mathbb{E}_G$ and $\mathbb{P}_G$ respectively denote the expectation and probability law with respect to G. Given any prior distribution G on the Poisson mean parameter θ let $Y \sim f_G$ as before. The mean squared error of estimating $\theta \sim G$ based on a single observation $Y \sim f_\theta$ of an estimator $\widehat{\theta}$ is $\mathbb{E}_G\left[(\widehat{\theta}(Y)-\theta)^2\right]$. The Bayes error, also known as the minimum mean squared error (mmse), is given by

$$\mathrm{mmse}(G) \triangleq \inf_{\widehat{\theta}} \mathbb{E}_G\left[\left(\widehat{\theta}(Y)-\theta\right)^2\right] = \mathbb{E}_G\left[(\widehat{\theta}_G(Y)-\theta)^2\right].$$

Given any estimator $\widehat{G}^{(n)}$ of G based on $\{Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} f_G$, define the regret of $\widehat{\theta}_{\widehat{G}^{(n)}}$ as the excess mean squared error over $\mathrm{mmse}(G)$

$$\mathsf{Regret}(\widehat{G}^{(n)}; G) = \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}^{(n)}}(Y)-\theta\right)^2\right] - \mathrm{mmse}(G) = \mathbb{E}_G\left[\left(\widehat{\theta}_{\widehat{G}^{(n)}}(Y)-\widehat{\theta}_G(Y)\right)^2\right],$$

where the last equality followed using the orthogonality principle: the average risk of any estimator $\widehat{\theta}$ can be decomposed as

$$\mathbb{E}_G[(\widehat{\theta}-\theta)^2] = \mathrm{mmse}(G) + \mathbb{E}_G[(\widehat{\theta}-\widehat{\theta}_G)^2]. \tag{3.4}$$

Similarly we define the maximum regret of $\widehat{G}^{(n)}$ over the class of model distributions $\mathcal{G}$

$$\mathsf{Regret}(\widehat{G}^{(n)}; \mathcal{G}) = \sup_{G\in\mathcal{G}} \mathsf{Regret}(\widehat{G}^{(n)}; G).$$

Let $\mathcal{Z}$ be the class of all probability mass functions on $\mathbb{Z}_+$. Consider the class of divergence $\mathcal{D}$ such that any $d \in \mathcal{D}$ satisfies:

(P1) There exist $c_1, c_2 > 0$ such that $c_1\mathsf{H}^2(q_1,q_2) \le d(q_1\|q_2) \le c_2\chi^2(q_1\|q_2)$ for all discrete mass functions $q_1, q_2 \in \mathcal{Z}$

(P2) **(t,ℓ)-representation**: There exist maps $t : \mathcal{Z} \mapsto \mathbb{R}, \ell : \mathbb{R}^2 \mapsto \mathbb{R}$ such that for any two

distributions $q_1, q_2 \in \mathcal{Z}$

$$d\left(q_1 \| q_2\right) = t(q_1) + \sum_{y \in \mathcal{Y}} \ell(q_1(y), q_2(y)),$$

where $\ell(a, b)$ is decreasing and strictly convex as a function of b for each $a \geq 0$ and $\ell(0, b) = 0$ for each $b \geq 0$.

Remark 5. The divergences Kullback Leibler (KL), squared Hellinger and Chi squared belong to $\mathcal{D}$. This follows from noting that $2\mathsf{H}^2 \leq \mathrm{KL} \leq \chi^2$ and each of the divergences obtain $(t, \ell)-$representation: for squared Hellinger $t \equiv 2, \ell(a, b) = -2\sqrt{ab}$, for KL divergence $t \equiv 0, \ell(a, b) = a \log \frac{a}{b}$, for χ^2-divergence $t \equiv -1, \ell(a, b) = \frac{a^2}{b}$.

3.2.2 Results

Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f_G$ for G in some distribution class $\mathcal{G}$ and p_n^{emp} be defined as in (3.3). Let $d \in \mathcal{D}$ be such that it satisfies (P1) for some $c_1, c_2 > 0$. Consider the minimum d-distance estimator

$$\widehat{G}^{(n)} = \operatorname*{argmin}_{Q \in \mathcal{G}'} d(p_n^{\mathsf{emp}} \| f_Q) \tag{3.5}$$

where $\mathcal{G}'$ is some distribution class chosen according to the problem (when the minimization is performed over the set of all distribution functions we use the notation argmin_Q). For the purpose of this paper we will consider distribution classes that are superset of either $\mathcal{P}[0, h]$, the set of all distributions supported on $[0, h]$, or $\mathsf{SubE}(s)$, the set of all s-subexponential distributions on $\mathbb{R}_+$: $\mathsf{SubE}(s) = \left\{G : G([t, \infty)]) \leq 2e^{-t/s}, \forall t > 0\right\}$. We will analyze the empirical Bayes estimators of the form $\widehat{\theta}_{\widehat{G}^{(n)}}$. Our main results are the following.

Theorem 9 (Density estimation)**.** *Given any $h, s > 0$, there exist absolute constants $C_1 = C_1(h, c_1, c_2)$ and $C_2 = C_1(s, c_1, c_2)$ such that the following are satisfied.*

(a) Let $G \in \mathcal{P}[0, h], \mathcal{G} \supseteq \mathcal{P}[0, h]$ and $\widehat{G}^{(n)} = \operatorname{argmin}_{Q \in \mathcal{G}} d(p_n^{\mathsf{emp}} \| f_Q)$, then $\mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\right] \leq \frac{C_1}{n} \frac{\log n}{\log \log n}$.

(b) Let $G \in \mathsf{SubE}(s), \mathcal{G} \supseteq \mathsf{SubE}(s)$ and $\widehat{G}^{(n)} = \operatorname{argmin}_{Q \in \mathcal{G}} d(p_n^{\mathsf{emp}} \| f_Q)$, then $\mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\right] \leq$

$\frac{C_2}{n}\log n$.

Theorem 10 (Empirical Bayes). *Given any $h, s > 0$, there exist absolute constants $C_1 = C_1(h, c_1, c_2)$ and $C_2 = C_1(s, c_1, c_2)$ such that the following are satisfied.*

(a) If $\widehat{G}^{(n)} = \operatorname{argmin}_{Q\in\mathcal{P}[0,h]} d(p_n^{\mathsf{emp}}\|f_Q)$, then $\mathsf{Regret}(\widehat{G}^{(n)}; \mathcal{P}[0,h]) \le \frac{C_1}{n}\left(\frac{\log n}{\log\log n}\right)^2$.

(b) If $\widehat{G}^{(n)} = \operatorname{argmin}_{Q} d(p_n^{\mathsf{emp}}\|f_Q)$, then $\mathsf{Regret}(\widehat{G}^{(n)}; \mathsf{SubE}(s)) \le \frac{C_2}{n}(\log n)^3$.

The following remarks are in order:

1. For any $d \in \mathcal{D}$, both $\operatorname{argmin}_{Q\in\mathcal{P}[0,h]} d(p_n^{\mathsf{emp}}\|f_Q)$ and $\operatorname{argmin}_{Q} d(p_n^{\mathsf{emp}}\|f_Q)$ exist and are unique. The proof is given in Appendix 3.6.1.

2. The condition (P2) is only required to prove Theorem 10(b). More specifically, under the above condition on the divergence d, the minimum-distance estimator $\widehat{G}^{(n)} = \operatorname{argmin}_{Q} d(p_n^{\mathsf{emp}}\|f_Q)$ can be chosen to be supported on the bounded interval $[Y_{\min}^{(n)}, Y_{\max}^{(n)}]$ where $Y_{\min}^{(n)} = \min_{i=1}^n Y_i, Y_{\max}^{(n)} = \max_{i=1}^n Y_i$. See Lemma 12 for details. This result is essential for demonstrating the optimal regret rate of the unrestricted minimizer.

3. As mentioned before, in recent work Polyanskiy and Wu (2021) has shown that for fixed sample size n the minimax regret guarantee scales as $\Theta\left(\frac{(\log n)^2}{n(\log\log n)^2}\right)$ for $\mathcal{P}[0,h]$ class of priors and as $\Theta\left(\frac{(\log n)^3}{n}\right)$ when we consider $\mathsf{SubE}(s)$ class of priors. This establishes rate optimality of our estimators for any given h, s.

4. When d denotes the Kullback-Leibler (KL) divergence, i.e. $d(p\|q) \triangleq D(q\|p) = \sum_{y\in\mathcal{Y}} q(y)\log\frac{q(y)}{p(y)}$, the minimum-distance estimator $\widehat{G}^{(n)} = \operatorname{argmin}_{Q} d(p\|q)$ gives us the NPMLE. This follows from the expansion

$$D(p_n^{\mathsf{emp}}\|f_Q) = \sum_{y\in\mathcal{Y}} p_n^{\mathsf{emp}}(y)\log\frac{p_n^{\mathsf{emp}}(y)}{f_Q(y)} = \sum_{y\in\mathcal{Y}} p_n^{\mathsf{emp}}(y)\log p_n^{\mathsf{emp}}(y) - \frac{1}{n}\sum_{i=1}^n \log f_Q(Y_i),$$

   as minimizing the above with respect to Q is equivalent to maximizing $\frac{1}{n}\sum_{i=1}^n \log f_Q(Y_i)$ and hence the likelihood $\prod_{i=1}^n f_Q(Y_i)$.

5. The results demonstrated above hold for each of the NPMLE, the minimum H^2 distance estimator and the minimum χ^2 divergence estimator as their corresponding divergences

belong to class $\mathcal{D}$.

6. Our results hold for much more general class of estimators. Consider a generalization of the estimator in (3.5) as

$$d(p_n^{\mathsf{emp}} \| f_{\widehat{G}^{(n)}}) \leq \inf_{Q \in \mathcal{G}'} d(p_n^{\mathsf{emp}} \| f_Q) + \delta. \tag{3.6}$$

for some $\delta > 0$. Then the results corresponding to the bounded priors setup hold for $\delta \lesssim \frac{\log n}{n \log \log n}$ and the results corresponding to the subexponential priors setup hold for $\delta \lesssim \frac{\log n}{n}$. Note that $\widehat{G}^{(n)}$ is the NPMLE over $\mathcal{G}'$ if $\delta = 0$ and d is given by KL divergence. In case of NPMLE, (3.6) translates to obtaining $\widehat{G}^{(n)}$ that satisfies

$$\prod_{i=1}^{n} \left\{ \frac{f_{\widehat{G}^{(n)}}(Y_i)}{f_Q(Y_i)} \right\} \geq e^{-n\delta}, \quad \forall Q \in \mathcal{G}'.$$

This type of relaxed estimators is well known in the literature; for example, Jiang and Zhang (2009); Zhang (2009) used similar estimators for constructing the generalized maximum likelihood Empirical Bayes (GMLEB) estimators to analyze the Normal location-mixture model. These estimators enjoy a much less conservative optimization routine than the typical NPMLE while attaining similar regret guarantees.

3.3 Proof of error upper bound in density estimation

The central idea in the proof of Theorem 9 is as follows. It is straightforward to show that the density estimation error for any minimum d-distance estimator can be bounded from above by the expected distance between the empirical mass function p_n^{emp} and the data generating distribution f_G, which can be further bounded from above by the expected χ^2-distance between said quantities $\chi^2(p_n^{\mathsf{emp}} \| f_G)$. We show that the significant contribution in $\chi^2(p_n^{\mathsf{emp}} \| f_G)$ comes from the "essential support-set" of the data generating distribution, i.e. the set of minimal support-points outside which the probability of observing a sample point is vanishing at a rate $o(\frac{1}{n})$. For the the prior classes $\mathcal{P}[0, h]$ and $\mathsf{SubE}(s)$ such support-sets have sizes respectively of orders $O\left(\frac{\log n}{\log \log n}\right)$ and $O(\log n)$. Each support point from

the essential support-set has at most $\frac{1}{n}$ contribution to $\chi^2(p_n^{\mathsf{emp}}\|f_G)$ from which our results follow. The technical details are provided below.

Proof of Theorem 9. For any $K \geq 1$ and distribution G denote the tail probabilities of the Poisson mixture as $\epsilon_K(G) \triangleq \sum_{y=K}^{\infty} f_G(y)$. We will prove the following inequality

$$\mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\right] \leq \frac{4c_2}{c_1}\frac{K}{n} + \left(\frac{4c_2}{c_1} + 2n\right)\epsilon_K(G). \tag{3.7}$$

Specializing it to different distribution classes we achieve our results.

Note that $d \in \mathcal{D}$ satisfies $c_1\mathsf{H}^2(q_1, q_2) \leq d(q_1\|q_2) \leq c_2\chi^2(q_1\|q_2)$ for all $q_1, q_2 \in \mathcal{Z}$. Using the triangle inequality for the Hellinger distance, the elementary result $(a+b)^2 \leq 2(a^2+b^2)$, and utilizing the minimizer property of $\widehat{G}^{(n)}$ we get

$$\begin{aligned}
\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}}) &\leq \left(\mathsf{H}(p_n^{\mathsf{emp}}, f_{\widehat{G}^{(n)}}) + \mathsf{H}(p_n^{\mathsf{emp}}, f_G)\right)^2 \\
&\leq 2[\mathsf{H}^2(p_n^{\mathsf{emp}}, f_{\widehat{G}^{(n)}}) + \mathsf{H}^2(p_n^{\mathsf{emp}}, f_G)] \\
&\leq \frac{2}{c_1}(d(p_n^{\mathsf{emp}}\|f_{\widehat{G}^{(n)}}) + d(p_n^{\mathsf{emp}}\|f_G)) \leq \frac{4}{c_1}d(p_n^{\mathsf{emp}}\|f_G).
\end{aligned} \tag{3.8}$$

Bounding $d(p_n^{\mathsf{emp}}\|f_G)$ by $c_2\chi^2(p_n^{\mathsf{emp}}\|f_G)$ we get

$$\begin{aligned}
\mathbb{E}\left[d(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\left\{Y_{\max}^{(n)}<K\right\}}\right] &\leq c_2\sum_y \frac{\mathbb{E}\left[(p_n^{\mathsf{emp}}(y) - f_G(y))^2\mathbf{1}_{\left\{Y_{\max}^{(n)}<K\right\}}\right]}{f_G(y)} \\
&\overset{(a)}{\leq} c_2\sum_{y<K}\frac{\mathrm{Var}(p_n^{\mathsf{emp}}(y))}{f_G(y)} + c_2\sum_{y\geq K} f_G(y) \\
&= c_2\sum_{y<K}\frac{f_G(y)(1-f_G(y))}{nf_G(y)} + c_2\sum_{y\geq K} f_G(y) \leq \frac{c_2K}{n} + c_2\epsilon_K(G).
\end{aligned}$$

where (a) followed by using $\mathrm{Var}(p_n^{\mathsf{emp}}(y)) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\mathbf{1}_{\{Y_i=y\}}) = \frac{f_G(y)(1-f_G(y))}{n}$ and for all $y > Y_{\max}^{(n)}$ we get $p_n^{\mathsf{emp}}(y) = 0$. Using the union bound and the fact $\mathsf{H}^2(p,q) \leq 2, p, q \in \mathcal{Z}$ we have

$$\mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\mathbf{1}_{\left\{Y_{\max}^{(n)}\geq K\right\}}\right] \leq 2\mathbb{P}\left[Y_{\max}^{(n)} \geq K\right] \leq 2n\epsilon_K(G).$$

Combining this with the last display we use (3.8) to get

$$\mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\right] \le \mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\mathbf{1}_{\left\{Y_{\max}^{(n)}<K\right\}}\right] + \mathbb{E}\left[\mathsf{H}^2(f_G, f_{\widehat{G}^{(n)}})\mathbf{1}_{\left\{Y_{\max}^{(n)}\ge K\right\}}\right]$$
$$\le \frac{4}{c_1}\mathbb{E}\left[d(p_n^{\mathsf{emp}}\|f_G)\mathbf{1}_{\left\{Y_{\max}^{(n)}<K\right\}}\right] + 2n\epsilon_K(G) \le \frac{4c_2}{c_1}\frac{K}{n} + \left(\frac{4c_2}{c_1}+2n\right)\epsilon_K(G)$$

as required.

(a) Note that we have $G \in \mathcal{P}[0,h]$. For any $K > 2h$ using the fact that for each $y > 0$ the function $e^{-\theta}\theta^y$ is increasing in $\theta \in [0,y]$ we have

$$\epsilon_K(G) = \sum_{y=K}^{\infty}\int_0^h \frac{e^{-\theta}\theta^y}{y!}G(d\theta) \le \sum_{y=K}^{\infty}\frac{e^{-h}h^y}{y!} \le \frac{e^{-h}h^K}{K!}\sum_{y-K=0}^{\infty}\left(\frac{h}{K}\right)^{y-K} \le \frac{2e^{-h}h^K}{K!}.$$

We choose $K = \left\lceil \frac{2(2+he)\log n}{\log\log n}\right\rceil$. Using $K! \ge \left(\frac{K}{e}\right)^K$ from Stirling's formula and the fact $\log x < \frac{x}{2}$ with $x = \log\log n$ we continue the last display to get

$$\epsilon_K(G) \le 2\left(\frac{he}{K}\right)^K$$
$$\le 2\left(\frac{\log\log n}{2\log n}\right)^{\frac{2(2+he)\log n}{\log\log n}} \le 2e^{-(\log\log n - \log\log\log n)\frac{2(2+he)\log n}{\log\log n}} \le 2e^{-2\log n} \le \frac{2}{n^2} \tag{3.9}$$

as required.

(b) Choose $K = \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)}$. Then the properties of Poisson mixture (Appendix 3.6.2) imply that $\epsilon_K(G) \le \frac{3}{2n^2}$. Plugging this in (3.7) we get the result.

□

3.4 Proof of error upper bound for the empirical Bayes estimators

3.4.1 General program for regret upper bound via density estimation

The proof of our bounds in Theorem 10 relies on relating the regret to the mixture density estimation error in the squared Hellinger distance. This idea has been previously noted in

(Jiang and Zhang, 2009, Theorem 3) in the context of the Gaussian location models. For the above continuous model the authors bounded the regret from above using the squared Hellinger distance between the data generating mixture and NPMLE based estimator. We use a similar technique for general minimum distance estimators. The added benefit in our approach is that our class of minimum-distance based estimators are optimal for estimating the mixture distribution in a discrete setup. This turn out to be essential for achieving optimal upper bounds on the regret.

As a first step to achieving regret upper bounds we bound the error of estimating $\widehat{\theta}_G(Y)$ by $\widehat{\theta}_{\widehat{G}}(Y)$, where $Y \sim f_G$ and $\widehat{G}$ is any distribution function. More specifically we want to bound

$$\sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_G(y))^2 f_G(y)$$

for any given estimator $\widehat{G}$ of G. Then to obtain bound on regret for any estimator based on the training sample we average the above formula over the relevant distribution. Given any $h > 0$ let G_h denote its restriction on $[0, h]$, i.e. $G_h(A) = \frac{G(A\cap[0,h])}{G([0,h])}, A \subseteq \mathbb{R}$. Then we have the following result.

Lemma 11. *Let G be a distribution such that $\mathbb{E}_G[\theta^4] \leq M$ for some constant M. Then given any arbitrary distribution $\widehat{G}$ supported on $[0, \widehat{h}]$ and any $h > 0, K \geq 1$*

$$\sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_G(y))^2 f_G(y) \leq \left\{6(h^2 + \widehat{h}^2) + 24(h + \widehat{h})K\right\} \mathsf{H}^2(f_{\widehat{G}}, f_{G_h}) + (h + \widehat{h})^2 \epsilon_K(G_h) + \frac{(1 + 2\sqrt{2})\sqrt{(M + \widehat{h}^4)G((h, \infty))}}{G([0, h])}.$$

The core idea of the proof is as follows. It is relatively easier to bound the estimation error if the corresponding Bayes estimator is also bounded. One can force the Bayes estimator to be bounded if we use priors with bounded support. Thankfully, the extra price we pay to change the problem from estimating $\widehat{\theta}_G$ to estimating $\widehat{\theta}_{G_h}$ can be controlled for our chosen classes of priors. More specifically, using properties of the mmse and the mean squared error

from (Wu and Verdú, 2010, Lemma 2) we show

$$\sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y)-\widehat{\theta}_{G}(y))^2 f_G(y) \le \sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y)-\widehat{\theta}_{G_h}(y))^2 f_{G_h}(y) + \frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}.$$

Then we use the structure of the Bayes estimator in the Poisson case to relate the error of estimating $\widehat{\theta}_G(Y)$ by $\widehat{\theta}_{\widehat{G}}(Y)$ to the squared Hellinger distance between f_{G_h} and $f_{\widehat{G}}$

$$\sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y)-\widehat{\theta}_{G_h}(y))^2 f_{G_h}(y) \le \left\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\right\}\mathsf{H}^2(f_{\widehat{G}},f_{G_h})+(h+\widehat{h})^2\epsilon_K(G_h)$$

from which the result follows. The technical details have been provided in Appendix 3.6.3.

3.4.2 Proof of Theorem 10

Proof of Theorem 10(a). Using Lemma 11, with $\widehat{G}^{(n)} = \operatorname{argmin}_{Q\in\mathcal{P}[0,h]} d\left(p_n^{\mathsf{emp}}\|f_Q\right), \widehat{h}=h$, and Theorem 9 we get for $K=\left\lceil\frac{2(2+he)\log n}{\log\log n}\right\rceil, G\in\mathcal{P}[0,h]$

$$\begin{aligned}\mathsf{Regret}(\widehat{G}^{(n)};G) &\le \left\{12h^2+48hK\right\}\mathbb{E}_G\left[\mathsf{H}^2\left(f_G,f_{\widehat{G}^{(n)}}\right)\right]+4h^2\epsilon_K(G)\\ &\le \frac{C_0}{n}\left(\frac{\log n}{\log\log n}\right)^2+\frac{4h^2}{n^2}\end{aligned}$$

where the bound on $\epsilon_K(G)$ used in last inequality follows from (3.9) and C_0 is some absolute constant depending on h. □

Proof of Theorem 10(b). To use Lemma 11 for our purpose we need to show that there exists an $\widehat{h}$ such that $[0,\widehat{h}]$ includes the support of $\widehat{G}^{(n)}$. Indeed for any divergence d satisfying property (P2), the support of the minimum distance estimator $\widehat{G}^{(n)}$, based on training sample $Y_1,\ldots,Y_n$, is a subset of $[0,Y_{\max}^{(n)}]$ as the following lemma indicates.

Lemma 12. *Let d is a divergence satisfying (P2). Then the search of the minimizer* $\operatorname{argmin}_Q d(p_n^{\mathsf{emp}}\|f_Q)$ *reduces to minimization over distributions supported on* $[Y_{\min}^{(n)},Y_{\max}^{(n)}]$.

A proof of the above lemma is provided at the end of this section. Next, define

$$h = 4s\log n, \quad K = \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)}, \quad M = 30s^4$$

and note that from Appendix 3.6.2, for all $n \geq 2$

$$\mathbb{E}_G[\theta^4] \leq M, \quad G((h,\infty)) \leq \frac{2}{n^4}, \quad \epsilon_K(G_h) \leq \epsilon_K(G)\frac{1}{G([0,h])} \leq \frac{3}{2n^2(1-2/n^4)} \leq \frac{3}{n^2}.$$

Observe that given any d satisfying (P2) the minimimum distance estimator $\widehat{G}^{(n)} = \operatorname{argmin}_Q d(p_n^{\mathsf{emp}} \| f_Q)$ that is supported on $[0,\widehat{h}]$ where $\widehat{h} = Y_{\max}^{(n)}$. Define $B_{n,s} = \frac{64(\log n)^4+45}{\left(\log\left(1+\frac{1}{2s}\right)\right)^4}$ and note that $\mathbb{E}_G[(Y_{\max}^{(n)})^4] \leq B_{n,s}$ (see Appendix 3.6.2 for a proof). Then using Lemma 11 we get

$$\begin{aligned}
&\mathsf{Regret}(\widehat{G}^{(n)}; G) \\
&\leq \mathbb{E}_G\left[\left\{6(h^2+(Y_{\max}^{(n)})^2) + 24K(h+Y_{\max}^{(n)})\right\}\mathsf{H}^2(f_{G_h}, f_{\widehat{G}^{(n)}}) + \frac{3(h+Y_{\max}^{(n)})^2}{n^2}\right. \\
&\quad \left. + \frac{\sqrt{2}(1+2\sqrt{2})\sqrt{M+(Y_{\max}^{(n)})^4}}{n^2}\right] \\
&\overset{(a)}{\leq} \mathbb{E}_G\left[\left\{6(h^2+(Y_{\max}^{(n)})^2) + 24K(h+Y_{\max}^{(n)})\right\}\mathsf{H}^2(f_{G_h}, f_{\widehat{G}^{(n)}})\right] + \frac{6(h^2+\mathbb{E}_G[(Y_{\max}^{(n)})^2])}{n^2} \\
&\quad + \frac{\sqrt{2}(1+2\sqrt{2})\sqrt{M+\mathbb{E}_G[(Y_{\max}^{(n)})^4]}}{n^2} \\
&\overset{(b)}{\leq} 6\mathbb{E}_G\left[\left\{(h^2+(Y_{\max}^{(n)})^2) + 4K(h+Y_{\max}^{(n)})\right\}\mathsf{H}^2(f_{G_h}, f_{\widehat{G}^{(n)}})\right] + \frac{C}{n} \qquad (3.10)
\end{aligned}$$

for some absolute constant C, where inequality (a) followed from Jensen's inequality $\mathbb{E}_G\left[\sqrt{Z}\right] \leq \sqrt{\mathbb{E}_G[Z]}$ with random variable $Z = M + (Y_{\max}^{(n)})^4$ and inequality $(x+y)^2 \leq 2(x^2+y^2)$, and (b) followed by using the result $(\mathbb{E}_G[(Y_{\max}^{(n)})^2])^2 \leq \mathbb{E}_G[(Y_{\max}^{(n)})^4] \leq B_{n,s} \leq Cn$. Next we bound the expectation term in the last line of (3.10). Using the fact that the squared Hellinger

distance between any two distributions is less than 2 we get

$$\begin{aligned}
&\mathbb{E}_G\left[\left\{(h^2+(Y_{\max}^{(n)})^2)+4K(h+Y_{\max}^{(n)})\right\}\mathsf{H}^2(f_{G_h},f_{\widehat{G}^{(n)}})\right]\\
&=\mathbb{E}_G\left[\left\{(h^2+(Y_{\max}^{(n)})^2)+4K(h+Y_{\max}^{(n)})\right\}\mathsf{H}^2(f_{G_h},f_{\widehat{G}^{(n)}})\mathbf{1}_{\left\{Y_{\max}^{(n)}\le 2K\right\}}\right]\\
&\quad+\mathbb{E}_G\left[\left\{(h^2+(Y_{\max}^{(n)})^2)+4K(h+Y_{\max}^{(n)})\right\}\mathsf{H}^2(f_{G_h},f_{\widehat{G}^{(n)}})\mathbf{1}_{\left\{Y_{\max}^{(n)}>2K\right\}}\right]\\
&\le(h^2+4KH+12K^2)\mathbb{E}_G\left[\mathsf{H}^2\left(f_{G_h},f_{\widehat{G}^{(n)}}\right)\right]\\
&\quad+2\mathbb{E}_G\left[\left\{(h^2+(Y_{\max}^{(n)})^2)+4K(h+Y_{\max}^{(n)})\right\}\mathbf{1}_{\left\{Y_{\max}^{(n)}>2K\right\}}\right] \qquad (3.11)
\end{aligned}$$

For bounding the expectation in the first term of (3.11) we use triangle inequality and $(x+y)^2\le 2(x^2+y^2)$ to get

$$\mathsf{H}^2(f_{G_h},f_{\widehat{G}^{(n)}})\le 2\left\{\mathsf{H}^2(f_G,f_{\widehat{G}^{(n)}})+\mathsf{H}^2(f_{G_h},f_G)\right\}.$$

We get $\mathbb{E}_G\left[\mathsf{H}^2(f_G,f_{\widehat{G}^{(n)}})\right]\le\frac{C'\log n}{n}$ using Theorem 9 for absolute constant C'. To bound $\mathsf{H}^2(f_{G_h},f_G)$ for any $G\in\mathsf{SubE}(s)$ we use the total variation distance between f_{G_h},f_G to get

$$\begin{aligned}
\mathsf{H}^2(f_{G_h},f_G)&\le\sum_{y=0}^{\infty}|f_{G_h})(y)-f_G(y)|\\
&\le\sum_{y=0}^{K-1}|f_{G_h}(y)-f_G(y)|+\sum_{y=K}^{\infty}f_G(y)+\sum_{y=K}^{\infty}f_{G_h}(y)\\
&=\sum_{y=0}^{K-1}\left|\frac{\int_{\theta=0}^{h}\frac{e^{-\theta}\theta^y}{y!}G(d\theta)}{G([0,h])}-f_G(y)\right|+\epsilon_K(G)+\sum_{y=K}^{\infty}\frac{\int_{\theta=0}^{h}\frac{e^{-\theta}\theta^y}{y!}G(d\theta)}{G([0,h])}\\
&\le\sum_{y=0}^{K-1}\left(\left|\frac{f_G(y)}{G([0,h])}-f_G(y)\right|+\frac{\int_{\theta>h}^{\infty}\frac{e^{-\theta}\theta^y}{y!}G(d\theta)}{G([0,h])}\right)+\epsilon_K(G)+\frac{\epsilon_K(G)}{G([0,h])}\\
&\le\sum_{y=0}^{K-1}\frac{G((h,\infty))}{G([0,h])}f_G(y)+\frac{G((h,\infty))}{G([0,h])}+\epsilon_K(G)\left(1+\frac{1}{G([0,h])}\right)\le\frac{7}{n^2}.
\end{aligned}$$

Next we bound the second term in (3.11). Note that by union bound and Appendix 3.6.2 we have $\mathbb{P}_G\left[Y_{\max}^{(n)}\ge 2K\right]\le n\epsilon_{2K}(G)\le\frac{3}{2n^3}$. Using this and the Cauchy-Schwarz inequality

we get

$$\begin{aligned}
&\mathbb{E}_G\left[\left\{6(h^2+(Y_{\max}^{(n)})^2)+24K(h+Y_{\max}^{(n)})\right\}\mathbf{1}_{\left\{Y_{\max}^{(n)}>2K\right\}}\right]\\
\leq&\sqrt{\mathbb{E}_G\left[\left\{6(h^2+(Y_{\max}^{(n)})^2)+24K(h+Y_{\max}^{(n)})\right\}^2\right]\mathbb{P}_G\left[Y_{\max}^{(n)}\geq 2K\right]}\\
\leq&\sqrt{\mathbb{E}_G\left[\left\{6(h^2+(Y_{\max}^{(n)})^2)+24K(h+(Y_{\max}^{(n)}))\right\}^2\right]n\epsilon_{2K}(G)}\\
\leq&\frac{\sqrt{3}}{\sqrt{2}n^{3/2}}\sqrt{\mathbb{E}_G\left[\left\{4(h^4+(Y_{\max}^{(n)})^4)+16K^2(h^2+(Y_{\max}^{(n)})^2)\right\}\right]}\\
\leq&\frac{\sqrt{3}}{\sqrt{2}n^{3/2}}\sqrt{h^4+B_{n,s}+4K^2(h^2+\sqrt{B_{n,s}})}\leq\frac{c}{n^{3/2}}
\end{aligned}$$

for some absolute constant c. Plugging the bounds back in (3.11) in view of (3.10) we get the result. □

Proof of Lemma 12. Given any distribution Q define $\widetilde{Q}$ as

$$d\widetilde{Q}(\theta)=\begin{cases}Q([0,Y_{\min}^{(n)}]), & \theta=Y_{\min}^{(n)},\\ dQ(\theta), & Y_{\min}^{(n)}<\theta<Y_{\max}^{(n)},\\ Q([Y_{\max}^{(n)},\infty)), & \theta=Y_{\max}^{(n)}.\end{cases}$$

In other words, $\widetilde{Q}$ accumulates masses of Q over the intervals $[0,Y_{\min}],[Y_{\max},\infty)$ on the points $Y_{\min}$ and $Y_{\max}$ respectively. As $v(\theta)=e^{-\theta}\theta^x$ is increasing in $\theta\in[0,x]$ and decreasing in $\theta\in[x,\infty)$ we get for each $i=1,\ldots,n$

$$\begin{aligned}
f_Q(Y_i)&=\int\frac{e^{-\theta}\theta^{Y_i}}{Y_i!}dQ(\theta)\\
&=\int_{0<\theta\leq Y_{\min}^{(n)}}\frac{e^{-\theta}\theta^{Y_i}}{Y_i!}dQ(\theta)+\int_{Y_{\min}^{(n)}<\theta<Y_{\max}^{(n)}}\frac{e^{-\theta}\theta^{Y_i}}{Y_i!}dQ(\theta)+\int_{Y_{\max}^{(n)}\geq\theta}\frac{e^{-\theta}\theta^{Y_i}}{Y_i!}dQ(\theta)\\
&\leq Q([0,Y_{\min}^{(n)}])\frac{e^{-Y_{\min}^{(n)}}\left(Y_{\min}^{(n)}\right)^{Y_i}}{Y_i!}+\int_{Y_{\min}^{(n)}<\theta<Y_{\max}^{(n)}}\frac{e^{-\theta}\theta^{Y_i}}{Y_i!}dQ(\theta)\\
&\quad+Q([Y_{\min}^{(n)},\infty))\frac{e^{-Y_{\max}^{(n)}}\left(Y_{\max}^{(n)}\right)^{Y_i}}{Y_i!}\\
&=\int\frac{e^{-\theta}\theta^{Y_i}}{Y_i!}d\widetilde{Q}(\theta)=f_{\widetilde{Q}}(Y_i).
\end{aligned}$$

Note that we can write

$$
\begin{aligned}
d(p_n^{\mathsf{emp}} \| f_G) &= t(p_n^{\mathsf{emp}}) + \sum_{y \in \mathcal{Y}} \ell(p_n^{\mathsf{emp}}(y), f_G(y)) \\
&= t(p_n^{\mathsf{emp}}) + \frac{1}{n} \sum_{y \in \mathcal{Y}: p_n^{\mathsf{emp}}(y) > 0} \frac{\sum_{i=1}^n \mathbf{1}_{\{Y_i = y\}} \ell(p_n^{\mathsf{emp}}(y), f_G(y))}{p_n^{\mathsf{emp}}(y)} \\
&= t(p_n^{\mathsf{emp}}) + \frac{1}{n} \sum_{i=1}^n \sum_{y \in \mathcal{Y}: p_n^{\mathsf{emp}}(y) > 0} \frac{\mathbf{1}_{\{Y_i = y\}} \ell(p_n^{\mathsf{emp}}(Y_i), f_G(Y_i))}{p_n^{\mathsf{emp}}(Y_i)} \\
&= t(p_n^{\mathsf{emp}}) + \frac{1}{n} \sum_{i=1}^n \frac{\ell(p_n^{\mathsf{emp}}(Y_i), f_G(Y_i))}{p_n^{\mathsf{emp}}(Y_i)}, \qquad (3.12)
\end{aligned}
$$

which implies

$$
\begin{aligned}
\sum_{y \in \mathcal{Y}} \ell(p_n^{\mathsf{emp}}(y), f_Q(y)) &= \frac{1}{n} \sum_{i=1}^n \frac{\ell\left(p_n^{\mathsf{emp}}(Y_i), f_Q(Y_i)\right)}{p_n^{\mathsf{emp}}(Y_i)} \\
&\geq \frac{1}{n} \sum_{i=1}^n \frac{\ell\left(p_n^{\mathsf{emp}}(Y_i), f_{\widetilde{Q}}(Y_i)\right)}{p_n^{\mathsf{emp}}(Y_i)} = \sum_{y \in \mathcal{Y}} \ell(p_n^{\mathsf{emp}}(y), f_{\widetilde{Q}}(y)),
\end{aligned}
$$

and hence given any Q we can produce $\widetilde{Q}$ such that $d(p_n^{\mathsf{emp}} \| f_{\widetilde{Q}}) \leq d(p_n^{\mathsf{emp}} \| f_Q)$. As a result we can choose $\widehat{G}^{(n)} = \operatorname{argmin}_Q d(p_n^{\mathsf{emp}} \| f_Q)$ to be supported on $[Y_{\min}^{(n)}, Y_{\max}^{(n)}]$. □

3.5 Numerical experiments

In this section we analyze the performances of the empirical Bayes estimators based on the minimum-H^2, the minimum-χ^2, and the minimum-KL divergence estimator (i.e., the NPMLE). We compare them against the Robbins estimator and also draw comparisons among their individual performances. Unlike the Robbins estimator the minimum-distance based estimators usually do not have any closed form solutions. To compute the solution we rely on the optimality conditions.

3.5.1 First-order optimality condition and algorithm

In the numerical studies we consider the unrestricted minimizers of the form $\widehat{G}^{(n)} = \operatorname{argmin}_Q d(p_n^{\mathsf{emp}} \| f_Q)$. Given any divergence d let $\widehat{G}^{(n)}$ be any such minimum d-distance estimator. Given $d \in \mathcal{D}$

the support of any such minimizer has size at most n (see Appendix 3.6.1) and Lemma 12 implies that the support points lie in the interval $[0, Y_{\max}]$. For any $\theta \in \mathbb{R}_+$ let δ_θ denote the degenerate distribution at θ. Then the first-order optimality condition for $\widehat{G}^{(n)}$ is given by (both necessary and sufficient for establishing optimality of $\widehat{G}^{(n)}$)

$$d(p_n^{\mathsf{emp}} \| f_{\widehat{G}^{(n)}}) \leq d(p_n^{\mathsf{emp}} \| f_{(1-\epsilon)\widehat{G}^{(n)} + \epsilon\delta_\theta}), \quad \theta \in \mathbb{R}_+, \epsilon \in [0,1].$$

When $d \in \mathcal{D}$ is either of H^2, χ^2 or the KL divergence $d(p_n^{\mathsf{emp}} \| f_{(1-\epsilon)\widehat{G}^{(n)} + \epsilon\delta_\theta})$ is differentiable in ϵ, a consequence of the corresponding $t - \ell$ representation via smooth ℓ functions. This implies

$$\frac{d}{d\epsilon} d(p_n^{\mathsf{emp}} \| f_{(1-\epsilon)\widehat{G}^{(n)} + \epsilon\delta_\theta}) \Big|_{\epsilon=0} \geq 0$$

Let $y_1, \ldots, y_q$ be the distinct values in the training sample $\{Y_1, \ldots, Y_n\}$. Then in view of the last inequality we get

$$\sum_{i=1}^{q} \frac{d}{df} \ell(p_n^{\mathsf{emp}}(y_i), f) \Big|_{f = f_{\widehat{G}^{(n)}}(y_i)} (f_\theta(y_i) - f_{\widehat{G}^{(n)}}(y_i)) \geq 0.$$

The locations of θ where the above inequality is satisfied are given by the solutions to

$$\frac{d}{d\theta} \left\{ \sum_{i=1}^{q} \frac{d}{df} \ell(p_n^{\mathsf{emp}}(y_i), f) \Big|_{f = f_{\widehat{G}^{(n)}}(y_i)} (f_\theta(y_i) - f_{\widehat{G}^{(n)}}(y_i)) \right\} = 0.$$

Simplifying the above equation we get that the support points of $\widehat{G}^{(n)}$ satisfies the following polynomial equation in θ

$$e^{-\theta} \sum_{i=1}^{q} w_i(\widehat{G}^{(n)}) \left(y_i \theta^{y_i - 1} - \theta^{y_i} \right) = 0, \quad w_i(\widehat{G}^{(n)}) = \frac{\frac{d}{df} \ell(p_n^{\mathsf{emp}}(y_i), f) \Big|_{f = f_{\widehat{G}^{(n)}}(y_i)}}{y_i!}.$$

Based on the above conditions we construct the following algorithm to approximate the minimum-distance estimators.

Algorithm 1 Computing the minimum d-distance estimators

Input: Data-points $Y_1, \ldots, Y_n$. Target distribution $G_{\boldsymbol{\theta},\boldsymbol{\mu}} = \sum_j \mu_j \delta_{\theta_j}$. Divergence d with $t-\ell$ decomposition $d(q_1\|q_2) = t(q_1) + \sum_{y\in\mathcal{Y}} \ell(q_1(y), q_2(y))$. Objective function to minimize

$$v(\boldsymbol{\mu},\boldsymbol{\theta}) = d(p_n^{\mathsf{emp}}\|f_G) = t(p_n^{\mathsf{emp}}) + \sum_{i=1}^{q} \ell(p_n^{\mathsf{emp}}(y_i), \sum_j \mu_j f_{\theta_j}(y_i)).$$

Initiate 1000 equidistant points in $[0, Y_{\max}]$ as $\boldsymbol{\theta}$ and initiate probability assignment $\boldsymbol{\mu}$ randomly. Number of steps N to guarantee convergence of optimization rule.

Output: $\boldsymbol{\theta}, \boldsymbol{\mu}$ such that $d(p_n^{\mathsf{emp}}\|f_{G_{\boldsymbol{\theta},\boldsymbol{\mu}}})$ is minimized.

1: **for** $t = 1, \ldots, N$ **do**
2: Update $\boldsymbol{\mu} \leftarrow \text{argmin}_{\boldsymbol{\mu}} \sum_{i=1}^{q} \ell(p_n^{\mathsf{emp}}(y_i), \sum_j \mu_j f_{\theta_j^{(t-1)}}(y_i))$
3: $\widehat{G}_{\boldsymbol{\theta},\boldsymbol{\mu}} = \sum_j \widehat{\mu}_j \delta_{\theta_j}$
4: Update $\boldsymbol{\theta}$ as roots of the polynomial $\sum_{i=1}^{q} w_i(\widehat{G}_{\boldsymbol{\theta},\boldsymbol{\mu}}) \left(y_i \theta^{y_i - 1} - \theta^{y_i}\right)$.
5: $t \leftarrow t + 1$.
6: **end for**

3.5.2 Real-data analysis: Prediction of hockey goals

We study the data on total number of goal scored in the National Hockey League for the seasons 2017-18 and 2018-19 (the data is available at https://www.hockey-reference.com/). We consider the statistics of $n = 745$ players, for whom the data were collected for both the seasons. Let Y_i be the total number of goal scored by the i^{th} player in the season 2017-18. We model Y_i as independently distributed $\text{Poi}(\theta_i)$, where θ_i's are independently distributed according to some distribution G on the non-negative integers. Based on the observations we intend to predict the goal scored by the i^{th} player in the season 2018-19 using $\widehat{\theta}_G$. In Fig. 3.2 we plot the EB estimators based on the Robbins method, the minimum H^2, the minimum-χ^2 distance estimator and the NPMLE against the 2017-18 data (denoted as "Past" on the x-axis) and compare their estimates against the real values of goals in 2018-19 (denoted by "Future" on the y-axis). On the left figure we show the comparison for all the estimators, whereas the figure on the right presents more detailed comparison for the minimum distance based estimators.

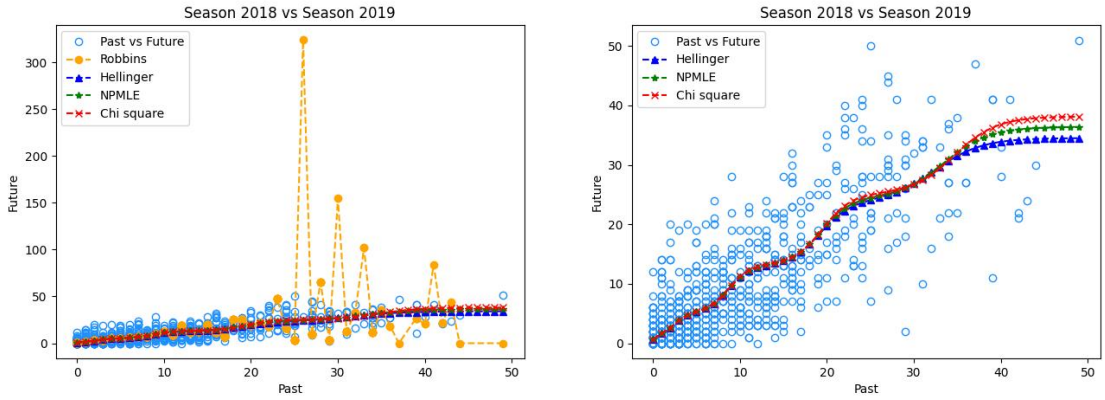

Figure 3.2: Robbins vs. minimum-distance EB: Experiments with hockey goals

From the first figure it is visible that there exist many individuals for whom the Robbins methods produces significantly worse predictions compared to all the minimum distance methods. This difference is significant for values of scored goals which have lower sample representations. On the other hand the separate comparison of the minimum distance based predictions shows that their behavior are mostly comparable except near the tail end of the data-points. Upon computing the root mean squared error (RMSE) and the Mean absolute deviation (MAD) with respect to the true goal values in 2018-19 it turns out that the minimum H^2-distance based EB estimator marginally outperforms the other minimum-distance based EB estimators.

Methods	Robbins	minimum-H^2	NPMLE	minimum-χ^2
RMSE	15.589	6.023	6.0386	6.053
MAD	6.639	4.368	4.381	4.389

Table 3.1: Robbins vs. minimum-distance: Prediction error comparison

3.5.3 Simulation studies: Unbounded priors

Next we extend the simulation studies presented in the Introduction section to priors with unbounded supports. The three g-modeling estimators provide comparatively better performance than the Robbins estimator in the unbounded prior setup as well. To demonstrate this we consider both discrete and continuous priors. For the discrete setup we choose the mixture of $\text{Poi}(1), \text{Poi}(2)$ and $\text{Poi}(8)$ with weights $(0.2, 0.3, 0.5)$ and for the continuous setup we choose Gamma distribution with scale parameter 2 and shape parameter 4, i.e. with

prior density $f(x) = \frac{1}{96}x^3e^{-\frac{x}{2}}$. In both of the cases we simulate $\{\theta_i\}_{i=1}^{600}$ independently from the prior distribution and correspondingly generate data $Y_i \sim \text{Poi}(\theta_i)$. For each of the priors we calculate the Bayes estimator numerically (denoted by the black dashed line in the plots). Then from the generated datasets we compute the Robbins estimator, the NPMLE based EB estimator, the H^2-distance based EB estimator and the χ^2-distance based EB estimator. All the estimators are then plotted against θ and the data (Fig. 3.3). As expected, the Robbins estimator shows high deviation from the true θ values in many instances whereas the minimum-distance based estimators are much more stable.

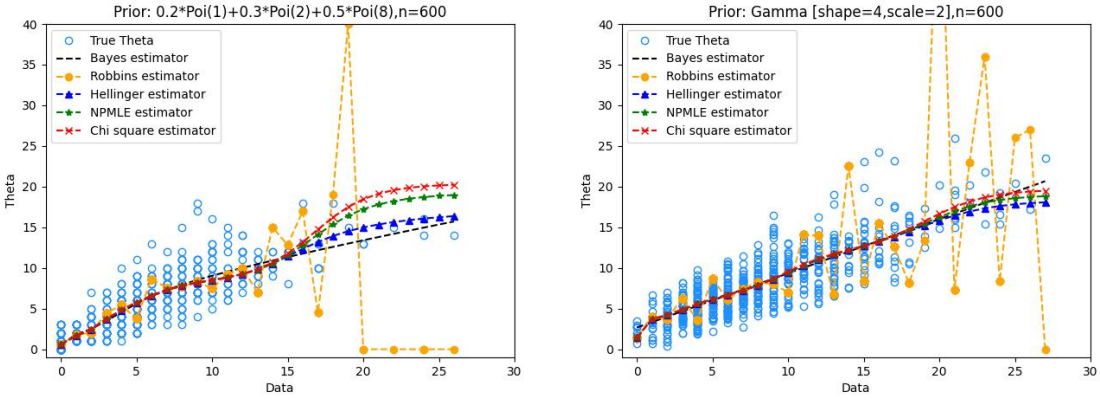

Figure 3.3: Robbins vs. minimum-distance: Unbounded priors

A natural follow up question in this context is to search for characteristics of the priors that might help to differentiate between performances of the minimum-distance estimators. We aim to analyze the effect of heavy tail distributions in this context. For this purpose we focus on the exponential distributions parameterized by scale (α) and with density $g_\alpha(x) = \frac{1}{\alpha}e^{-x/\alpha}$. Note that the higher values of α generate distributions with heavier tails. We consider three values of α: 0.3,1.05 and 2. At each such value we estimate the regret for training sample sizes n in the range $[50, 300]$. Write the regret as $\mathbb{E}_G[(\widehat{\theta}(Y) - \widehat{\theta}_{\widehat{G}^{(n)}}(Y))^2]$, where $\widehat{G}^{(n)}$ is any of the three minimum-distance estimates based on training sample size n. The Bayes estimator $\widehat{\theta}_G(y)$ is computed for each y by numerically calculating the marginals. For every triplet (α, n) we replicate the following experiment 500 times for each minimum-distance method:

- Generate $\{\theta_i\}_{i=1}^n$ and $Y_i \sim \text{Poi}(\theta_i)$,

- Calculate $\widehat{G}^{(n)}$,
- Generate independently $\theta \sim G, Y \sim \text{Poi}(Y)$,
- Calculate $(\widehat{\theta}_G(Y) - \widehat{\theta}_{\widehat{G}^{(n)}}(Y))^2$.

Then we take the average of $(\widehat{\theta}_G(Y) - \widehat{\theta}_{\widehat{G}^{(n)}}(Y))^2$ values from all the 500 replications to estimate $\mathbb{E}_G[(\widehat{\theta}(Y) - \widehat{\theta}_{\widehat{G}^{(n)}}(Y))^2]$. We plot the regrets against training sample size below (Fig. 3.4).

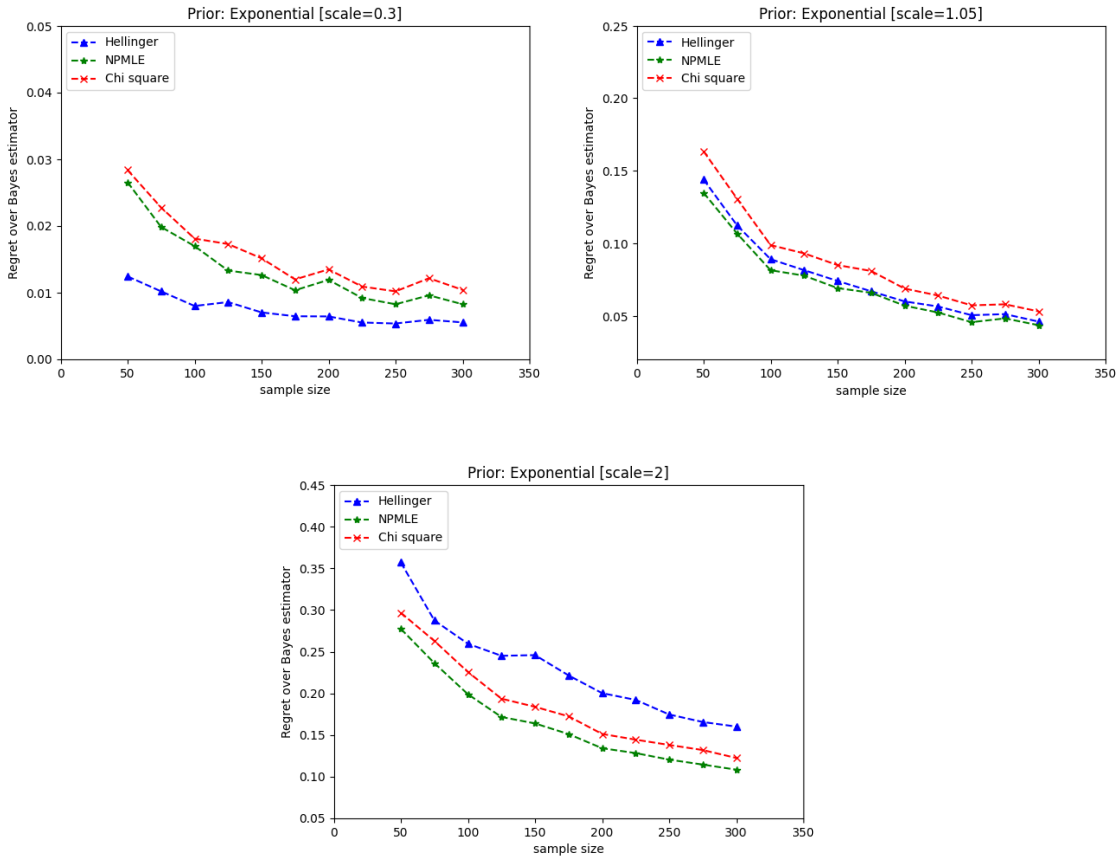

Figure 3.4: Comparison of minimum-distance estimators

We observe that that minimum-H^2 based estimator outperforms the other estimators when the scale of the exponential distribution is small. As the tails of the prior distributions become heavier, the performance of the minimum-H^2 based estimator gets worse and the NPMLE based estimator comes out as a better choice.

3.6 Appendix

3.6.1 Existence and uniqueness of minimum distance estimators

Theorem 13. *Fix any divergence $d \in \mathcal{D}$. Given training sample $Y_1, \ldots, Y_n$ and corresponding empirical distribution p_n^{emp} there exist unique choices for* $\operatorname{argmin}_Q d(p_n^{\mathsf{emp}} \| f_Q)$ *and* $\operatorname{argmin}_{Q \in \mathcal{P}[0,h]} d(p_n^{\mathsf{emp}} \| f_Q)$*, and the solutions have support set sizes bounded from above by the number of distinct sample points.*

Proof. We first note that in view of Lemma 12 we have

$$\operatorname*{argmin}_Q d(p_n^{\mathsf{emp}} \| f_Q) = \operatorname*{argmin}_{\mathcal{P}[0,Y_{\max}]} d(p_n^{\mathsf{emp}} \| f_Q).$$

Hence given any sample observations, without loss of generality, it suffices to analyze existence and uniqueness of $\operatorname{argmin}_{Q \in \mathcal{P}[0,h]} d(p_n^{\mathsf{emp}} \| f_Q)$ for any arbitrary $h > 0$. Suppose that the distinct values in the sample $Y_1, \ldots, Y_n$ are given by $y_1, \ldots, y_q, q \leq N$. Consider the strictly convex function $v(f_1, \ldots, f_n) = \sum_{i=1}^q \ell(p_n^{\mathsf{emp}}(y_i), f_i)$ (which follows from the strict convexity of the ℓ function in its second argument) on the set

$$S = \{(f_G(y_1), \ldots, f_G(y_q)) : G \in \mathcal{P}[0,h]\}.$$

Note that S is convex and compact, which follow from the fact that (see existence argument of (Simar, 1976, Section 3.1) for details) S can be written as closed convex hull of the bounded set $\{f_\theta(y_1), \ldots, f_\theta(y_q) : \theta \in [0,h]\}$. Due to strict convexity of $v(f_1, \ldots, f_q)$ we get that there exists a unique point $(f_1^*, \ldots, f_q^*)$ on S where $v(f_1, \ldots, f_q)$ achieves its minimum. In view of (3.12) we get that any G that minimizes $d(p_n^{\mathsf{emp}} \| f_G)$ (need not be unique) also satisfies

$$f_G(y_j) = f_j^*, j = 1, \ldots, q \tag{3.13}$$

Let $\widehat{G}^{(n)}$ be one such minimizer (it is known that an q-atomic minimizer exists thanks to the Fenchel-Eggleston-Caratheodory theorem (Eggleston, 1966)). For any $\theta \in \mathbb{R}_+$ let δ_θ denote the atomic distribution on θ. Then from the first order optimality condition we get that for

all $\theta \in \mathbb{R}_+, \epsilon \in [0,1]$

$$d(p_n^{\mathsf{emp}} \| f_{\widehat{G}^{(n)}}) \leq d(p_n^{\mathsf{emp}} \| f_{(1-\epsilon)\widehat{G}^{(n)}+\epsilon\delta_\theta}).$$

This implies $\frac{d}{d\epsilon} d(p_n^{\mathsf{emp}} \| f_{(1-\epsilon)\widehat{G}^{(n)}+\epsilon\delta_\theta})\Big|_{\epsilon=0} \geq 0$, and hence

$$\sum_{i=1}^{q} \frac{d}{df} \ell(p_n^{\mathsf{emp}}(y_i), f)\Big|_{f=f_i^*} (f_\theta(y_i) - f_i^*) \geq 0.$$

As ℓ is decreasing in second coordinate (and hence $\frac{d}{df}\ell(p_n^{\mathsf{emp}}(y_i), f) \leq 0$ for all $f \in \mathbb{R}_+, i = 1, \ldots, q$), rearranging the terms we get

$$\sum_{i=1}^{q} \left(\frac{\frac{1}{y_i!} \frac{d}{df} \ell(p_n^{\mathsf{emp}}(y_i), f)\Big|_{f=f_i^*}}{\sum_{i=1}^{q} \frac{d}{df} \ell(p_n^{\mathsf{emp}}(y_i), f)\Big|_{f=f_i^*} f_i^*} \right) \theta^{y_i} \leq e^\theta.$$

Hence given any optimizer $\widehat{G}^{(n)}$ its support points satisfy equality in the last display. Using (Simar, 1976, Lemma 3.1) we get that there are at most $m(\leq q)$ different θ_i's (denote them by $\theta_1 \ldots, \theta_m$) for which equality holds in the last display. This implies given any optimizer $\widehat{G}^{(n)}$ its support points form a subset of $\{\theta_1 \ldots, \theta_m\}$. Let w_i be the weight $\widehat{G}^{(n)}$ puts on θ_i. Then in view of (3.13) we get that

$$\sum_{j=1}^{m} w_j e^{-\theta_j} \theta_j^{y_i} = f_i^* y_i!, \quad i = 1, \ldots, q.$$

The matrix $\{\theta_j^{y_i} : j = 1, \ldots, m, i = 1, \ldots, q\}$ has full column rank, and hence the vector $\{w_1 e^{-\theta_i}\}_{i=1}^m$ can be solve uniquely. This implies unique solution of $(w_i, \ldots, w_m)$ and as a consequence uniqueness of the optimizer $\widehat{G}^{(n)}$ as well. This finishes the proof. □

3.6.2 Properties of the subexponential distributions

Lemma 14. *For any $G \in \mathsf{SubE}(s)$ random variables the followings are satisfied.*

(i) If $\theta \sim G$

$$\mathbb{P}[\theta \geq c_1 \log n] \leq 2n^{-\frac{c_1}{s}}, \quad \mathbb{E}[\theta^4] \leq 30s^4.$$

(ii) If $\{Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} f_G$

$$\epsilon_K(G) = \mathbb{P}\left[Y_1 \geq K\right] \leq \frac{3}{2} e^{-K \log\left(1+\frac{1}{2s}\right)}, \quad \mathbb{E}\left[\left(\max_{i=1}^n Y_i\right)^4\right] \leq \frac{64(\log n)^4 + 45}{\left(\log\left(1+\frac{1}{2s}\right)\right)^4}.$$

Proof. (i) Using tail property of $\mathsf{SubE}(s)$ distributions

$$\mathbb{P}\left[\theta \geq c_1 \log n\right] \leq 2e^{-\frac{c_1 \log n}{s}} \leq 2n^{-\frac{c_1}{s}}.$$

For the expectation term using $z^3 \leq 15 e^{\frac{z}{2}}, z \in \mathbb{R}$ we have

$$\mathbb{E}[\theta^4] = 4\int y^3 \mathbb{P}\left[\theta > y\right] dy \leq 2\int y^3 e^{-\frac{y}{s}} dy \leq 30 s^4 \int e^{-\frac{z}{2}} dz. \leq 30 s^4.$$

(ii) Using $\mathbb{E}_{Z\sim \mathrm{Poi}(\theta)}\left[e^{Zt}\right] = e^{\theta(e^t-1)}, t > 0$ and denoting $c(s) = \log\frac{1+2s}{2s}$ we have

$$\mathbb{E}\left[e^{Y_1 c(s)}\right] = \mathbb{E}_{\theta\sim G}\left[\mathbb{E}_{Y_1\sim \mathrm{Poi}(\theta)}\left[e^{Y_1 c(s)}\middle|\theta\right]\right] = \mathbb{E}_G\left[e^{\frac{\theta}{2s}}\right] = \int e^{\theta/2s} G(d\theta)$$
$$= \int_\theta \int_{x<\theta} \frac{e^{x/2s}}{2s} dx G(d\theta) = \int_x \frac{e^{x/2s}}{2s} G([x,\infty)) dx \overset{(a)}{\leq} \int_{x<0} \frac{e^{x/2s}}{2s} dx + \int_{x>0} \frac{e^{-x/2s}}{s} dx \leq \frac{3}{2}$$

where (a) followed by using tail bound for $\mathsf{SubE}(s)$ distribution G. In view of Markov inequality

$$\mathbb{P}\left[Y_1 \geq K\right] \leq \mathbb{E}\left[e^{Y_1 c(s)}\right] e^{-c(s)K} \leq \frac{3}{2} e^{-K\log\left(1+\frac{1}{2s}\right)}.$$

For the expectation term we have for any $L > 0$

$$\begin{aligned}
\mathbb{E}\left[\left(\max_{i=1}^{n} Y_i\right)^4\right] &= 4\int y^3 \mathbb{P}\left[\max_{i=1}^{n} Y_i > y\right] \\
&\leq 4L^4 + n\int_{y>L} y^3 \mathbb{P}\left[Y_1 > y\right] dy \\
&\leq 4L^4 + \frac{3n}{2}\int_{y>L} y^3 e^{-y\log\left(1+\frac{1}{2s}\right)} dy \\
&\leq 4L^4 + \frac{3n}{2\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4}\int_{z>L\log\left(1+\frac{1}{2s}\right)} z^3 e^{-z} dz \\
&\leq 4L^4 + \frac{45n}{2\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4}\int_{z>L\log\left(1+\frac{1}{2s}\right)} e^{-z/2} dz \leq 4L^4 + \frac{45ne^{-\frac{L}{2}\log\left(1+\frac{1}{2s}\right)}}{\left\{\log\left(1+\frac{1}{2s}\right)\right\}^4}
\end{aligned}$$

where c_2 is an absolute constant. Choosing $L = \frac{2\log n}{\log\left(1+\frac{1}{2s}\right)}$ we get the desired result.

□

3.6.3 Proof of Lemma 11

Let $\theta \sim G, Y|\theta \sim f_\theta$ and $\{Y, \theta\}$. Then we can write

$$\mathbb{E}_G\left[\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_G(Y)\right]^2 = \sum_{y=0}^{\infty}(\widehat{\theta}_{\widehat{G}}(y) - \widehat{\theta}_G(y))^2 f_G(y).$$

Fix $h > 0$ and note the following

- (Wu and Verdú, 2010, Lemma 2) $\text{mmse}(G_h) \leq \frac{\text{mmse}(G)}{G([0,h])}$,
- $\text{mmse}(G) \leq \sqrt{\mathbb{E}[\theta^4]} \leq \sqrt{M}$, and
- For any fixed distribution $\widehat{G}$

$$\begin{aligned}
\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] &\leq \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2 \mathbf{1}_{\{\theta \leq h\}}\right] + \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2 \mathbf{1}_{\{\theta > h\}}\right] \\
&\overset{(a)}{\leq} \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2 \middle| \theta \leq h\right] + \sqrt{\mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^4\right]\mathbb{E}_G\left[\mathbf{1}_{\{\theta > h\}}\right]} \\
&\overset{(b)}{\leq} \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] + \sqrt{8(\widehat{h}^4 + \mathbb{E}_G[\theta^4])G((h,\infty))} \\
&= \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] + \sqrt{8(\widehat{h}^4 + M)G((h,\infty))}.
\end{aligned}$$

where step (a) followed by the Cauchy-Schwarz inequality and step (b) followed as $(x+y)^4 \le 8(x^4+y^4)$ for any $x, y \in \mathbb{R}$.

In view of the above and the orthogonality relation (3.4), with $\widehat{\theta} = \widehat{\theta}_{\widehat{G}}$, we have

$$
\begin{aligned}
&\mathbb{E}_G\left[\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_G(Y)\right]^2 \\
&= \mathbb{E}_G\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] - \mathrm{mmse}(G) \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \theta)^2\right] - \mathrm{mmse}(G_h) + \mathrm{mmse}(G_h) - \mathrm{mmse}(G) + \sqrt{8(\widehat{h}^4 + M)G((h,\infty))} \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2\right] + \left(\frac{1}{G([0,h])} - 1\right)\mathrm{mmse}(G) + \sqrt{8(\widehat{h}^4 + M)G((h,\infty))} \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2\right] + \frac{G((h,\infty))}{G([0,h])}\sqrt{M} + \sqrt{8(\widehat{h}^4 + M)G((h,\infty))} \\
&\le \mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2\right] + \frac{(1+2\sqrt{2})\sqrt{(\widehat{h}^4 + M)G((h,\infty))}}{G([0,h])}. \qquad (3.14)
\end{aligned}
$$

Next we fix $K \ge 1$. Using $\widehat{\theta}_{G_h}(y) \le h, \widehat{\theta}_{\widehat{G}}(y) \le \widehat{h}$ we have

$$
\begin{aligned}
&\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y) - \widehat{\theta}_{G_h}(Y))^2 \mathbf{1}_{\{Y \le K-1\}}\right] \\
&= \sum_{y=0}^{K-1}(y+1)^2 f_{G_h}(y)\left(\frac{f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)} - \frac{f_{G_h}(y+1)}{f_{G_h}(y)}\right)^2 \\
&\overset{(a)}{\le} \sum_{y=0}^{K-1}(y+1)^2 f_{G_h}(y)\left\{3\left(\frac{f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)} - \frac{2f_{\widehat{G}}(y+1)}{f_{G_h}(y) + f_{\widehat{G}}(y)}\right)^2 + 3\left(\frac{f_{G_h}(y+1)}{f_{G_h}(y)} - \frac{2f_{G_h}(y+1)}{f_{G_h}(y) + f_{\widehat{G}}(y)}\right)^2\right. \\
&\quad \left. + 3\left(\frac{2f_{G_h}(y+1) - 2f_{\widehat{G}}(y+1)}{f_{G_h}(y) + f_{\widehat{G}}(y)}\right)^2\right\} \\
&\le 3\sum_{y=0}^{K-1}\left\{\left(\frac{(y+1)f_{\widehat{G}}(y+1)}{f_{\widehat{G}}(y)}\right)^2 \frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)} + \left(\frac{(y+1)f_{G_h}(y+1)}{f_{G_h}(y)}\right)^2 \frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)}\right. \\
&\quad \left. + 4(y+1)^2 \frac{(f_{G_h}(y+1) - f_{\widehat{G}}(y+1))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)}\right\} \\
&= 3(\{\widehat{\theta}_{G_h}(y)\}^2 + \{\widehat{\theta}_{\widehat{G}}(y)\}^2)\sum_{y=0}^{K-1}\frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)} + 12\sum_{y=0}^{K-1}(y+1)^2\frac{(f_{G_h}(y+1) - f_{\widehat{G}}(y+1))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)} \\
&\le 3(h^2 + \widehat{h}^2)\sum_{y=0}^{K-1}\frac{(f_{G_h}(y) - f_{\widehat{G}}(y))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)} + 12\sum_{y=0}^{K-1}(y+1)^2\frac{(f_{G_h}(y+1) - f_{\widehat{G}}(y+1))^2}{f_{G_h}(y) + f_{\widehat{G}}(y)}
\end{aligned}
$$

where (a) followed from $(x+y+z)^2 \le 3(x^2+y^2+z^2)$ for any $x, y, z \in \mathbb{R}$. Using $(\sqrt{f_{G_h}(x)}+\sqrt{f_{\widehat{G}}(x)})^2 \le 2(f_{G_h}(x)+f_{\widehat{G}}(x))$ for $x = y, y+1$ we continue the last display to get

$$\begin{aligned}
&\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\mathbf{1}_{\{Y\le K-1\}}\right] \\
&\le 6(h^2+\widehat{h}^2)\sum_{y=0}^{K-1}(\sqrt{f_{G_h}(y)}-\sqrt{f_{\widehat{G}}(y)})^2 \\
&\quad + 24K\max_{y=0}^{K-1}\frac{(y+1)f_{G_h}(y+1)+(y+1)f_{\widehat{G}}(y+1)}{f_{G_h}(y)+f_{\widehat{G}}(y)}\sum_{y=0}^{K-1}(\sqrt{f_{G_h}(y+1)}-\sqrt{f_{\widehat{G}}(y+1)})^2 \\
&\le \Big(6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\Big)\,\mathsf{H}^2(f_{\widehat{G}}, f_{G_h}).
\end{aligned}$$

Again using $\widehat{\theta}_{G_h}(y) \le h, \widehat{\theta}_{\widehat{G}}(y) \le \widehat{h}$ we bound $\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\mathbf{1}_{\{Y\ge K\}}\right]$ by $(h+\widehat{h})^2\epsilon_K(G_h)$. Combining this with the last display we get

$$\mathbb{E}_{G_h}\left[(\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_{G_h}(Y))^2\right] \le \Big\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\Big\}\,\mathsf{H}^2(f_{\widehat{G}}, f_{G_h}) + (h+\widehat{h})^2\epsilon_K(G_h).$$

In view of above continuing (3.14) we have

$$\begin{aligned}
\mathbb{E}_G\left[\widehat{\theta}_{\widehat{G}}(Y)-\widehat{\theta}_G(Y)\right]^2 \le &\Big\{6(h^2+\widehat{h}^2)+24(h+\widehat{h})K\Big\}\,\mathsf{H}^2(f_{\widehat{G}}, f_{G_h}) \\
&+ (h+\widehat{h})^2\epsilon_K(G_h) + \frac{(1+2\sqrt{2})\sqrt{(M+\widehat{h}^4)G((h,\infty))}}{G([0,h])}.
\end{aligned}$$

This finishes the proof.

Chapter 4

Optimal prediction of Markov chains with and without spectral gap

(This is a joint-work with Yanjun Han and Yihong Wu)

4.1 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 Falahatgar et al. (2016); Hao et al. (2018), in this paper we study the following prediction problem with dependent data modeled as Markov chains.

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

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

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 (4.1) is distinct from the parameter estimation problem such as estimating the transition matrix (Anderson and Goodman, 1957; Bartlett, 1951; Billingsley, 1961; Wolfer and Kontorovich, 2019) or its properties (Csiszár and Shields, 2000; Han et al., 2018a; Hsu et al., 2019; Kamath and Verdú, 2016) 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 $n \to \infty$ (Braess et al., 2002) and $\Theta(\frac{k}{n})$ for $k = O(n)$ (Kamath et al., 2015; Paninski, 2004).[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 number of parameters and inverse proportionally to the desired accuracy.

In the setting of Markov chains, however, the prediction problem is much less understood

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

especially for large state space. Recently the seminal work Falahatgar et al. (2016) showed the surprising result that for stationary Markov chains on two states, the optimal prediction risk satisfies

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

which has a nonparametric rate even when the problem has only two parameters. The follow-up work Hao et al. (2018) 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 Hao et al. (2018) 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 15 (Optimal rates without spectral gap)**.** *There exists a universal constant $C > 0$ such that for all $3 \le k \le \sqrt{n}/C$,*

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

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

Remark 6. 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 $n = 0, 1$)

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

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

$$\widehat{M}^{+1}_{x^n}(j|i) \triangleq \frac{N_{ij}+1}{N_i+k}, \tag{4.1}$$

which is an additively smoothed version of the empirical frequency. Finally, the optimal rate in (15) can be achieved by the following estimator $\widehat{M}$ defined as an average of add-one estimators over different sample sizes:

$$\widehat{M}_{x^n}(x_{n+1}|x_n) \triangleq \frac{1}{n}\sum_{t=1}^{n} \widehat{M}^{+1}_{x_{n-t+1}^n}(x_{n+1}|x_n). \tag{4.2}$$

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., Yang and Barron (1999)). It remains open whether the usual add-one estimator (namely, the last term in (4.2) 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 (4.1) is attained by a hybrid strategy Falahatgar et al. (2016), 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 (4.2) 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 15 shows that the departure from the parametric rate of $\frac{k^2}{n}$, first discovered in Falahatgar et al. (2016); Hao et al. (2018) 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

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

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 15 as well as (4.1) 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 Dean et al. (2019); Simchowitz et al. (2018). 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 15 is proved by constructing chains with spectral gap as small as $O(\frac{1}{n})$ (see Section 4.3). 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 $1 = \lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq -1$. The *absolute spectral gap* of M, defined as

$$\gamma_* \triangleq 1 - \max\{|\lambda_i| : i \neq 1\},$$

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 Levin and Peres (2017a) 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 possible to use more general notions such as pseudo spectral gap to quantify the memory of the process, which is beyond

[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.

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 (4.1) to this subcollection and noticing the stationary distribution here is uniquely determined by M, we define the corresponding minimax risk:

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

Extending the result (4.1) of Falahatgar et al. (2016), the following theorem characterizes the optimal prediction risk for two-state chains with prescribed spectral gaps (the case $\gamma_0 = 0$ correspond to the minimax rate in Falahatgar et al. (2016) over all binary Markov chains):

Theorem 16 (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 16 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 18 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. Similar conclusion has been established previously for the sample complexity of estimating the entropy rate of Markov chains in (Han et al., 2018a, Theorem 1).

Theorem 17. *The add-one estimator in* (4.1) *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 18. *For any fixed* $k \geq 2$, $\mathsf{Risk}_{k,n}(\gamma_0) = O(\frac{1}{n})$ *if and only if* $\gamma_0 = \Omega(1)$.

Finally, we address the optimal prediction risk for higher-order Markov chains. In contrast to the $\Theta\left(\frac{\log\log n}{n}\right)$ rate in the binary first-order Markov chains, for all higher-order

chains the minimax rate for binary state space is significantly higher: $\Theta\left(\frac{\log n}{n}\right)$. In general, we show that for the m^{th}-order chain the minimax prediction error is $\frac{k^{m+1}}{n}\log\frac{n}{k^{m+1}}$ for any given $m \geq 2$, upto constant multiples depending on m. A Cesàro-mean of add-one type estimators, similar to the first-order construction and generalized for the m^{th}-order chains, achieves the upper bound. The lower bound construction is based on a similar mutual information based characterization of the prediction error. The construction of prior on the transition matrices, which are $k^m \times k$ dimensional, to achieve the required bounds on the mutual information, is more involved but uses a certain generalization of the symmetry structure of the first-order transition matrices.

Theorem 19. *There is a constant C_m depending on m such that for any $2 \leq k \leq \sqrt[m+1]{n}/C_m$ and constant $m \geq 2$ the minimax prediction rate for m^{th}-order Markov chains with stationary initialization is $\Theta_m\left(\frac{k^{m+1}}{n}\log\frac{n}{k^{m+1}}\right)$.*

4.1.1 Proof techniques

The proof of Theorem 15 deviates from existing approaches based on concentration inequalities for Markov chains. For instance, the standard program for analyzing the add-one estimator (4.1) 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 Lezaud (1998); Paulin (2015), which have been widely used in recent work on statistical inference with Markov chains (Han et al., 2018a; Hao et al., 2018; Hsu et al., 2019; Kamath and Verdú, 2016; Wolfer and Kontorovich, 2019). 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 Falahatgar et al. (2016) 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 (Whittle, 1955) which takes an unwieldy determinantal form.

Eschewing concentration-based arguments, the crux of our proof of Theorem 15, for both

the upper and lower bound, revolves around the following quantity known as *redundancy*:

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

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 (4.1.1) corresponds to minimax cumulative risk (namely, the total prediction risk when the sample size ranges from 1 to n), while (4.1) is the individual minimax risk at sample size n – see Section 4.2 for a detailed discussion. We prove the following reduction between prediction risk and redundancy:

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

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 (4.2) 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 (4.1.1), and representing the Bayes risk as conditional mutual information and applying the chain rule.

Given the above reduction in (4.1.1), 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 (4.1.1) by the maximum over all trajectories. Following Csiszár and Shields (2004); Davisson et al. (1981), 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 16 relies on the key observation in Falahatgar et al. (2016) 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 (Hao et al., 2018, Theorem 1). The proof of Theorem 17 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 Lezaud (1998); Paulin (2015) falls short of establishing the parametric rate of $O(\frac{k^2}{n})$ (see Remark 9 in Section 4.4.2 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 Billingsley (1961); Han et al. (2018a). 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 $[k]$:

$$\mathbb{P}\left(D(P\|\widehat{P}^{+1}_{n,k}) \geq c \cdot \frac{k}{n} + \frac{\mathsf{polylog}(n) \cdot \sqrt{k}}{n}\right) \leq \frac{1}{\mathsf{poly}(n)}, \tag{4.3}$$

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 (4.3) crucially improves this to $\mathsf{polylog}(n)\cdot\sqrt{k}/n$. Such an improvement has been recently observed by Agrawal (2020); Guo and Richardson (2020); Mardia et al. (2020) 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 (4.3) in which the true distribution P appears as the first argument in the KL divergence.

The nonasymptotic 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, achieving nonasymptotic bounds for redundancy in the higher-order regimes is more

challenging. The geometry of the eigenvalues and the eigenvectors of asymmetric transition matrices are comparatively less understood. As a result, the spectral gaps are not well defined, which makes it challenging to borrow related analysis for the first-order case from Tatwawadi et al. (2018). To bypass this issue, we deduced results based on the pseudo spectral gap of the transition matrix of the first-order chain $\{(X_{t+1}, \ldots, X_{t+m-1})\}_{t=0}^{n-m+1}$. Our analysis entails the construction of a prior that retains $\Theta\left(k^{m+1}\right)$ degrees of freedom for the transition matrix and achieves pseudo spectral gap of constant order, which ensures the required estimation bounds for the transition kernel and attains the required minimax rate.

4.1.2 Related work

While the exact prediction problem studied in this paper has recently been in focus since Falahatgar et al. (2016); Hao et al. (2018), 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 Davisson (1973); Davisson et al. (1981); Rissanen (1984); Ryabko (1988); Shtarkov (1987). Asymptotic characterization of the minimax redundancy for Markov sources, both average and pointwise, were obtained in Atteson (1999); Davisson (1983); Jacquet and Szpankowski (2002), in the regime of fixed alphabet size k and large sample size n. Non-asymptotic characterization was obtained in Davisson (1983) for $n \gg k^2 \log k$ and recently extended to $n \asymp k^2$ in Tatwawadi et al. (2018), 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 Anderson and Goodman (1957); Bartlett (1951); Billingsley (1961); Sinkhorn (1964) and testing the order of Markov chains Csiszár and Shields (2000) have been well studied in the large-sample regime. More recently attention has been shifted towards large state space and nonasymptotics. For example, Wolfer and Kontorovich (2019) 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 Hsu et al. (2019), entropy rate (Han et al., 2018a; Kamath and Verdú, 2016; Obremski and Skorski, 2020), graph statistics based on random walk (Ben-Hamou et al., 2018), as well as identity testing (Cherapanamjeri and Bartlett, 2019; Daskalakis et al., 2018; Fried and Wolfer, 2021; Wolfer and Kontorovich, 2020). 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 15, 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.

4.1.3 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 (Cover and Thomas, 2006; Csiszár and Körner, 1982). The conditional KL divergence is defined as as an average of KL divergence between conditional distributions:

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

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

$$I(A;B) = \min_{Q_B} D(P_{B|A}\|Q_B|P_A), \quad I(A;B|C) = \min_{Q_{B|C}} D(P_{B|A,C}\|Q_{B|C}|P_{AC}).$$

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

4.1.4 Organization

The rest of the paper is organized as follows. In Section 4.2 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 15. Section 4.3 presents the lower bound construction, starting from three states and then extending to k states. Spectral-gap dependent risk bounds in Theorems 16 and 17 are given in Section 4.4. Section 4.5 presents the results and proofs for m^{th}-order Markov chains. Section 4.6 discusses the assumptions and implications of our results and related open problems.

4.2 Two general paradigms

4.2.1 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.

"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

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

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

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

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 (Cover and Thomas, 2006, Chapter 13).

The capacity-redundancy theorem (see Kemperman (1974) for a very general result) provides the following mutual information characterization of (4.2.1):

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

where the supremum is over all distributions (priors) P_θ on Θ. In view of the variational representation (4.1.3), this result can be interpreted as a minimax theorem:

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

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 Rissanen (1984) 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 Davisson (1973) and $\mathsf{Red}_n = \frac{k^m(k-1)}{2} \log n (1 + o(1))$ for m-order Markov models Trofimov (1974), with more refined asymptotics shown in Szpankowski and Weinberger (2012); Xie and Barron (1997). 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 $n \gtrsim k^2$ Tatwawadi et al. (2018).

"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

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

where the conditional KL divergence is defined in (4.1.3). The minimax prediction risk is then defined as

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

While (4.2.1) does not directly correspond to a statistical estimation problem, (4.2.1) 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 (4.1.3):

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

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

$$Q^{\mathsf{Bayes}}_{X_{n+1}|X^n} = P_{X^{n+1}|X^n} = \frac{\int_\Theta P_{X^{n+1}|\theta}\, dP_\theta}{\int_\Theta P_{X^n|\theta}\, dP_\theta}, \tag{4.4}$$

known as the Bayes predictive density Davisson (1973); Liang and Barron (2004). These representations play a crucial role in the lower bound proof of Theorem 15. Under appropriate conditions which hold for Markov models (see Lemma 46 in Appendix 4.7.1), the minimax prediction risk (4.2.1) also admits a dual representation analogous to (4.2.1):

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

which, in view of (4.2.1), 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 20. *For any model $\mathcal{P}$,*

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

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

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

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

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

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

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

Remark 7. Note that the upper bound (20) on redundancy, known as the "estimation-compression inequality" Falahatgar et al. (2016); Kamath et al. (2015), holds without conditions, while the lower bound (20) 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 (20) differ by an unbounded factor of $\Theta(\log\log n)$ for $k=2$ and $\Theta(\log n)$ for fixed $k \ge 3$ – see (4.1) and Theorem 15. On the other hand, Markov chains with at least three states offers a rare instance where (20) is tight, namely, $\mathsf{Risk}_n \asymp \frac{\mathsf{Red}_n}{n}$ (cf. Lemma 21).

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

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

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 $t=1,\dots,n$) yields (20).

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 (20), namely,

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

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

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

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 (4.2.1) of KL divergence. Since the above holds for any $\theta \in \Theta$, the desired (20) follows which implies that $\mathsf{Risk}_{n-1} \leq \frac{\mathsf{Red}_n}{n-m}$. Finally, $\mathsf{Risk}_{n-1} \leq \mathsf{Risk}_n$ 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 $(X_2, \ldots, X_n)$ and $(X_1, \ldots, X_{n-1})$ are equal in law. □

Remark 8. Alternatively, Lemma 20 also follows from the mutual information representation (4.2.1) and (4.2.1). Indeed, by the chain rule for mutual information,

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

taking the supremum over π (the distribution of θ) on both sides yields (4.2.1). For (4.2.1), it suffices to show that $I(\theta; X_t|X^{t-1})$ is decreasing in t: for any $\theta \sim \pi$,

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

and the first term is

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

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 (8), we have $I(\theta; X^n) \ge (n-m) I(\theta; X_n|X^{n-1})$, yielding $\mathsf{Risk}_{n-1} \le \frac{\mathsf{Red}_n}{n-m}$.

4.2.2 Proof of the upper bound part of Theorem 15

Specializing to first-order stationary Markov chains with k states, we denote the redundancy and prediction risk in (4.2.1) and (4.2.1) by $\mathsf{Red}_{k,n}$ and $\mathsf{Risk}_{k,n}$, the latter of which is precisely the quantity previously defined in (4.1). Applying Lemma 20 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:

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

where $\widehat{M}^{+1}$ is the add-one estimator defined in (4.1). 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 21. $\mathsf{Red}(Q) \le k(k-1)\left[\log\left(1 + \frac{n-1}{k(k-1)}\right) + 1\right] + \log k.$

Combined with Lemma 20, Lemma 21 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 (4.2), which is obtained by applying the rule (20) to (4.2.2).

It remains to show Lemma 21. To do so, we in fact bound the pointwise redundancy of the add-one probability assignment (4.2.2) over all (not necessarily stationary) Markov chains on k states. The proof is similar to those of (Csiszár and Shields, 2004, Theorems 6.3 and 6.5), which, in turn, follow the arguments of (Davisson et al., 1981, 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

$$\log \frac{\pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)}{Q(x_1, \cdots, x_n)} \leq k(k-1)\left[\log\left(1 + \frac{n}{k(k-1)}\right) + 1\right] + \log k, \tag{4.5}$$

where we abbreviate the add-one estimator $M_{x^t}(x_{t+1}|x_t)$ defined in (4.1) as $M(x_{t+1}|x_t)$.

To establish (4.5), note that $Q(x_1, \cdots, x_n)$ could be equivalently expressed using the empirical counts N_i and N_{ij} in (6) 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

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

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

$$\frac{\pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)}{Q(x_1, \cdots, x_n)} \leq k \cdot \prod_{i=1}^{k} \frac{k \cdot (k+1) \cdot \cdots \cdot (N_i + k - 1)}{N_i^{N_i}} \prod_{j=1}^{k} \frac{N_{ij}^{N_{ij}}}{N_{ij}!}. \tag{4.6}$$

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

$$\prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i} \leq \frac{\prod_{i=1}^{k} n_i!}{n!}, \tag{4.7}$$

with the understanding that $(\frac{0}{n})^0 = 0! = 1$. Applying this claim to (4.6) gives

$$\begin{aligned}
\log \frac{\pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)}{Q(x_1,\cdots,x_n)} &\le \log k + \sum_{i=1}^{k} \log \frac{k\cdot(k+1)\cdot\cdots\cdot(N_i+k-1)}{N_i!} \\
&= \log k + \sum_{i=1}^{k}\sum_{\ell=1}^{N_i} \log\left(1+\frac{k-1}{\ell}\right) \\
&\le \log k + \sum_{i=1}^{k}\int_0^{N_i} \log\left(1+\frac{k-1}{x}\right) dx \\
&= \log k + \sum_{i=1}^{k}\left((k-1)\log\left(1+\frac{N_i}{k-1}\right) + N_i \log\left(1+\frac{k-1}{N_i}\right)\right) \\
&\overset{\text{(a)}}{\le} k(k-1)\log\left(1+\frac{n-1}{k(k-1)}\right) + k(k-1) + \log k,
\end{aligned}$$

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 (4.7), 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

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

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 $(\frac{n_1}{n},\ldots,\frac{n_k}{n})$. □

4.3 Optimal rates without spectral gap

In this section, we prove the lower bound part of Theorem 15, which shows the optimality of the average version of the add-one estimator (20). We first describe the lower bound construction for three-state chains, which is subsequently extended to k states.

4.3.1 Warmup: an $\Omega(\frac{\log n}{n})$ lower bound for three-state chains

Theorem 22. $\mathsf{Risk}_{3,n} = \Omega\left(\frac{\log n}{n}\right)$.

To show Theorem 22, consider the following one-parameter family of transition matrices:

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

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. 4.1.

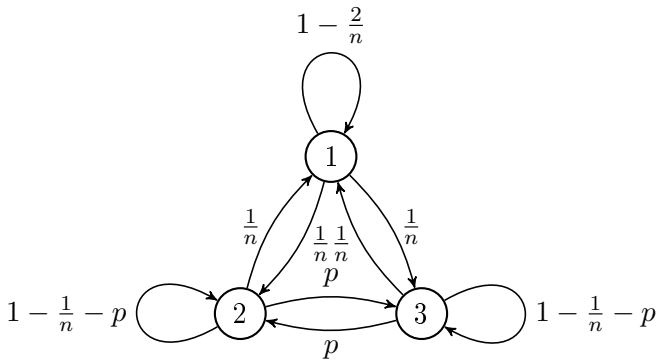

Figure 4.1: Lower bound construction for three-state chains.

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

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

Intuitively speaking, the construction in Fig. 4.1 "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 22, which actually stems from the harmonic sum in $\mathbb{E}[\frac{1}{\text{Uniform}([n])}]$. We will fully explore this embedding idea in Section 4.3.2 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 23. *Let $q \sim \text{Uniform}(0,1)$. Conditioned on q, let $N \sim \text{Binom}(m,q)$. 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 22. 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 (4.3.1). Define

$$\begin{gathered}\mathcal{X}_t = \{x^n : x_1 = \ldots = x_t = 1, x_i \neq 1, i = t+1, \ldots, n\}, \quad t = 1, \ldots, n-1, \\ \mathcal{X} = \cup_{t=1}^{n-1} \mathcal{X}_t.\end{gathered} \tag{4.8}$$

Let $\mu(x^n|p) = \mathbb{P}[X = x^n]$. Then

$$\mu(x^n|p) = \frac{1}{3}\left(1 - \frac{2}{n}\right)^{t-1} \frac{2}{n} p^{N(x^n)} \left(1 - \frac{1}{n} - p\right)^{n-t-1-N(x^n)}, \quad x^n \in \mathcal{X}_t,$$

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

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

and hence

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

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 (4.3.1) we have

$$\begin{aligned}\widehat{p}(x^n) &= \frac{\mathbb{E}\left[p^{N(x^n)+1}\left(1-\frac{1}{n}-p\right)^{n-t-1-N(x^n)}\right]}{\mathbb{E}\left[p^{N(x^n)}\left(1-\frac{1}{n}-p\right)^{n-t-1-N(x^n)}\right]}, \quad p\sim\text{Uniform}\left(0,\frac{n-1}{n}\right) \\ &= \frac{n-1}{n}\frac{\mathbb{E}\left[U^{N(x^n)+1}\left(1-U\right)^{n-t-1-N(x^n)}\right]}{\mathbb{E}\left[U^{N(x^n)}\left(1-U\right)^{n-t-1-N(x^n)}\right]}, \quad U\sim\text{Uniform}(0,1) \\ &= \frac{n-1}{n}\frac{N(x^n)+1}{n-t+1},\end{aligned}$$

where the last step follows from Lemma 23. From (4.3.1), 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 23 (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 (4.3.1) and (4.3.1) we get

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

Thus

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

Finally, we relate (4.3.1) 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

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

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

$$\begin{aligned}&\sum_{i=1}^{3} \mathbb{E}[D(M(\cdot|i)\|\widehat{M}(\cdot|i))\mathbf{1}_{\{X_n=i\}}] \\ &\geq \frac{1}{2}\mathbb{E}[(p-\widehat{p}_2)^2 \mathbf{1}_{\{X_n=2\}} + (p-\widehat{p}_3)^2 \mathbf{1}_{\{X_n=3\}}] \\ &\geq \frac{1}{2}\sum_{x \in \mathcal{X}} \mu(x^n)\left(\mathbb{E}[(p-\widehat{p}_2)^2 | X = x^n]\mathbf{1}_{\{x_n=2\}} + \mathbb{E}[(p-\widehat{p}_3)^2 | X = x^n]\mathbf{1}_{\{x_n=3\}}\right) \\ &\geq \frac{1}{2}\sum_{x^n \in \mathcal{X}} \mu(x^n)\mathbb{E}[(p-\widehat{p}(x^n))^2 | X = x^n](\mathbf{1}_{\{x_n=2\}} + \mathbf{1}_{\{x_n=3\}}) \\ &= \frac{1}{2}\sum_{x^n \in \mathcal{X}} \mu(x^n)\mathbb{E}[(p-\widehat{p}(x^n))^2 | X = x^n] \\ &= \frac{1}{2}\mathbb{E}[(p-\widehat{p}(X))^2 \mathbf{1}_{\{X \in \mathcal{X}\}}] \overset{(4.3.1)}{=} \Theta\left(\frac{\log n}{n}\right).\end{aligned}$$

□

4.3.2 k-state chains

The lower bound construction for 3-state chains in Section 4.3.1 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 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. 4.2 in Section 4.3.2 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

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

where the last equality follows from the representation (4.2.1) of Bayes prediction risk as conditional mutual information. Lower bounding the minimax risk by the Bayes risk, we

have

$$\begin{aligned}
\mathsf{Risk}_{k,n} &\geq \inf_{\widehat{M}} \mathbb{E}_T \left[\mathbb{E}[D(M(\cdot|X_n) \| \widehat{M}(\cdot|X_n))] \right] \\
&\geq \inf_{\widehat{M}} \sum_{t=1}^{n-1} \mathbb{E}_M \left[\mathbb{E}[D(M(\cdot|X_n) \| \widehat{M}(\cdot|X_n))|\mathcal{X}_t] \cdot \mathbb{P}(\mathcal{X}_t) \right] \\
&\geq \frac{c}{n} \cdot \sum_{t=1}^{n-1} \inf_{\widehat{M}} \mathbb{E}_M \left[\mathbb{E}[D(M(\cdot|X_n) \| \widehat{M}(\cdot|X_n))|\mathcal{X}_t] \right] \\
&\approx \frac{c}{n} \cdot \sum_{t=1}^{n-1} I(T; Y_{n-t+1}|Y^{n-t}) = \frac{c}{n} \cdot (I(T;Y^n) - I(T;Y_1)). \qquad (4.9)
\end{aligned}$$

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 (4.2.1), the above inequality (4.9) 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 (4.9) yields the desired lower bound $\mathsf{Risk}_{k,n} = \Omega(\frac{k^2}{n} \log \frac{n}{k^2})$ in Theorem 15.

Next we rigorously carry out the lower bound proof sketched above: In Section 4.3.2, we explicitly construct the k-state chain which satisfies the desired properties in Section 4.3.2. In Section 4.3.2, we make the steps in (4.9) precise and bound the Bayes risk from below by an appropriate mutual information. In Section 4.3.2, 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 15, with the added bonus that the construction is restricted to irreducible and reversible chains.

Construction of the k-state chain

We construct a k-state chain with the following transition probability matrix:

$$M = \begin{bmatrix} 1-\frac{1}{n} & \frac{1}{n(k-1)} & \frac{1}{n(k-1)} & \cdots & \frac{1}{n(k-1)} \\ 1/n & & & & \\ 1/n & & \left(1-\frac{1}{n}\right) T & & \\ \vdots & & & & \\ 1/n & & & & \end{bmatrix}, \tag{4.10}$$

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 4.2. One can also verify that the spectral gap of M is $\Theta(\frac{1}{n})$.

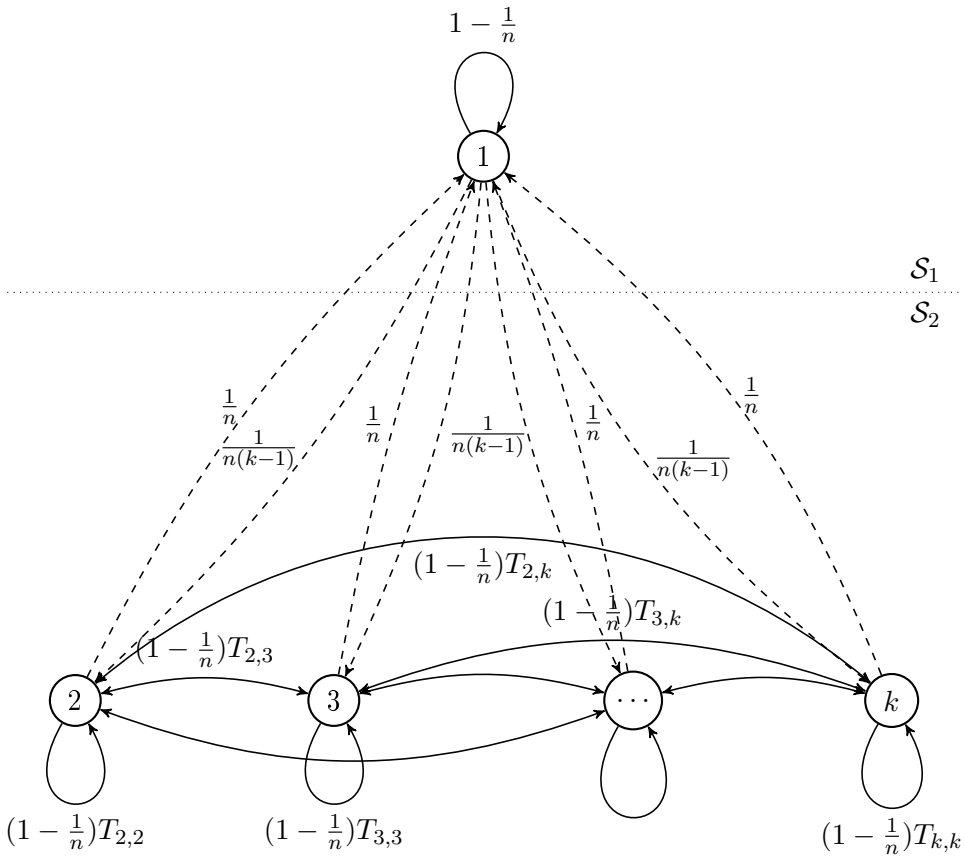

Figure 4.2: Lower bound construction for k-state chains. Solid arrows represent transitions within $\mathcal{S}_1$ and $\mathcal{S}_2$, and dashed arrows represent transitions between $\mathcal{S}_1$ and $\mathcal{S}_2$. The double-headed arrows denote transitions in both directions with equal probabilities.

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

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

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,

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

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 (4.10) 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 24. *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*

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

Lemma 24 is the formal statement of the inequality (4.9) presented in the proof sketch. Maximizing the lower bound over the prior on T and in view of the mutual information representation (4.2.1), we obtain the following corollary.

Corollary 25. *Let* $\mathsf{Risk}_{k,n}^{\mathsf{sym}}$ *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 4.3.2 to prove Lemma 24.

Proof of Lemma 24. 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 (4.10), 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 (4.4) of the Bayes estimator, we have, for each $j \in \mathcal{S}_2$,

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

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 (4.10),

$\widehat{M}_n(1|x^n) = \frac{1}{n}$ is deterministic. In all, we have

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

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

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

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

$$\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X^n))]\right] \geq \frac{1}{2en}\sum_{t=1}^{n-1}\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X^n))|X^n \in \mathcal{X}_t]\right] \quad (4.11)$$

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

$$\begin{aligned}\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_n)\|\widehat{M}(\cdot|X^n))|X^n \in \mathcal{X}_t]\right] &\overset{(a)}{=} \left(1 - \frac{1}{n}\right)\cdot\mathbb{E}_T\left[\mathbb{E}[D(T(\cdot|Y_{n-t})\|\widehat{T}(\cdot|Y^{n-t}))]\right] \\ &\overset{(b)}{=} \left(1 - \frac{1}{n}\right)\cdot I(T; Y_{n-t+1}|Y^{n-t}), \end{aligned} \quad (4.12)$$

where (a) follows from (4.3.2), (4.3.2), 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 (4.2.1) of the Bayes prediction risk. Finally, the lemma follows from (4.11), (4.12), and the chain rule

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

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

Prior construction and lower bounding the mutual information

In view of Lemma 24, 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 Davisson (1983); Davisson et al. (1981); Rissanen (1984) between estimation error and mutual information (see also (Csiszár and Shields, 2004, 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 26. *In the setting of Lemma 24, 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*

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

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

We show how Lemma 26 leads to the desired lower bound on the mutual information $I(T; Y^n)$. Since $k \ge 3$, we may assume that $k - 1 = 2k_0$ is an even integer. Consider the following prior distribution π on T: let $u = (u_{i,j})_{i,j\in[k_0],i\le j}$ 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

$$T_{2i-1,2j-1} = T_{2i,2j} = u_{i,j}, \quad T_{2i-1,2j} = T_{2i,2j-1} = \frac{1}{k_0} - u_{i,j}, \quad \forall i, j \in [k].$$

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 26 is fulfilled, so there exist estimators $\widehat{T}(Y^n)$ and $\widehat{u}(Y^n)$ such that

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

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. Cover and Thomas (2006)). Then

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

Then

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

where (a) is because u and T are in one-to-one correspondence by (4.3.2); (b) follows from the data processing inequality; (c) is because $h(\cdot)$ is translation invariant and concave; (d) follows from the maximum entropy principle Cover and Thomas (2006): $h(u-\widehat{u}(Y^n)) \le \frac{k_0(k_0+1)}{4}\log\left(\frac{2\pi e}{k_0(k_0+1)/2}\cdot\mathbb{E}[\|\widehat{u}(Y^n)-u\|_2^2]\right)$, which in turn is bounded by (4.5.2). Plugging this lower bound into Lemma 24 completes the lower bound proof of Theorem 15.

Proof of Lemma 26. 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 (Tatwawadi et al., 2018, Lemma 7, Lemma 8) using the concentration inequality of Paulin (2015):

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

where $\gamma_*(T)$ is the absolute spectral gap of T defined in (4.1). 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. □

4.4 Spectral gap-dependent risk bounds

4.4.1 Two states

To show Theorem 16, let us prove a refined version. In addition to the absolute spectral gap defined in (4.1), define the spectral gap

$$\gamma \triangleq 1 - \lambda_2$$

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 (4.1), define $\mathsf{Risk}'_{k,n}(\gamma_0)$ as the minimax prediction risk restricted to $M \in \mathcal{M}'_k(\gamma_0)$ Since $\gamma \geq \gamma^*$, we have $\mathcal{M}_k(\gamma_0) \subseteq \mathcal{M}'_k(\gamma_0)$ and hence $\mathsf{Risk}'_{k,n}(\gamma_0) \geq \mathsf{Risk}_{k,n}(\gamma_0)$. Nevertheless, the next result shows that for $k = 2$ they have the same rate:

Theorem 27 (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

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

Indeed, for any $\gamma_0 \leq n^{-0.9}$, the upper bound $\mathcal{O}(\log\log n/n)$ proven in Falahatgar et al. (2016), 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 (4.4.1). Following Falahatgar et al. (2016), con-

sider trajectories with a single transition, namely, $\{2^{n-\ell}1^\ell, 1^{n-\ell}2^\ell : 1 \le \ell \le n-1\}$, 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 (4.1), 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 (6); for step sequences of the form $2^{n-\ell}1^\ell$, we estimate by

$$\widehat{M}_\ell(2|1) = 1/(\ell \log(1/\gamma_0)), \quad \widehat{M}_\ell(1|1) = 1 - \widehat{M}_\ell(2|1). \tag{4.13}$$

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

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

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 (Falahatgar et al., 2016, Lemma 7,8) we have that the total risk of other non-step sequences is bounded from above by $\mathcal{O}\left(\frac{1}{n}\right)$ and hence it is enough to analyze the risk for step sequences, and further by symmetry, those in $\{2^{n-\ell}1^\ell : 1 \le \ell \le n-1\}$. The desired upper bound (4.4.1) then follows from Lemma 28 next.

Lemma 28. *For any $n^{-0.9} \le \gamma_0 \le e^{-e^5}$, $\widehat{M}_\ell(\cdot|1)$ in (4.13) 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 $x = \frac{1}{\ell \log(1/\gamma_0)}$,

$$\begin{aligned} D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1)) &= M(1|1)\log\left(\frac{M(1|1)}{1-\frac{1}{\ell\log(1/\gamma_0)}}\right) + M(2|1)\log\left(M(2|1)\ell\log(1/\gamma_0)\right) \\ &\lesssim \frac{1}{\ell\log(1/\gamma_0)} + M(2|1)\log(M(2|1)\ell) + M(2|1)\log\log(1/\gamma_0) \\ &\le \frac{1}{\ell\log(1/\gamma_0)} + M(2|1)\log_+(M(2|1)\ell) + M(2|1)\log\log(1/\gamma_0), \end{aligned} \tag{4.14}$$

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 $x, y \in [0,1]$,

$$\begin{aligned} \sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}y(1-y)^{\ell-1} &\le \frac{1}{n-1}\left(\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}\right)\left(\sum_{\ell=1}^{n-1} y(1-y)^{\ell-1}\right) \\ &\le \frac{1}{n-1}, \end{aligned} \tag{4.15}$$

$$\begin{aligned} \sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}y(1-y)^{\ell-1}\log_+(\ell y) &\le \frac{1}{n-1}\left(\sum_{\ell=1}^{n-1} x(1-x)^{n-\ell-1}\right)\left(\sum_{\ell=1}^{n-1} y(1-y)^{\ell-1}\log_+(\ell y)\right) \\ &\le \frac{1}{n-1}\sum_{\ell=1}^{n-1} y(1-y)^{\ell-1}(1+\ell y) \le \frac{2}{n-1}, \end{aligned} \tag{4.16}$$

where in (4.16) 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,

$$
\begin{aligned}
&\sum_{\ell=1}^{n-1} \mathbb{E}\left[\mathbf{1}_{\{X^n=2^{n-\ell}1^\ell\}} D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1))\right] \\
&\le \sum_{\ell=1}^{n-1} M(2|2)^{n-\ell-1} M(1|2) M(1|1)^{\ell-1} D(M(\cdot|1)\|\widehat{M}_\ell(\cdot|1)) \\
&\overset{\text{(a)}}{\lesssim} \sum_{\ell=1}^{n-1} M(2|2)^{n-\ell-1} M(1|2) M(1|1)^{\ell-1} \left(\frac{1}{\ell \log(1/\gamma_0)} + M(2|1)\log_+(M(2|1)\ell) + M(2|1)\log\log(1/\gamma_0)\right) \\
&\overset{\text{(b)}}{\le} \sum_{\ell=1}^{n-1} \frac{M(2|2)^{n-\ell-1} M(1|2) M(1|1)^{\ell-1}}{\ell \log(1/\gamma_0)} + \frac{2+\log\log(1/\gamma_0)}{n-1}, \qquad (4.17)
\end{aligned}
$$

where (a) is due to (4.14), (b) follows from (4.15) and (4.16) 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

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

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 (4.17) 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 i.i.d. setting (see, e.g., Braess et al. (2002)), 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

$$\alpha = \log(1/\gamma_0), \quad \beta = \left\lceil \frac{\alpha}{5 \log \alpha} \right\rceil, \tag{4.18}$$

and consider the prior distribution

$$\mathscr{M} = \text{Uniform}(\mathcal{M}), \quad \mathcal{M} = \left\{ M : M(1|2) = \frac{1}{n}, M(2|1) = \frac{1}{\alpha^m} : m \in \mathbb{N} \cap (\beta, 5\beta) \right\} \tag{4.19}$$

Then the lower bound part of Theorem 16 follows from the next lemma.

Lemma 29. *Assume that* $n^{-0.9} \leq \gamma_0 \leq e^{-e^5}$. *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 $\left\{2^{n-\ell}1^\ell : 1 \leq \ell \leq n-1\right\}$. Our argument closely follows that of (Hao et al., 2018, Theorem 1). Since $\gamma_0 \geq n^{-1}$, for each $M \in \mathcal{M}$, the corresponding stationary distribution π satisfies

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

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

$$\mathbb{P}\left[X^n = 2^{n-\ell}1^\ell\right] = \pi_2 \left(1 - \frac{1}{n}\right)^{n-\ell-1} \frac{1}{n} M(1|1)^{\ell-1} \geq \frac{1}{2en} M(1|1)^{\ell-1}.$$

Then

$$\begin{aligned}\mathsf{Risk}(\mathscr{M}) &\geq \mathbb{E}_{M\sim\mathscr{M}}\left[\sum_{\ell=1}^{n-1}\mathbb{E}\left[\mathbf{1}_{\left\{X^n=2^{n-\ell}1^\ell\right\}}D(M(\cdot|1)\|\widehat{M}_\ell^{\mathsf{B}}(\cdot|1))\right]\right]\\ &\geq \mathbb{E}_{M\sim\mathscr{M}}\left[\sum_{\ell=1}^{n-1}\frac{M(1|1)^{\ell-1}}{2en}D(M(\cdot|1)\|\widehat{M}_\ell^{\mathsf{B}}(\cdot|1))\right]\\ &= \frac{1}{2en}\sum_{\ell=1}^{n-1}\mathbb{E}_{M\sim\mathscr{M}}\left[M(1|1)^{\ell-1}D(M(\cdot|1)\|\widehat{M}_\ell^{\mathsf{B}}(\cdot|1))\right]. \qquad (4.20)\end{aligned}$$

Recalling the general form of the Bayes estimator in (4.4) and in view of (4.4.1), we get

$$\widehat{M}_\ell^{\mathsf{B}}(2|1)=\frac{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]},\quad \widehat{M}_\ell^{\mathsf{B}}(1|1)=1-\widehat{M}_\ell^{\mathsf{B}}(2|1). \qquad (4.21)$$

Plugging (4.21) into (4.20), and using

$$D((x,1-x)\|(y,1-y))=x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}\geq x\max\left\{0,\log\frac{x}{y}-1\right\},$$

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

$$\begin{aligned}&\mathsf{Risk}(\mathscr{M})\\ \geq&\frac{1}{2en}\sum_{\ell=1}^{n-1}\mathbb{E}_{M\sim\mathscr{M}}\left[M(1|1)^{\ell-1}M(2|1)\max\left\{0,\log\left(\frac{M(2|1)\cdot\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]}\right)-1\right\}\right] \qquad (4.22)\end{aligned}$$

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

We further lower bound (4.22) 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 $\gamma_0\leq e^{-e^5}$, our choice of α ensures that the intervals $\{[\ell_1(m),\ell_2(m)]\}_{\beta\leq m\leq 3\beta}$ 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

$$\frac{\alpha^{-m}\cdot\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}]}{\mathbb{E}_{M\sim\mathscr{M}}[M(1|1)^{\ell-1}M(2|1)]}\gtrsim\frac{\log(1/\gamma_0)}{\log\log(1/\gamma_0)}. \tag{4.23}$$

We first complete the proof of the Bayes risk bound assuming (4.23). Using (4.22) and (4.23), we have

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

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

Next we prove the claim (4.23). Expanding the expectation in (4.19), we write the LHS of (4.23) 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

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

We bound each of the terms individually. Clearly, $X_\ell\in(0,1)$ and $A_\ell\geq0$. Thus it suffices to show that $B_\ell\gtrsim\beta$ and $C_\ell,D_\ell\lesssim1$, for $m\in[\beta,3\beta]$ and $\ell_1(m)\leq\ell\leq\ell_2(m)$. Indeed,

- For $j\geq m+1$, we have

$$\left(1-\alpha^{-j}\right)^\ell\geq\left(1-\alpha^{-j}\right)^{\ell_2(m)}\overset{\text{(a)}}{\geq}(1/4)^{\frac{\ell_2(m)}{\alpha^j}}\geq(1/4)^{\frac{\log\alpha}{\alpha}}\geq1/4,$$

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

$$\left(1-\alpha^{-j}\right)^{\ell} \leq \left(1-\alpha^{-j}\right)^{\ell_1(m)} \overset{\text{(b)}}{\leq} e^{-\frac{\alpha^{m-j}}{\log \alpha}} = \gamma_0^{\frac{\alpha^{m-j-1}}{\log \alpha}},$$

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 (4.18);

- $D_\ell \leq \sum_{j=m+1}^{5\beta} \alpha^{m-j} \leq 1$, since $\alpha = \log \frac{1}{\gamma_0} \geq e^5$.

Combining the above bounds completes the proof of (4.23). □

4.4.2 k states

Proof of Theorem 17 (i)

Notice that the prediction problem consists of k sub-problems of estimating the individual rows of M, so it suffices show the contribution from each of them is $O\left(\frac{k}{n}\right)$. In particular, assuming the chain terminates in state 1 we bound the risk of estimating the first row by the add-one estimator $\widehat{M}^{+1}(j|1) = \frac{N_{1j}+1}{N_1+k}$. Under the absolute spectral gap condition of $\gamma_* \geq \gamma_0$, we show

$$\mathbb{E}\left[\mathbf{1}_{\{X_n=1\}} D\left(M(\cdot|1) \| \widehat{M}^{+1}(\cdot|1)\right)\right] \lesssim \frac{k}{n}\left(1+\sqrt{\frac{\log k}{k\gamma_0^4}}\right). \tag{4.24}$$

By symmetry, 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 basic steps of our analysis are as follows:

- When N_1 is substantially smaller than its mean, we can bound the risk using the worst-case risk bound for add-one estimators and the probability of this rare event.

- Otherwise, we decompose the prediction risk as

$$D(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)) = \sum_{j=1}^{k}\left[M(j|1)\log\left(\frac{M(j|1)(N_1+k)}{N_{1j}+1}\right) - M(j|1) + \frac{N_{1j}+1}{N_1+k}\right].$$

  We then analyze each term depending on whether N_{1j} is typical or not. Unless N_{1j} is atypically small, the add-one estimator works well whose risk can be bounded quadratically.

To analyze the concentration of the empirical counts we use the following moment bounds. The proofs are deferred to Appendix 4.7.2.

Lemma 30. *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 i.i.d. sampling. For i.i.d. models Kamath et al. (2015) showed that the add-one estimator achieves $\mathcal{O}\left(\frac{k}{n}\right)$ risk bound which we aim here too. In addition, dependency of the above moments on γ_* gives rise to sufficient conditions that guarantees parametric rate. The technical details are given below.

We decompose the left hand side in (4.24) based on N_1 as

$$\mathbb{E}\left[\mathbf{1}_{\{X_n=1\}} D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] = \mathbb{E}\left[\mathbf{1}_{\{A^{\leq}\}} D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] + \mathbb{E}\left[\mathbf{1}_{\{A^{>}\}} D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right]$$

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

$$A^{\leq} \triangleq \{X_n = 1, N_1 \leq (n-1)\pi_1/2\}, \quad A^{>} \triangleq \{X_n = 1, N_1 > (n-1)\pi_1/2\}.$$

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{A^{\le}\}} D\left(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1)\right)\right] \\
&\le \log(n\pi_1+k)\,\mathbb{P}\left[X_n=1, N_1 \le (n-1)\pi_1/2\right] \\
&\overset{(a)}{\lesssim} \mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\pi_1\log(n\pi_1+k) + \mathbf{1}_{\{n\pi_1\gamma_*>10\}}\pi_1\log(n\pi_1+k)\,\frac{\mathbb{E}\left[(N_1-(n-1)\pi_1)^4\,|X_n=1\right]}{n^4\pi_1^4} \qquad (4.25) \\
&\overset{(b)}{\le} \mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\frac{10}{n\gamma_*}\log\left(\frac{10}{\gamma_*}+k\right) + \mathbf{1}_{\{n\pi_1\gamma_*>10\}}\log(n\pi_1+k)\left(\frac{1}{n^2\pi_1\gamma_*^2}+\frac{1}{n^4\pi_1^3\gamma_*^4}\right) \\
&\overset{(c)}{\lesssim} \frac{1}{n}\left\{\mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\frac{\log(1/\gamma_*)+\log k}{\gamma_*} + \mathbf{1}_{\{n\pi_1\gamma_*>10\}}(n\pi_1+\log k)\left(\frac{1}{n\pi_1\gamma_*^2}+\frac{1}{n^3\pi_1^3\gamma_*^4}\right)\right\} \\
&\lesssim \frac{1}{n}\left\{\mathbf{1}_{\{n\pi_1\gamma_*\le 10\}}\left(\frac{1}{\gamma_*^2}+\frac{\log k}{\gamma_*}\right) + \mathbf{1}_{\{n\pi_1\gamma_*>10\}}\left(\frac{1}{\gamma_*^2}+\frac{\log k}{\gamma_*}\right)\right\} \lesssim \frac{1}{n\gamma_0^2}+\frac{\log k}{n\gamma_0}. \qquad (4.26)
\end{aligned}$$

where we got (a) from Markov inequality, (b) from Lemma 30(iii) and (c) using $x+y \le xy, x, y \ge 2$.

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

Case (a) $n\pi_1 \le r$ **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

$$\Delta_i \le \frac{(M(i|1)N_1 - N_{1i} + M(i|1)k - 1)^2}{(N_1+k)(N_{1i}+1)}. \qquad (4.27)$$

This implies

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

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

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

Hence

$$\mathbb{E}\left[\mathbf{1}_{\{A^{>}\}} D(M(\cdot|1)\|\widehat{M}^{+1}(\cdot|1))\right] = \sum_{i=1}^{k} \mathbb{E}\left[\mathbf{1}_{\{A^{>}\}}\Delta_i\right] \lesssim \frac{rk}{n} + \frac{1}{\gamma_0^2}. \tag{4.28}$$

Case(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.

Bound on $\mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}}\Delta_i\right]$ Using $\widehat{M}^{+1}(i|1) \geq \frac{1}{N_1+k}$ and $N_{1i} < N_1 M(i|1)/2$ in $B^{\leq}$ we get

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}}\Delta_i\right] &= \mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}} M(i|1)\log\left(\frac{M(i|1)(N_1+k)}{N_{1i}+1}\right)\right] + \mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}}\left(\frac{N_{1i}+1}{N_1+k} - M(i|1)\right)\right] \\
&\leq \mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right] + \mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}}\left(\frac{N_{1i}}{N_1} - M(i|1)\right)\right] + \mathbb{E}\left[\frac{\mathbf{1}_{\{B^{\leq}\}}}{N_1}\right] \\
&\lesssim \mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right] + \frac{1}{n} \qquad (4.29)
\end{aligned}$$

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

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

Applying this identity with $t_0 = \lceil (n-1)\pi_1/2 \rceil$, we can bound the expectation term in (4.29) as

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right] \\
&= M(i|1)\log\left(M(i|1)(t_0+k)\right)\mathbb{P}\left[N_1 \geq t_0, N_{1i} \leq \frac{n\pi_1 M(i|1)}{4}, X_n = 1\right] \\
&\quad + M(i|1)\sum_{t=t_0+1}^{n-1} \log\left(1 + \frac{1}{t-1+k}\right)\mathbb{P}\left[N_1 \geq t+1, N_{1i} \leq \frac{n\pi_1 M(i|1)}{4}, X_n = 1\right] \\
&\leq \pi_1 M(i|1)\log\left(M(i|1)(t_0+k)\right)\mathbb{P}\left[M(i|1)N_1 - N_{1i} \geq \frac{M(i|1)t_0}{4}\middle| X_n = 1\right] \\
&\quad + \frac{M(i|1)}{n}\sum_{t=t_0+1}^{n-1} \mathbb{P}\left[M(i|1)N_1 - N_{1i} \geq \frac{M(i|1)t}{4}\middle| X_n = 1\right] \qquad (4.30)
\end{aligned}$$

where last inequality uses $\log\left(1 + \frac{1}{t-1+k}\right) \leq \frac{1}{t} \lesssim \frac{1}{n\pi_1}$ for all $t \geq t_0$. Using Markov inequality $\mathbb{P}[Z > c] \leq c^{-4}\mathbb{E}\left[Z^4\right]$ for $c > 0$, Lemma 30(ii) and $x + x^4 \leq 2(1+x^4)$ with $x = \sqrt{M(i|1)}/\gamma_*$

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

In view of above continuing (4.30) we get

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{B^{\leq}\}} M(i|1)\log\left(M(i|1)(N_1+k)\right)\right] \\
&\lesssim \left((n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}\right)\left(\frac{\pi_1 M(i|1)\log(M(i|1)(n\pi_1+k))}{(n\pi_1 M(i|1))^4} + \frac{1}{n(M(i|1))^3}\sum_{t=t_0+1}^{n}\frac{1}{t^4}\right) \\
&\lesssim \left(\frac{(n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}}{n}\right)\left(\frac{\log(n\pi_1 M(i|1)+kM(i|1))}{(n\pi_1 M(i|1))^3} + \frac{1}{(n\pi_1 M(i|1))^3}\right) \\
&\lesssim \frac{1}{n}\left((n\pi_1 M(i|1))^2 + \frac{M(i|1)^2}{\gamma_*^4}\right)\frac{\log(n\pi_1 M(i|1)+kM(i|1))}{(n\pi_1 M(i|1))^3} \\
&\lesssim \frac{1}{n}\left(\frac{\log(n\pi_1 M(i|1)+kM(i|1))}{n\pi_1 M(i|1)} + \frac{M(i|1)\log(n\pi_1 M(i|1)+k)}{n\pi_1\gamma_*^4(n\pi_1 M(i|1))^2}\right) \\
&\overset{(a)}{\lesssim} \frac{1}{n}\left(\frac{n\pi_1 M(i|1)+kM(i|1)}{n\pi_1 M(i|1)} + \frac{M(i|1)\log(n\pi_1 M(i|1))}{n\pi_1\gamma_*^4(n\pi_1 M(i|1))^2} + \frac{M(i|1)\log k}{n\pi_1\gamma_*^4(n\pi_1 M(i|1))^2}\right) \\
&\overset{(b)}{\lesssim} \frac{1}{n}\left(1 + kM(i|1) + \frac{M(i|1)\log k}{r\gamma_0^4}\right)
\end{aligned}
$$

where (a) followed using $x+y \leq xy$ for $x, y \geq 2$ and (b) followed as $n\pi_1 \geq r, n\pi_1 M(i|1) \geq 10$ and $\log(n\pi_1 M(i|1)) \leq n\pi_1 M(i|1)$. In view of (4.29) this implies

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

Bound on $\mathbb{E}\left[\mathbf{1}_{\{B^{>}\}}\Delta_i\right]$

Using the inequality (4.27)

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

where (a) follows using properties of the set $B^{>}$ along with $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$.

Using Lemma 30(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 we get for $n\pi_1 > r, n\pi_1 M(i|1) > 10$

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

Combining this with (4.28) 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 (4.26) this implies the required bound.

Remark 9. We explain the subtlety of the concentration bound in Lemma 30 based on fourth moment and why existing Chernoff bound or Chebyshev inequality falls short. For example, the risk bound in (4.26) 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 (Lezaud, 1998, Theorem 1.1) or (Paulin, 2015, Proposition 3.10 and Theorem 3.3))

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

in contrast, the fourth moment bound in (4.25) yields $\mathbb{P}\left[N_1 \leq (n-1)\pi_1/2 | X_1 = 1\right] = O(\frac{1}{(n\pi_1\gamma_*)^2})$. Although the exponential tail in (9) 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 (9) leads to $\mathbb{P}\left[N_1 \leq (n-1)\pi_1/2, X_n = 1\right] = O(\frac{1}{\sqrt{n}})$ as opposed to the desired $O(\frac{1}{n})$.)

In the same context it is also insufficient to use 2nd moment based bound (Chebyshev), which leads to $\mathbb{P}\left[N_1 \leq (n-1)\pi_1/2 | X_1 = 1\right] = O(\frac{1}{n\pi_1\gamma_*})$. This bound is too loose, which,

upon substitution into (4.25), results in an extra $\log n$ factor in the final risk bound when π_1 and γ_* are large.

Proof of Theorem 17 (ii)

Let $k \geq (\log n)^6$ and $\gamma_0 \geq \frac{(\log(n+k))^2}{k}$. We prove a stronger result using spectral gap as opposed to the absolute spectral gap. Fix M such that $\gamma \geq \gamma_0$. Denote its stationary distribution by π. For absolute constants $\tau > 0$ to be chosen later and c_0 as in Lemma 31 below, define

$$\epsilon(m) = \frac{2k}{m} + \frac{c_0(\log n)^3\sqrt{k}}{m}, \quad c_n = 100\tau^2\frac{\log n}{n\gamma},$$
$$n_i^{\pm} = n\pi_i \pm \tau \max\left\{\frac{\log n}{n\gamma}, \sqrt{\frac{\pi_i \log n}{n\gamma}}\right\}, \quad i = 1, \ldots, k.$$

Let N_i be the number of visits to state i as in (6). We bound the risk by accounting for the contributions from different ranges of N_i and π_i separately:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{k} \mathbf{1}_{\{X_n=i\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right] \\
&= \sum_{i:\pi_i \geq c_n} \mathbb{E}\left[\mathbf{1}_{\{X_n=i, n_i^- \leq N_i \leq n_i^+\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right] \\
&+ \sum_{i:\pi_i \geq c_n} \mathbb{E}\left[\mathbf{1}_{\{X_n=i, N_i > n_i^+ \text{ or } N_i < n_i^-\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right] + \sum_{i:\pi_i < c_n} \mathbb{E}\left[\mathbf{1}_{\{X_n=i\}} D\left(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)\right)\right] \\
&\leq \log(n+k) \sum_{i:\pi_i \geq c_n} \mathbb{P}\left[D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)) > \epsilon(N_i), n_i^- \leq N_i \leq n_i^+\right] + \sum_{i:\pi_i \geq c_n} \mathbb{E}\left[\mathbf{1}_{\{X_n=i, n_i^- \leq N_i \leq n_i^+\}} \epsilon(N_i)\right] \\
&+ \log(n+k) \sum_{i:\pi_i \geq c_n} \left[\mathbb{P}\left[N_i \geq n_i^+\right] + \mathbb{P}\left[N_i \leq n_i^-\right]\right] + \sum_{i:\pi_i \leq c_n} \pi_i \log(n+k) \\
&\lesssim \log(n+k) \sum_{i:\pi_i \geq c_n} \mathbb{P}\left[D(M(\cdot|i)\|\widehat{M}^{+1}(\cdot|i)) > \epsilon(N_i), n_i^- \leq N_i \leq n_i^+\right] + \sum_{i:\pi_i \geq c_n} \pi_i \max_{n_i^- \leq m \leq n_i^+} \epsilon(m) \\
&\quad + \log(n+k) \sum_{i:\pi_i \geq c_n} \left(\mathbb{P}\left[N_i > n_i^+\right] + \mathbb{P}\left[N_i < n_i^-\right]\right) + \frac{k\left(\log(n+k)\right)^2}{n\gamma}. \qquad (4.31)
\end{aligned}$$

where the first inequality uses the worst-case bound (4.4.2) for add-one estimator. We analyze the terms separately as follows.

For the second term, given any i such that $\pi_i \geq c_n$, we have, by definition in (4.4.2),

$n_i^- \geq 9n\pi_i/10$ and $n_i^+ - n_i^- \leq n\pi_i/5$, which implies

$$\sum_{i:\pi_i \geq c_n} \pi_i \max_{n_i^- \leq m \leq n_i^+} \epsilon(m) \leq \sum_{i:\pi_i \geq c_n} \pi_i \left(\frac{2k}{0.9n\pi_i} + \frac{10}{9} \frac{c_0 (\log n)^3 \sqrt{k}}{n\pi_i} \right) \lesssim \frac{k^2}{n} + \frac{(\log n)^3 k^{3/2}}{n} \tag{4.32}$$

For the third term, applying (Han et al., 2018a, Lemma 16) (which, in turn, is based on ther Bernstein inequality in Paulin (2015)), we get $\mathbb{P}\left[N_i > n_i^+\right] + \mathbb{P}\left[N_i < n_i^-\right] \leq 2n^{\frac{-\tau^2}{4+10\tau}}$.

To bound the first term in (4.31), we follow the method in Billingsley (1961); Han et al. (2018a) 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 $W = \{W_{i\ell} : i = 1, \ldots, k \text{ and } \ell = 1, 2, \ldots\}$ of independent random variables, such that X and W are independent and $W_{i\ell} \overset{\text{i.i.d.}}{\sim} M(\cdot|i)$ for each i. Starting with generating X_1 from π, at every step $i \geq 2$ 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 $\{X_1, \ldots, X_n\}$ 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 $\{W_{i1}, \ldots, W_{im}\}$ 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 i.i.d. sample of size m. Lemma 31 below provides a high-probability bound for the add-one estimator in this iid setting. Using this result and the union bound, we have

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

where the second inequality applies Lemma 31 with $t = n \geq n_i^+ \geq m$ and uses $n_i^+ - n_i^- \leq n\pi_i/5$ for $\pi_i \geq c_n$.

Combining the above with (4.32), we continue (4.31) with $\tau = 25$ to get

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

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

Lemma 31 (KL risk bound for add-one estimator). *Let $V_1, \ldots, V_m \overset{iid}{\sim} Q$ 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 $t \geq m$,*

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

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

$$\begin{aligned}
D(Q\|\widehat{Q}^{+1}) &= \sum_{i=1}^k \left(Q_i \log \frac{Q_i}{\widehat{Q}_i^{+1}} - Q_i + \widehat{Q}_i^{+1}\right) \\
&= \sum_{i=1}^k \left(Q_i \log \frac{Q_i(m+k)}{m\widehat{Q}_i + 1} - Q_i + \frac{m\widehat{Q}_i + 1}{m+k}\right) \\
&= \sum_{i=1}^k \left(Q_i \log \frac{Q_i}{\widehat{Q}_i + \frac{1}{m}} - Q_i + \widehat{Q}_i + \frac{1}{m}\right) + \sum_{i=1}^k \left(Q_i \log \frac{m+k}{m} - \frac{k\widehat{Q}_i}{m+k} - \frac{k}{m(m+k)}\right) \\
&\leq \sum_{i=1}^k \left(Q_i \log \frac{Q_i}{\widehat{Q}_i + \frac{1}{m}} - Q_i + \widehat{Q}_i + \frac{1}{m}\right) + \frac{k}{m}
\end{aligned}$$

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

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

$$\mathbb{P}\left[\sum_{i=1}^k \left(Q_i \log \frac{Q_i}{\widehat{Q}_i^{\mathsf{poi}} + \frac{1}{m}} - Q_i + \widehat{Q}_i^{\mathsf{poi}} + \frac{1}{m}\right) > \frac{k}{m} + \frac{c_0(\log t)^3\sqrt{k}}{m}\right] \leq \frac{1}{t^4} \tag{4.33}$$

where $m\widehat{Q}_i^{\mathsf{poi}}, i = 1, \ldots, k$ are distributed independently as $\mathrm{Poi}(mQ_i)$. (Here and below $\mathrm{Poi}(\lambda)$ denotes the Poisson distribution with mean λ.) To see why (4.33) implies the desired

result, letting $w = \frac{k}{m} + \frac{c_0(\log t)^3\sqrt{k}}{m}$ and $Y = \sum_{i=1}^k m\widehat{Q}_i^{\mathsf{poi}} \sim \mathrm{Poi}(m)$, we have

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

where (a) followed from the fact that conditioned on their sum independent Poisson random variables follow a multinomial distribution; (b) applies (4.33); (c) follows from Stirling's approximation.

To prove (4.33) 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

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

Sub-exponential random variables satisfy the following properties (Wainwright, 2019, Sec. 2.1.3):

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

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq v\right] \leq \begin{cases} 2e^{-v^2/(2\sigma^2)}, & 0 < v \leq \frac{\sigma^2}{b} \\ 2e^{-v/(2b)}, & v > \frac{\sigma^2}{b}. \end{cases} \tag{4.34}$$

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

$$\mathbb{E}\left[|X - \mathbb{E}[X]|^\ell\right] \leq \frac{1}{2}\ell!\sigma^2 b^{\ell-2}, \quad \ell = 2, 3, \ldots. \tag{4.35}$$

- 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 32 below shows that X_i's are independent $\mathsf{SE}(\sigma^2, b)$ with $\sigma^2 = \frac{c_1(\log m)^4}{m^2}, b = \frac{c_2(\log m)^2}{n}$ 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 (4.34) for the choice $c_0 = 8(c_1 + c_2)$

this implies

$$\mathbb{P}\left[\sum_{i=1}^{k}(X_i-\mathbb{E}[X_i])\geq c_0\frac{(\log t)^3\sqrt{k}}{m}\right]\leq 2e^{-\frac{c_0^2k(\log t)^6}{2m^2\sigma^2}}+2e^{-\frac{c_0\sqrt{k}(\log t)^3}{2mb}}\leq\frac{1}{t^3}. \tag{4.36}$$

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

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

Combining the above with (4.36) we get (4.33) as required. □

Lemma 32. *There exist absolute constants c_1, c_2 such that the following holds. For any $p\in(0,1)$ and $nY\sim\mathrm{Poi}(np)$, $X=p\log\frac{p}{Y+\frac{1}{n}}-p+Y+\frac{1}{n}$ 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 (4.35), 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 $c_4\geq 24$.

Case I $p\geq\frac{c_4\ell\log n}{n}$: Using Chernoff bound for Poisson (Janson, 2002, Theorem 3)

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

we get

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

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 $\mathbb{E}[X^\ell]\lesssim\frac{\left(\sqrt{c_4\ell p\log n/4n}\right)^{2\ell}}{(p/2)^\ell}+\frac{n^\ell}{n^{2\ell}}\lesssim\left(\frac{c_4\ell\log n}{n}\right)^\ell$.

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

- On the event $\{Y > p\}$, we have $X \le Y + \frac{1}{n} \le 2Y$, where the last inequality follows because nY takes non-negative integer values. Since $X \ge 0$, we have $X^\ell \mathbf{1}_{\{Y>p\}} \le (2Y)^\ell \mathbf{1}_{\{Y>p\}}$ for any $\ell \ge 2$. Using the Chernoff bound (4.4.2), we get $Y \le \frac{2c_4 \ell \log n}{n}$ with probability at least $1 - n^{-2\ell}$, which implies

$$\begin{aligned}\mathbb{E}\left[X^\ell \mathbf{1}_{\{Y \ge p\}}\right] &\le \mathbb{E}\left[(2Y)^\ell \mathbf{1}_{\left\{Y>p, Y \le \frac{2c_4 \ell \log n}{n}\right\}}\right] + \mathbb{E}\left[(2Y)^\ell \mathbf{1}_{\left\{Y>p, Y > \frac{2c_4 \ell \log n}{n}\right\}}\right] \\ &\le \left(\frac{4c_4 \ell \log n}{n}\right)^\ell + 2^\ell \left(\mathbb{E}[Y^{2\ell}] \mathbb{P}\left[Y > \frac{2c_4 \ell \log n}{n}\right]\right)^{\frac{1}{2}} \le \left(\frac{c_5 \ell \log n}{n}\right)^\ell\end{aligned}$$

  for absolute constant c_5. Here, the last inequality follows from Cauchy-Schwarz and using the Poisson moment bound (Ahle, 2021, Theorem 2.1):[3] $\mathbb{E}[(nY)^{2\ell}] \le \left(\frac{2\ell}{\log\left(1+\frac{2\ell}{np}\right)}\right)^{2\ell} \le (c_6 \ell \log n)^{2\ell}$ 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.

□

Proof of Corollary 18

We show the following monotonicity result of the prediction risk. In view of this result, Corollary 18 immediately follows from Theorem 16 and Theorem 17 (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 33. $\mathsf{Risk}_{k+1,n}(\gamma_0) \ge \mathsf{Risk}_{k,n}(\gamma_0)$ *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

[3] For a result with less precise constants, see also (Ahle, 2021, Eq. (1)) based on (Latała, 1997, Corollary 1).

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 $(X_1, \ldots, X_n)$ and $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ be stationary Markov chains with transition matrices M and $\widetilde{M}$, respectively. Then as $\delta \to 0$, $(X_1, \ldots, X_n)$ converges to $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ 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 $i = 1, \ldots, k$. Then

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

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 $\alpha = \sum q_i \leq 1$, we have $D(P\|Q) = D(P\|\bar{Q}) + \log\frac{1}{\alpha} \geq D(P\|\bar{Q})$. 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 $\mathsf{Risk}_{k+1,n}(\gamma_0) \geq \mathsf{Risk}_{k,n}(\gamma_0)$. □

4.5 Higher-order Markov chains

In this section we prove Theorem 19. We start with some basic definitions for higher-order Markov chains. Let $m \geq 1$. Let $X_1, X_2, \ldots$ be an m^{th}-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 $t \geq m$. Clearly, the joint distribution of the process is specified by the transition matrix and the initial distribution, which is a joint distribution for $(X_1, \ldots, X_m)$.

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

$$(X_{i_1+t}, \ldots, X_{i_n+t}) \overset{\text{law}}{=} (X_{i_1}, \ldots, X_{i_n}), \quad \forall n, i_1, \ldots, i_n \in \mathbb{N}, t \in \mathbb{Z}_+.$$

It is clear that (4.5) is equivalent to $(X_1, \ldots, X_m) \overset{\text{law}}{=} (X_2, \ldots, X_{m+1})$. In other words, π is the solution to the linear system:

$$\pi(x_1, \ldots, x_m) = \sum_{x_0 \in \mathcal{S}} \pi(x_0, x_1, \ldots, x_{m-1}) M(x_m | x_1, \ldots, x_{m-1}), \quad \forall x_1, \ldots, x_m \in \mathcal{S} \quad (4.37)$$

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

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

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

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

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 34. *An m^{th}-order stationary Markov chain is reversible if and only if (4.5) holds*

for $n = m + 1$, namely

$$\pi(x_1, \ldots, x_m)M(x_{m+1}|x_1, \ldots, x_m) = \pi(x_{m+1}, \ldots, x_2)M(x_1|x_{m+1}, \ldots, x_2), \quad \forall x_1, \ldots, x_{m+1} \in \mathcal{S}. \tag{4.38}$$

Proof. First, we show that (4.5) for $n = m + 1$ implies that for $n \leq m$. Indeed,

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

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 (4.5) for $n = m + 2$ and the rest follows from induction on n. Indeed,

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

where (a) and (c) apply (4.5) for $n = m+1$, namely, (4.38); (b) applies (4.5) for $n = m$. □

Note that any distribution π on $\mathcal{S}^m$ and m^{th}-order transition matrix M that satisfy $\pi(x^m) = \pi(\overline{x^m})$ and (4.38) also satisfy (4.37). This implies such a π with be the stationary distribution for M. In view of Lemma 34 the above conditions also guarantee reversibility. This observation can be summarized in the following lemma, which we will use to prove reversibility of specific Markov chains later.

Lemma 35. *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 $\pi(x^m) = \pi(\overline{x^m})$ and $\pi(x^m)M(x_{m+1}|x^m) = \pi(\overline{x_2^{m+1}})M(x_1|\overline{x_2^{m+1}})$. Then π is the stationary distribution of M and the resulting chain is reversible.*

We define the prediction risk as

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

where the suppremum is taken over all $k^m \times k$ stochastic matrices M and the trajectory is initiated from the stationary distribution. Based on these definitions we will show the following.

Theorem 36. *There exist constants $C_m > 0$ such that for all $m \geq 2, 2 \leq k \leq \sqrt[m+1]{n}/C_m$*

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

The lower bound is achieved over the class of all reversible Markov chains.

4.5.1 Upper bound

We prove the upper bound part of the preceding theorem, using only stationarity (not reversibility). Our proof uses techniques in (Csiszár and Shields, 2004, Chapter 6, Page 486) for proving redundancy bounds for the m^{th}-order chains. Let Q be the probability assignment given by

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

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

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

Then in view of Lemma 20, the following lemma generates the desired upper bound in Theorem 36.

Lemma 37. *Let* $\mathsf{Red}(Q_{X^n})$ *be the redundancy of the* m^{th}*-order Markov chain, as defined in Section 4.2.1, and* X^m *be the corresponding observed trajectory. Then*

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

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

$$\log\frac{\pi(x_1^m)\prod_{t=m}^{n-1}M(x_{t+1}|x_{t-m+1}^t)}{Q(x_1,\cdots,x_n)} \leq k^m(k-1)\left[\log\left(1+\frac{n-m}{k^m(k-1)}\right)+1\right]+m\log k,$$

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

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

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

$$\frac{\pi(x_1^m)\prod_{t=m}^{n-1}M(x_{t+1}|x_{t-m+1}^t)}{Q(x_1,\cdots,x_n)}\leq k^m\cdot\prod_{a^m\in\mathcal{S}^m}\frac{k\cdot(k+1)\cdot\cdots\cdot(N_{a^m}+k-1)}{N_{a^m}^{N_{a^m}}}\prod_{a_{m+1}\in\mathcal{S}}\frac{N_{a^{m+1}}^{N_{a^{m+1}}}}{N_{a^{m+1}}!}. \tag{4.39}$$

Using (4.7) we continue (4.39) to get

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

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 $\log(1+x) \le x$. □

4.5.2 Lower bound

Special case $m \ge 2, k = 2$

We only analyze the case $m = 2$, i.e. second-order Markov chains with binary states, as the lower bound for the case of $m \ge 3$ case is then implied. 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 $\ell \in \mathcal{S}$:

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

Our result is the following.

Theorem 38. $\mathsf{Risk}_{2,n,2} = \Theta\left(\frac{\log n}{n}\right)$.

Proof. The upper bound part follows from that of first-order Markov chains. Given any m^{th}-order chain $\{X_t\}_{t\ge1}$ on the state space $[k]$, the process $\{Y_t\}_{t\ge1}$, with $Y_t \triangleq X_t^{t+m-1}$ consisting of adjacent m-tuples, is a first-order Markov chain with states space $\mathcal{S}^m$. Applying the data processing inequality, the KL risk of estimating $P_{X_{n+1}|X_{n-m+1}^{n-1}}$ is at most that of estimating $P_{Y_{n+2-m}|Y_{n+1-m}}$. As such, we have $\mathsf{Risk}_{k,n,m} \le \mathsf{Risk}_{k^m,n-m+1}$. This implies that $\mathsf{Risk}_{2,n,2} \le \mathsf{Risk}_{4,n} = \Theta(\log n/n)$, in view of Theorem 15.

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)

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

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 (4.8) for analyzing the first-order risk, we define

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

In other words, the sequences in $\mathcal{V}$ starts with a sequence of 1's, then transitions into two consecutive 2's, has no consecutive 1's afterwards, and end with 2 as well. Suppose that the operation $\oplus$ combines any two blocks from $\{22, 212\}$ via merging the end digit of the first block and the first digit of the second block

$$22 \oplus 212 = 2212, 22 \oplus 22 = 222, 212 \oplus 22 = 2122, 212 \oplus 212 = 21212.$$

Then for any $x^n \in \mathcal{V}$ we can write it using the initial part containing all 1's, alternating run of blocks from $\{22, 212\}$ with the first run being of the block 22 (all the runs have positive lengths), and the merging operation $\oplus$

$$x^n = \underbrace{1 \ldots 1}_{\text{all ones}} \underbrace{22 \oplus 22 \cdots \oplus 22}_{p_1 \text{ many } 22} \oplus \underbrace{212 \oplus 212 \cdots \oplus 212}_{p_2 \text{ many } 212} \oplus \underbrace{22 \oplus 22 \cdots \oplus 22}_{p_3 \text{ many } 22} \oplus \underbrace{212 \oplus 212 \cdots \oplus 212}_{p_4 \text{ many } 212} \oplus 22 \oplus \ldots. \tag{4.40}$$

Suppose that the vector $(q_{22\to22}, q_{22\to212}, q_{212\to22}, q_{212\to212})$ gives the transition probabilities between blocks in $\{22, 212\}$ (note that the blocks share common adjacent random variable that joins them)

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

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

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

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

$$\begin{aligned}
&\mu(x^n|p) \\
&= \mathbb{P}(X^{F_{111}+2} = 1^{F_{111}+2}) M(2|11) M(2|12) \prod_{a,b\in\{22,212\}} q_{a\to b}^{F_{a\to b}} \\
&= \frac{1}{4}\left(1-\frac{1}{n}\right)^{F_{111}} \frac{1}{n} \cdot p \cdot p^{F_{212\to22}} \left\{p\left(1-\frac{1}{n}\right)\right\}^{F_{22\to212}} (1-p)^{F_{22\to22}} \left\{(1-p)\left(1-\frac{1}{n}\right)\right\}^{F_{212\to212}} \\
&= \frac{1}{4}\left(1-\frac{1}{n}\right)^{F_{111}+F_{22\to212}+F_{212\to212}} \frac{1}{n} p^{y+1}(1-p)^{f-y} \qquad (4.41)
\end{aligned}$$

where $y = F_{212\to22} + F_{22\to212}$ denotes the number of times we alternate between the run of 22 and the run of 212, and $f = F_{212\to22} + F_{22\to212} + F_{212\to212} + F_{22\to22}$ denotes the number of times we jump between blocks from $\{22, 212\}$.

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. The number of sequences in $\mathcal{V}$ for any given realization of (y, f) is $\binom{f}{y}$. A short proof is provided. 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 (4.40). The p_i's are all positive 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 in the string from $\{22, 212\}$ is one more than its total number of f-transitions. Each positive solution to this equation $\{p_i\}_{i=1}^{y+1}$ gives us one single sequence $x^n \in \mathcal{V}$ and vice versa. The total number of such sequences is $\binom{f}{y}$.

For any $x^n \in \mathcal{V}$ with given $\boldsymbol{F}$ the Bayes estimator of p (under squared error loss and $p \sim \text{Uniform}[0,1]$) is

$$\widehat{p}(x^n) = \mathbb{E}[p|x^n] = \frac{\mathbb{E}[p \cdot \mu(x^n|p)]}{\mathbb{E}[\mu(x^n|p)]} \overset{(4.41)}{=} \frac{y+2}{f+3}.$$

Note that the probabilities $\mu(x^n|p)$ in (4.41) 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 the particular sequence x^n

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

where last equality followed by noting that the integral is the variance of a $\text{Beta}(y+2, f-y+1)$ 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

$$D(M(\cdot|12)\|\widehat{M}(\cdot|12)) \geq \frac{1}{2}\|M(\cdot|12) - \widehat{M}(\cdot|12)\|_{\ell_1}^2 \geq \frac{1}{2}(p - \widehat{M}(2|12))^2$$

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. Then we have

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

□

General case $m \geq 2, k \geq 3$

We will prove the following.

Theorem 39. *For absolute constant C, we have*

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

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:

$$M(s|x^m) = \begin{array}{|c|c|c|}\hline \text{Starting string } x^m & \multicolumn{2}{c|}{\text{Next state}} \\ \cline{2-3} & s=1 & s\in\{2,\dots,k\} \\ \hline 1^m & 1-\frac{1}{n} & \frac{1}{n(k-1)} \\ \hline 1x^{m-1}, x^{m-1}\in\widetilde{\mathcal{S}}^{m-1} & 1-b & \frac{b}{(k-1)} \\ \hline x^m\in\widetilde{\mathcal{S}}^m & \frac{1}{n} & \left(1-\frac{1}{n}\right)T(s|x^m) \\ \hline x^m\notin\left\{1^m, 1\widetilde{\mathcal{S}}^{m-1}, \widetilde{\mathcal{S}}^m\right\} & \frac{1}{2} & \frac{1}{2(k-1)} \\ \hline \end{array}, \quad b=\frac{1}{2}-\frac{2^m-2}{n}. \tag{4.42}$$

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 40. *Under the condition (P), the transition matrix T is reversible and its stationary distribution is uniform on $\widetilde{\mathcal{S}}^m$. Under the condition (P), the m^{th}-order Markov chain with transition matrix T and uniform initial distribution on $\widetilde{\mathcal{S}}^m$ is reversible (and hence stationary).*

Proof. We prove this result using Lemma 35. 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 $\pi(x^m) = \pi(\overline{x^m})$

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). □

Lemma 41. *M is a reversible transition matrix with the stationary distribution π given by*

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

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 (4.42).

Proof. Note that π is exchangeable, which enforces that all the lower dimensional distributions are well defined, and the choice of b guarantees that $\sum_{x^m \in \mathcal{S}^m} \pi(x^m) = 1$. The condition $\pi(x^m) = \pi(\overline{x^m})$ for all $x^m \in \mathcal{S}^m$ follows directly. 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

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

  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 (4.42). We observe the following properties:

(R1) This Markov chain is irreducible and reversible with mass $\frac{1}{2}$ on the state 1^m.

(R2) For $m \leq t \leq n-1$, 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 41

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

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,

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

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 (4.42) and stationary distribution π in (41). The following lemma lower bounds the Bayes prediction risk.

Lemma 42. *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*

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

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

(4.4) of the Bayes estimator, we have, for each $j \in \widetilde{\mathcal{S}}$,

$$\begin{aligned}
&\widehat{M}_n(j|x^n) \\
&= \frac{\mathbb{P}\left[X^{n+1} = (x^n, j)\right]}{\mathbb{P}\left[X^n = x^n\right]} \\
&= \frac{\mathbb{E}[\frac{1}{2} \cdot \left(1 - \frac{1}{n}\right)^{t-m} \cdot \frac{b}{n2^{m-2}(k-1)^m} \prod_{s=t+m}^{n-1} M(x_{s+1}|x_{s-m+1}^s) M(j|x_{n-m+1}^n)]}{\mathbb{E}[\frac{1}{2} \cdot \left(1 - \frac{1}{n}\right)^{t-m} \cdot \frac{b}{n2^{m-2}(k-1)^m} \prod_{s=t+m}^{n-1} M(x_{s+1}|x_{s-m+1}^s)]} \\
&= \left(1 - \frac{1}{n}\right) \frac{\mathbb{E}[\frac{1}{(k-1)^m} \prod_{s=t+m}^{n-1} T(x_{s+1}|x_{s-m+1}^s) T(j|x_{n-m+1}^n)]}{\mathbb{E}[\frac{1}{(k-1)^m} \prod_{s=t+m}^{n-1} T(x_{s+1}|x_{s-m+1}^s)]} \\
&= \left(1 - \frac{1}{n}\right) \frac{\mathbb{P}\left[Y^{n-t+1} = (x_{t+1}^n, j)\right]}{\mathbb{P}\left[Y^{n-t} = x_{t+1}^n\right]} \\
&= \left(1 - \frac{1}{n}\right) \widehat{T}_{n-t}(j|x_{t+1}^n).
\end{aligned}$$

Furthermore, since $M(1|x^m) = 1/n$ for all $x^m \in \widetilde{\mathcal{S}}$ in the construction (4.42), 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

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

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

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

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,

$$\begin{aligned}
&\mathbb{E}_T \left[\mathbb{E}[D(M(\cdot|X_{n-1}X_n) \| \widehat{M}(\cdot|X^n))]\right] \\
&\geq \frac{b}{n2^{m-1}} \left(1 - \frac{1}{n}\right)^{n-2m} \sum_{t=m}^{n-m} \mathbb{E}_T \left[\mathbb{E}[D(M(\cdot|X_{n-m+1}^n) \| \widehat{M}(\cdot|X^n))|X^n \in \mathcal{X}_t]\right]. \quad (4.43)
\end{aligned}$$

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

$$\mathbb{E}_T\left[\mathbb{E}[D(M(\cdot|X_{n-m+1}^n)\|\widehat{M}(\cdot|X^n))|X^n\in\mathcal{X}_t]\right]\overset{\text{(a)}}{=}\left(1-\frac{1}{n}\right)\cdot\mathbb{E}_T\left[\mathbb{E}[D(T(\cdot|Y_{n-t-m+1}^{n-t})\|\widehat{T}(\cdot|Y^{n-t}))]\right]$$
$$\overset{\text{(b)}}{=}\left(1-\frac{1}{n}\right)\cdot I(T;Y_{n-t+1}|Y^{n-t}),\qquad(4.44)$$

where (a) follows from (4.5.2), (4.5.2), 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 (4.2.1) of the Bayes prediction risk. Finally, the lemma follows from (4.43), (4.44), and the chain rule

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

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

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

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

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 43. *Suppose that T is an $\ell^m \times \ell$ transition matrix, on statespace $\mathcal{Y}^m$ with $|\mathcal{Y}| = \ell$, satisfying $T(x_{m+1}|x^m) = T(x_1|\overline{x_2^{m+1}})$, $\forall x^{m+1} \in [\ell]^{m+1}$ and $T(y_{m+1}|y^m) \in [\frac{c_1}{\ell}, \frac{c_2}{\ell}]$ with $0 < c_1 < c_2 < 1 < c_1$ 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*

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

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 Frobenious 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

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

Here and below, we identify $u = \left\{u_{ix^{m-1}j}\right\}_{i \leq j, x^{m-1} \in R_{m-1}}$ and $\widehat{u} = \left\{\widehat{u}_{ix^{m-1}j}\right\}_{i \leq j, x^{m-1} \in R_{m-1}}$ 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. Cover and Thomas (2006)). Then

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

Then

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

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 (4.5.2); (b) follows from the data processing inequality; (c) is because $h(\cdot)$ is translation invariant and concave; (d) follows from the maximum entropy principle Cover and Thomas (2006): $h(u - \widehat{u}(Y^n)) \leq \frac{|R_{m-1}|(k^2-1)}{16} \log\left(\frac{2\pi e}{|R_{m-1}|(k^2-1)/8} \cdot \mathbb{E}[\|\widehat{u}(Y^n) - u\|_2^2]\right)$, which in turn is bounded by (4.5.2). Plugging this lower bound into Lemma 42 completes the lower bound proof of Theorem 36.

Proof of Lemma 43 via pseudo spectral gap

In view of Lemma 40 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

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

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 $\mathbb{E}[\widehat{T}(y_{m+1}|y^m)] = \ell^m \mathbb{P}\left[y_{m+1}|y^m\right] = T(y_{m+1}|y^m)$. Next we observe that the sequence of random variables $\left\{Y_t^{t+m}\right\}_{t=1}^{n-m}$ 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 might must be non-reversible, the *pseudo spectral gap* $\gamma_{\mathrm{ps}}(T_{m+1})$ is defined as

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

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 (Paulin, 2015, Theorem 3.2) gives the

following variance bound:

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

The following lemma bounds the pseudo spectral gap from below.

Lemma 44. *Let $T \in \mathbb{R}^{\ell^m \times \ell}$ be the transition matrix of an m-th order Markov chain $(Y_t)_{t\ge 1}$ 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 $((Y_t, Y_{t+1}, \cdots, Y_{t+m}))_{t\ge 1}$. 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 44. 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

$$\min_{a^{2m+2} \in \mathcal{Y}^{2m+2}} T_{m+1}^{m+1}(a_{m+2}^{2m+2}|a^{m+1}) \ge \prod_{t=0}^{m} T(a_{2m+2-t}|a_{m+2-t}^{2m+1-t}) \ge \frac{c_1^{m+1}}{\ell^{m+1}}. \tag{4.45}$$

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

$$
\begin{aligned}
&(T^*_{m+1})^{m+1}(b^{m+1}|a^{m+1}) \\
&= \sum_{\boldsymbol{y}_1 \in \mathcal{Y}^{m+1}, \ldots, \boldsymbol{y}_m \in \mathcal{Y}^{m+1}} T^*_{m+1}(b^{m+1}|\boldsymbol{y}_m) \left\{ \prod_{t=1}^{m-1} T^*_{m+1}(\boldsymbol{y}_{m-t+1}|\boldsymbol{y}_{m-t}) \right\} T^*_{m+1}(\boldsymbol{y}_1|a^{m+1}) \\
&= \sum_{\boldsymbol{y}_1 \in \mathcal{Y}^{m+1}, \ldots, \boldsymbol{y}_m \in \mathcal{Y}^{m+1}} \frac{\pi(b^{m+1}) T_{m+1}(\boldsymbol{y}_m|b^{m+1})}{\pi(\boldsymbol{y}_m)} \left\{ \prod_{t=1}^{m-1} \frac{\pi(\boldsymbol{y}_{m-t+1}) T_{m+1}(\boldsymbol{y}_{m-t}|\boldsymbol{y}_{m-t+1})}{\pi(\boldsymbol{y}_{m-t})} \right\} \\
&\qquad\qquad \frac{\pi(\boldsymbol{y}_1) T_{m+1}(a^{m+1}|\boldsymbol{y}_1)}{\pi(a^{m+1})} \\
&= \frac{\pi(b^{m+1})}{\pi(a^{m+1})} \sum_{\boldsymbol{y}_1 \in \mathcal{Y}^{m+1}, \ldots, \boldsymbol{y}_m \in \mathcal{Y}^{m+1}} T_{m+1}(\boldsymbol{y}_m|b^{m+1}) \left\{ \prod_{t=1}^{m-1} T_{m+1}(\boldsymbol{y}_{m-t}|\boldsymbol{y}_{m-t+1}) \right\} T_{m+1}(a^{m+1}|\boldsymbol{y}_1) \\
&= \frac{\pi(b^{m+1})}{\pi(a^{m+1})} T^{m+1}_{m+1}(a^{m+1}|b^{m+1}) \\
&= \frac{\pi(b^m) T(b_{m+1}|b^m)}{\pi(b^m) T(a_{m+1}|a^m)} T^{m+1}_{m+1}(a^{m+1}|b^{m+1}) \geq \frac{c_1}{c_2} \cdot \frac{c_1^{m+1}}{\ell^{m+1}}. \qquad (4.46)
\end{aligned}
$$

Using (4.45),(4.46) we get

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

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 45 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

$$
\gamma_{\mathrm{ps}}(T_{m+1}) \geq \frac{\gamma((T^*_{m+1})^{m+1} T^{m+1}_{m+1})}{m+1} \geq \frac{c_1^{2m+3}}{c_2(m+1)}
$$

as required. A more generalized version of Lemma 45 can be found in from Hoffman (1967).

Lemma 45. *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

$$|\lambda v_j| = |\sum_{i=1}^d v_i A_{ij}| = |\sum_{i=1}^d v_i (A_{ij} - \epsilon)| \le \sum_{i=1}^d |v_i (A_{ij} - \epsilon)| = \sum_{i=1}^d |v_i| (A_{ij} - \epsilon)$$

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

□

4.6 Discussions and open problems

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

Very large state space. Theorem 15 determines the optimal prediction risk under the assumption of $k \lesssim \sqrt{n}$. When $k \gtrsim \sqrt{n}$, Theorem 15 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 (4.1.1), reduces to determining the redundancy for symmetric and general Markov chains. For iid data, the minimax *pointwise* redundancy is known to be $n \log \frac{k}{n} + O(\frac{n^2}{k})$ (Szpankowski and Weinberger, 2012, 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 (4.1.1), 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.

Other loss functions As mentioned in Section 4.1.1, standard arguments based on concentration inequalities inevitably rely on mixing conditions such as the spectral gap. In contrast, the risk bound in Theorem 15, 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[4] in (4.1), which allows one to use the mutual information representation and the chain rule to relate individual risks to the cumulative risk. More general loss in terms of f-divergence have been considered in Hao et al. (2018). Obtaining spectral gap-independent risk bound for these loss functions, this time without the aid of universal compression, is an open question.

Stationarity As mentioned above, the redundancy result in Lemma 21 (see also Davisson (1983); Tatwawadi et al. (2018)) holds for nonstationary Markov chains as well. However, our redundancy-based risk upper bound in Lemma 20 crucially relies on stationarity. It is unclear whether the result of Theorem 15 carries over to nonstationary chains.

4.7 Appendix

4.7.1 Mutual information representation of prediction risk

The following lemma justifies the representation (4.2.1) for the prediction risk as maximal conditional mutual information. Unlike (4.2.1) for redundancy which holds essentially without any condition Kemperman (1974), here we impose certain compactness assumptions which hold finite alphabets such as finite-state Markov chains studied in this paper.

Lemma 46. *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*

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

Then

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

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

[4] In fact, this connection breaks down if one swap M and $\widehat{M}$ in the KL divergence in (4.1).

Note that for stationary Markov chains, (4.2.1) follows from Lemma 46 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 (46) is equivalent to

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

By the variational representation (4.1.3) of conditional mutual information, we have

$$I(\theta; X_{n+1}|X^n) = \inf_{Q_{X_{n+1}|X^n}} D(P_{X_{n+1}|X^n,\theta} \| Q_{X_{n+1}|X^n} | P_{X^n,\theta}).$$

Thus (46) amounts to justifying the interchange of infimum and supremum in (46). It suffices to prove the upper bound.

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

$$\mathsf{Risk}_{n,\epsilon} \triangleq \inf_{Q_{X_{n+1}|X^n} \geq \frac{\epsilon}{K}} \sup_{P_\theta \in \mathcal{M}(\Theta)} D(P_{X_{n+1}|X^n,\theta} \| Q_{X_{n+1}|X^n} | P_{X^n,\theta}),$$

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

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

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

We first show that the infimum and supremum in (4.7.1) 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.,

(Fan, 1953, Theorem 2)) yields

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

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

$$D(P\|(1-\epsilon)Q+\epsilon U) \le (1-\epsilon)D(P\|Q) + \epsilon D(P\|U) \le (1-\epsilon)D(P\|Q) + \epsilon \log K,$$

which, in view of (4.7.1) and (4.7.1), 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, (46) follows. □

4.7.2 Proof of Lemma 30

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 47. *For any $1 \le t_1 < \cdots < t_m < \cdots < t_k$ and any $Z_2 = f(X_{t_k}, \ldots, X_{t_m}), Z_1 = g(X_{t_{m-1}}, \ldots, X_{t_1})$ we have*

$$\mathbb{E}\left[Z_2 \mathbf{1}_{\{X_{t_m}=j\}} Z_1 | X_1 = i\right] = \mathbb{E}\left[Z_2 | X_{t_m} = j\right] \mathbb{E}\left[\mathbf{1}_{\{X_{t_m}=j\}} Z_1 | X_1 = i\right]$$

For $t \ge 0$, denote the t-step transition probability by $\mathbb{P}[X_{t+1} = j | X_1 = i] = M^t(j|i)$, which is the ijth entry of M^t. The following result is standard (see, e.g., (Levin and Peres, 2017a, Chap. 12)). We include the proof mainly for the purpose of introducing the spectral decomposition.

Lemma 48. *Define* $\lambda_* \triangleq 1-\gamma_* = \max\{|\lambda_i| : i \neq 1\}$. *For any* $t \geq 0$, $\left|M^t(j|i) - \pi_j\right| \leq \lambda_*^t \sqrt{\frac{\pi_j}{\pi_i}}$.

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, $D_\pi^{\frac{1}{2}} M D_\pi^{-\frac{1}{2}}$, which shares the same spectrum with M, is a symmetric matrix and admits the spectral decomposition $D_\pi^{\frac{1}{2}} M D_\pi^{-\frac{1}{2}} = \sum_{a=1}^k \lambda_a u_a u_a^\top$ for some orthonormal basis $\{u_1, \ldots, u_k\}$; in particular, $\lambda_1 = 1$ and $u_{1i} = \sqrt{\pi_i}$. Then for each $t \geq 1$,

$$M^t = \sum_{a=1}^k \lambda_a^t D_\pi^{-\frac{1}{2}} u_a u_a^\top D_\pi^{\frac{1}{2}} = \mathbf{1}\pi + \sum_{a=2}^k \lambda_a^t D_\pi^{-\frac{1}{2}} u_a u_a^\top D_\pi^{\frac{1}{2}}. \tag{4.47}$$

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 $\sum_{a=2}^k u_{ab}^2 = 1 - u_{a1}^2 \leq 1$ for any $b = 1, \ldots, k$. Using this along with Cauchy-Schwarz inequality we get

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

as required. □

Lemma 49. *Fix states* i, j. *For any integers* $a \geq b \geq 1$, *define*

$$h_s(a,b) = \left|\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i\}} \left(\mathbf{1}_{\{X_a=j\}} - M(j|i)\right)^s | X_b = i\right]\right|, \quad s = 1,2,3,4.$$

Then

(i) $h_1(a,b) \leq 2\sqrt{M(j|i)}\lambda_*^{a-b}$

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

(iii) $h_3(a,b), h_4(a,b) \leq \pi_i M(j|i)(1-M(j|i)) + 4\sqrt{M(j|i)}\lambda_*^{a-b}$.

Proof. We apply Lemma 48 and time reversibility:

(i)

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

(ii)

$$\begin{aligned}
&|h_2(a,b) - \pi_i M(j|i)(1-M(j|i))| \\
=&\Big|\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i,X_a=j\}}|X_b=i\right] - \pi_i M(j|i) + (M(j|i))^2\left(\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i\}}|X_b=i\right]-\pi_i\right) \\
&- 2M(j|i)(\mathbb{E}\left[\mathbf{1}_{\{X_{a+1}=i,X_a=j\}}|X_b=i\right]-\pi_i M(j|i))\Big| \\
\le&\left|\mathbb{P}\left[X_{a+1}=i,X_a=j|X_b=i\right]-\pi_j M(i|j)\right| + (M(j|i))^2\left|\mathbb{P}\left[X_{a+1}=i|X_b=i\right]-\pi_i\right| \\
&+ 2M(j|i)\left|\mathbb{P}\left[X_{a+1}=i,X_a=j|X_b=i\right]-\pi_j M(i|j)\right| \\
=&M(i|j)\left|M^{a-b}(j|i)-\pi_j\right| + (M(j|i))^2\left|M^{a-b+1}(i|i)-\pi_i\right| + 2M(j|i)M(i|j)\left|M^{a-b}(j|i)-\pi_j\right| \\
\le&M(i|j)\sqrt{\frac{\pi_j}{\pi_i}}\lambda_*^{a-b} + (M(j|i))^2\lambda_*^{a-b+1} + 2M(j|i)M(i|j)\sqrt{\frac{\pi_j}{\pi_i}}\lambda_*^{a-b} \\
\le&\lambda_*^{a-b}\left(\sqrt{M(i|j)}\sqrt{\frac{M(i|j)\pi_j}{\pi_i}} + (M(j|i))^2 + 2M(j|i)\sqrt{M(i|j)}\sqrt{\frac{M(i|j)\pi_j}{\pi_i}}\right) \\
\le&4\sqrt{M(j|i)}\lambda_*^{a-b}.
\end{aligned}$$

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

Proof of Lemma 30(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

(6) 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

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

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 49.

Case I: Two distinct indices. For any $a > b$, using Lemma 47 we get

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

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.

Case II: Single index.

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

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 30(ii). We first note that due to reversibility we can write (similar as in proof of Lemma 30(i)) with $\eta_a = \mathbf{1}_{\{X_{a+1}=i\}}\left(\mathbf{1}_{\{X_a=j\}} - M(j|i)\right)$

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

We bound the sum over different combinations of $a \geq b \geq d \geq e$ 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 47: 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 49 to bound the h functions.

Case I: Four distinct indices. Using Lemma 49 we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Case IV: Single index. Bound on $\sum_{a=1}^{n-1}\mathbb{E}\left[\eta_a^4|X_1=i\right]$:

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

Combining all cases we get

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

as required. □

Proof of Lemma 30(iii). Throughout our proof we repeatedly use the spectral decomposi-

tion (4.47) 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 $a \geq b \geq d \geq e$,

$$\begin{aligned}
&\mathbb{E}\left[\xi_a\xi_b\xi_d\xi_e|X_1=i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b\left(\mathbf{1}_{\{X_d=i,X_e=i\}} - \pi_i\mathbf{1}_{\{X_d=i\}} - \pi_i\mathbf{1}_{\{X_e=i\}} + \pi_i^2\right)|X_1=i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b\mathbf{1}_{\{X_d=i,X_e=i\}}|X_1=i\right] - \pi_i\mathbb{E}\left[\xi_a\xi_b\mathbf{1}_{\{X_d=i\}}|X_1=i\right] \\
&\quad - \pi_i\mathbb{E}\left[\xi_a\xi_b\mathbf{1}_{\{X_e=i\}}|X_1=i\right] + \pi_i^2\mathbb{E}\left[\xi_a\xi_b|X_1=i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b|X_d=i\right]\mathbb{P}\left[X_d=i|X_e=i\right]\mathbb{P}[X_e=i|X_1=i] - \pi_i\mathbb{E}\left[\xi_a\xi_b|X_d=i\right]\mathbb{P}[X_d=i|X_1=i] \\
&\quad - \pi_i\mathbb{E}\left[\xi_a\xi_b|X_e=i\right]\mathbb{P}[X_e=i|X_1=i] + \pi_i^2\mathbb{E}\left[\xi_a\xi_b|X_1=i\right] \\
&= \mathbb{E}\left[\xi_a\xi_b|X_d=i\right]\left\{M^{d-e}(i|i)M^{e-1}(i|i) - \pi_i M^{d-1}(i|i)\right\} \\
&\quad - \left\{\pi_i\mathbb{E}\left[\xi_a\xi_b|X_e=i\right]M^{e-1}(i|i) - \pi_i^2\mathbb{E}\left[\xi_a\xi_b|X_1=i\right]\right\} \qquad (4.48)
\end{aligned}$$

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

$$\begin{aligned}
&\left| \mathbb{E}[\xi_a \xi_b | X_d = i] - \pi_i \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-b} \right| \\
&= \left| \mathbb{E}\left[\mathbf{1}_{\{X_a = i, X_b = i\}} - \pi_i \mathbf{1}_{\{X_a = i\}} - \pi_i \mathbf{1}_{\{X_b = i\}} + \pi_i^2 | X_d = i \right] - \pi_i \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-b} \right| \\
&= \left| M^{a-b}(i|i) M^{b-d}(i|i) - \pi_i M^{a-d}(i|i) - \pi_i M^{b-d}(i|i) + \pi_i^2 - \pi_i \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-b} \right| \\
&= \left| \left(\pi_i + \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-b} \right) \left(\pi_i + \sum_{v \ge 2} u_{vi}^2 \lambda_v^{b-d} \right) \quad - \pi_i \left(\pi_i + \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-d} \right) \right. \\
&\quad \left. - \pi_i \left(\pi_i + \sum_{v \ge 2} u_{vi}^2 \lambda_v^{b-d} \right) + \pi_i^2 - \pi_i \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-b} \right| \\
&= \left| \left(\sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-b} \right) \left(\sum_{v \ge 2} u_{vi}^2 \lambda_v^{b-d} \right) - \pi_i \sum_{v \ge 2} u_{vi}^2 \lambda_v^{a-d} \right| \\
&\le \lambda_*^{a-d} \left(\sum_{v \ge 2} u_{vi}^2 \right) \left(\sum_{v \ge 2} u_{vi}^2 \right) + \lambda_*^{a-d} \pi_i \sum_{v \ge 2} u_{vi}^2 \le 2 \lambda_*^{a-d}. \qquad (4.49)
\end{aligned}$$

We also get for $d \ge e$

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

This implies

$$
\begin{aligned}
&|\mathbb{E}\left[\xi_a\xi_b|X_d=i\right]|\left|M^{d-e}(i|i)M^{e-1}(i|i)-\pi_i M^{d-1}(i|i)\right| \\
&\leq \left(\pi_i\sum_{v\geq 2}u_{vi}^2\lambda_v^{a-b}+2\lambda_*^{a-d}\right)\left(2\lambda_*^{d-1}+\pi_i\lambda_*^{e-1}+\pi_i\lambda_*^{d-e}\right) \\
&\leq \left(\pi_i\lambda_*^{a-b}+2\lambda_*^{a-d}\right)\left(2\lambda_*^{d-1}+\pi_i\lambda_*^{e-1}+\pi_i\lambda_*^{d-e}\right) \\
&\leq 4\left[\pi_i^2\lambda_*^{a-b+d-e}+\pi_i^2\lambda_*^{a-b+e-1}+\pi_i\left(\lambda_*^{a-b+d-1}+\lambda_*^{a-d+e-1}+\lambda_*^{a-e}\right)+\lambda_*^{a-1}\right] \quad (4.50)
\end{aligned}
$$

Using (4.49) along with Lemma 48 for any $e\leq b\leq a$ we get

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

This together with (4.50) and (4.48) implies

$$
\begin{aligned}
|\mathbb{E}\left[\xi_a\xi_b\xi_d\xi_e|X_1=i\right]| \lesssim \pi_i^2\left(\lambda_*^{a-b+d-e}+\lambda_*^{a-b+e-1}\right)+\lambda_*^{a-1} \\
+\pi_i\left(\lambda_*^{a-b+d-1}+\lambda_*^{a-d+e-1}+\lambda_*^{a-e}\right)
\end{aligned} \quad (4.51)
$$

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.

Case I: four distinct indices. We sum (4.51) 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

$$\sum_{n-1\geq a>b>d>e\geq 1} |\mathbb{E}\left[\xi_a\xi_b\xi_d\xi_e | X_1 = i\right]| \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{n\pi_i}{\gamma_*^3} + \frac{1}{\gamma_*^4} \lesssim \frac{n^2\pi_i^2}{\gamma_*^2} + \frac{1}{\gamma_*^4}. \quad (4.52)$$

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 (4.51) 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

$$\begin{aligned}
&\sum_{n-1\geq a>b>e\geq 1}\sum\sum\left|\mathbb{E}\left[\xi_a\xi_b^2\xi_e|X_1=i\right]\right| \\
&\lesssim \sum_{n-1\geq a>b>e\geq 1}\sum\sum\left\{\pi_i\left(\lambda_*^{a-b+e-1}+\lambda_*^{a-e}\right)+\lambda_*^{a-1}\right\} \\
&\lesssim \frac{\pi_i}{\gamma_*}\sum_{n-1\geq a>b\geq 2}\sum\lambda_*^{a-b}+\pi_i\left(\sum_{n-1\geq a>b\geq 2}\sum\lambda_*^{a-b}\right)\left(\sum_{e\geq 1}^{b-1}\lambda_*^{b-e}\right)+\sum_{n-1\geq a\geq 3}a^3\lambda_*^{a-1} \\
&\lesssim \frac{n\pi_i}{\gamma_*^2}+\frac{1}{\gamma_*^3}\lesssim \frac{n^2\pi_i^2}{\gamma_*^2}+\frac{1}{\gamma_*^3} \qquad (4.53)
\end{aligned}$$

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 (4.51) with $e=d$ to get

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

Summing over a,b,e and applying (4.53), we get

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

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 (4.51) with $a=b$ we

get

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

which is equivalent to

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

For the first, second and fourth terms

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

and for summing the remaining terms we use (4.53), which implies

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

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 (4.51) 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

$$\sum_{n-1\geq a>e\geq 1}\sum\mathbb{E}\left[\xi_a^2\xi_e^2|X_1=i\right] \lesssim \sum_{n-1\geq a>e\geq 1}\sum\left\{\pi_i^2+\pi_i\left(\lambda_*^{e-1}+\lambda_*^{a-e}\right)+\lambda_*^{a-1}\right\} \lesssim n^2\pi_i^2+\frac{n\pi_i}{\gamma_*}+\frac{1}{\gamma_*^2}.$$

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 (4.51) 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

$$\sum_{n-1\geq a>e\geq 1}\sum \left|\mathbb{E}\left[\xi_a\xi_e^3|X_1=i\right]\right| \lesssim \pi_i \sum_{n-1\geq a>e\geq 1}\sum \lambda_*^{a-e} + \sum_{n-1\geq a>e\geq 1}\sum \lambda_*^{a-1} \lesssim \frac{n\pi_i}{\gamma_*}+\frac{1}{\gamma_*^2}.$$

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 (4.51) 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

$$\sum_{n-1\geq a>e\geq 1}\sum \left|\mathbb{E}\left[\xi_a^3\xi_e|X_1=i\right]\right| \lesssim \sum_{n-1\geq a>e\geq 1}\sum \left\{\pi_i\left(\lambda_*^{a-e}+\lambda_*^{e-1}\right)+\lambda_*^{a-1}\right\} \lesssim \frac{n\pi_i}{\gamma_*}+\frac{1}{\gamma_*^2}.$$

Case IV: single distinct index. We specialize (4.51) 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

$$\sum_{a=1}^{n-1}\mathbb{E}\left[\xi_a^4|X_1=i\right] \lesssim n\pi_i+\frac{1}{\gamma_*}. \tag{4.54}$$

Combining (4.52)–(4.54) 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 Falahatgar et al. (2016). The authors also thank David Pollard for insightful discussions on Markov chains at the initial stages of

the project.

Bibliography

Agrawal, R. (2020). Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302.

Ahle, T. (2021). Sharp and simple bounds for the raw moments of the binomial and poisson distributions. *arXiv:2103.17027*.

Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, pages 89–110.

Anevski, D., Gill, R. D., and Zohren, S. (2017). Estimating a probability mass function with unknown labels. *The Annals of Statistics*, 45(6):2708–2735.

Atteson, K. (1999). The asymptotic redundancy of Bayes rules for Markov chains. *IEEE Transactions on Information Theory*, 45(6):2104–2109.

Bartlett, M. S. (1951). The frequency goodness of fit test for probability chains. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 47, pages 86–95. Cambridge University Press.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference: the minimum distance approach*. CRC press.

Ben-Hamou, A., Oliveira, R. I., and Peres, Y. (2018). Estimating graph parameters via random walks with restarts. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1702–1714. SIAM.

Bennett, G. K. and Martz, H. (1972). A continuous empirical bayes smoothing technique. *Biometrika*, 59(2):361–368.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463.

Billingsley, P. (1961). Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, pages 12–40.

Braess, D., Forster, J., Sauer, T., and Simon, H. U. (2002). How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution. In *Algorithmic Learning Theory*, pages 380–394. Springer.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, 2(1):113–152.

Brown, L. D., Greenshtein, E., and Ritov, Y. (2013). The poisson compound decision problem revisited. *Journal of the American Statistical Association*, 108(502):741–749.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87.

Charikar, M., Chaudhuri, S., Motwani, R., and Narasayya, V. (2000). Towards estimation error guarantees for distinct values. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 268–279. ACM.

Chen, J. (2017). Consistency of the mle under mixture models. *Statistical Science*, 32(1):47–63.

Cherapanamjeri, Y. and Bartlett, P. L. (2019). Testing symmetric Markov chains without hitting. In *Conference on Learning Theory*, pages 758–785. PMLR.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108.

Csiszár, I. (1972). A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213.

Csiszár, I. and Körner, J. (1982). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc.

Csiszár, I. and Shields, P. (2004). Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–527.

Csiszár, I. and Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28(6):1601–1619.

Daskalakis, C., Dikkala, N., and Gravin, N. (2018). Testing symmetric Markov chains from a single trajectory. In *Conference On Learning Theory*, pages 385–409. PMLR.

Davisson, L. (1973). Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795.

Davisson, L. (1983). Minimax noiseless universal coding for Markov sources. *IEEE Transactions on Information Theory*, 29(2):211–215.

Davisson, L., McEliece, R., Pursley, M., and Wallace, M. (1981). Efficient universal noiseless source codes. *IEEE Transactions on Information Theory*, 27(3):269–279.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2019). On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47.

Doob, J. (1937). R. von mises, wahrscheinlichkeit statistik und wahrheit. *Bull. Amer. Math. Soc.*, 43(12):316–317.

Edelman, D. (1988). Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics*, 16(4):1609–1622.

Efron, B. (2014). Two modeling strategies for empirical bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285.

Efron, B. (2021). Empirical bayes: Concepts and methods. http://statweb.stanford.edu/~ckirby/brad/papers/2021EB-concepts-methods.pdf.

Efron, B. and Hastie, T. (2021). *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160.

Eggleston, H. G. (1966). Convexity.

Falahatgar, M., Orlitsky, A., Pichapati, V., and Suresh, A. (2016). Learning Markov distributions: Does estimation trump compression? In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2689–2693. IEEE.

Fan, K. (1953). Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47.

Fried, S. and Wolfer, G. (2021). Identity testing of reversible Markov chains. *arXiv preprint arXiv:2105.06347*.

Guo, F. R. and Richardson, T. S. (2020). Chernoff-type concentration of empirical probabilities in relative entropy. *IEEE Transactions on Information Theory*, 67(1):549–558.

Han, Y., Jana, S., and Wu, Y. (2021). Optimal prediction of markov chains with and without spectral gap. *Advances in Neural Information Processing Systems*, 34.

Han, Y., Jiao, J., Lee, C., Weissman, T., Wu, Y., and Yu, T. (2018a). Entropy rate estimation for Markov chains with large state space. In *Advances in Neural Information Processing Systems*, pages 9781–9792. arXiv:1802.07889.

Han, Y., Jiao, J., and Weissman, T. (2018b). Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. In *Proc. 2018 Conference On Learning Theory (COLT)*, pages 3189–3221.

Hao, Y., Orlitsky, A., and Pichapati, V. (2018). On learning Markov chains. *In Advances in Neural Information Processing Systems*, pages 648–657.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320.

Hoffman, A. (1967). Three observations on nonnegative matrices. *Journal of Research of the National Bureau of Standards B*, pages 39–41.

Houwelingen, J. v. and Stijnen, T. (1983). Monotone empirical bayes estimators for the continuous one-parameter exponential family. *Statistica Neerlandica*, 37(1):29–43.

Hsu, D., Kontorovich, A., Levin, D. A., Peres, Y., Szepesvári, C., and Wolfer, G. (2019). Mixing time estimation in reversible Markov chains from a single sample path. *Annals of Applied Probability*, 29(4):2439–2480.

Jacquet, P. and Szpankowski, W. (2002). A combinatorial problem arising in information theory: Precise minimax redundancy for Markov sources. In *Mathematics and Computer Science II*, pages 311–328. Springer.

Jana, S., Polyanskiy, Y., and Wu, Y. (2020). Extrapolating the profile of a finite population. In *Conference on Learning Theory*, pages 2011–2033. PMLR.

Janson, S. (2002). On concentration of probability. *Contemporary combinatorics*, 10(3):1–9.

Jewell, N. P. (1982). Mixtures of exponential distributions. *The annals of statistics*, pages 479–484.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684.

Juditsky, A. B. and Nemirovski, A. S. (2009). Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5A):2278–2300.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. (2015). On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100.

Kamath, S. and Verdú, S. (2016). Estimation of entropy rate and Rényi entropy rate for Markov chains. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 685–689. IEEE.

Kemperman, J. (1974). On the Shannon capacity of an arbitrary channel. In *Indagationes Mathematicae (Proceedings)*, volume 77, pages 101–115. North-Holland.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Kim, A. K. (2014). Minimax bounds for estimation of normal mixtures. *bernoulli*, 20(4):1802–1818.

Laird, N. M. (1982). Empirical bayes estimates using the nonparametric maximum likelihood estimate for the priort. *Journal of Statistical Computation and Simulation*, 15(2-3):211–220.

Lambert, D. and Tierney, L. (1984). Asymptotic properties of maximum likelihood estimates in the mixed poisson model. *The Annals of Statistics*, pages 1388–1399.

Latała, R. (1997). Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513.

Lemon, G. H. and Krutchkoff, R. G. (1969). An empirical bayes smoothing technique. *Biometrika*, 56(2):361–365.

Levin, D. and Peres, Y. (2017a). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Levin, D. A. and Peres, Y. (2017b). *Markov chains and mixing times*. American Mathematical Society.

Lezaud, P. (1998). Chernoff-type bound for finite Markov chains. *Annals of Applied Probability*, 8(3):849–867.

Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11):2708–2726.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *The annals of statistics*, pages 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics*, 11(3):783–792.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika*, 34(3):259–299.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. (2020). Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9(4):813–850.

Maritz, J. (1966). Smooth empirical bayes estimation for one-parameter discrete distributions. *Biometrika*, 53(3-4):417–429.

Maritz, J. (1968). On the smooth empirical bayes approach to testing of hypotheses and the compound decision problem. *Biometrika*, 55(1):83–100.

Maritz, J. (1969). Empirical bayes estimation for the poisson distribution. *Biometrika*, 56(2):349–359.

Maritz, J. S. and Lwin, T. (2018). *Empirical bayes methods*. Chapman and Hall/CRC.

Miao, Z., Kong, W., Vinayak, R. K., Sun, W., and Han, F. (2021). Fisher-pitman permutation tests based on nonparametric poisson mixtures with application to single cell genomics. *arXiv preprint arXiv:2106.03022*.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55.

Obremski, M. and Skorski, M. (2020). Complexity of estimating Rényi entropy of Markov chains. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2264–2269.

Orlitsky, A., Santhanam, N., Viswanathan, K., and Zhang, J. (2005). Convergence of profile based estimators. In *Proc. 2005*, pages 1843–1847. IEEE.

Orlitsky, A., Santhanam, N., Viswanathan, K., and Zhang, J. (2008). On estimating the probability multiset. draft.

Paninski, L. (2004). Variational minimax estimation of discrete distributions under KL loss. *Advances in Neural Information Processing Systems*, 17:1033–1040.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, pages 1–20.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Persaud, B., Lan, B., Lyon, C., and Bhim, R. (2010). Comparison of empirical bayes and full bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention*, 42(1):38–43.

Pfanzagl, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference*, 19(2):137–158.

Polyanskiy, Y., Suresh, A. T., and Wu, Y. (2017). Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory (COLT)*, Amsterdam, Netherland. arXiv:1702.05574.

Polyanskiy, Y. and Wu, Y. (2019). Dualizing Le Cam's method, with applications to estimating the unseens. *arxiv preprint arxiv:1804.05436*.

Polyanskiy, Y. and Wu, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Polyanskiy, Y. and Wu, Y. (2021). Sharp regret bounds for empirical bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*.

Raskhodnikova, S., Ron, D., Shpilka, A., and Smith, A. (2009). Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, 30(4):629–636.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pages 131–149. University of California Press.

Robbins, H. (1956). An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Ryabko, B. (1988). Prediction of random sequences and universal coding. *Prob. Pered. Inf.*, 24(2):87–96.

Saha, S. and Guntuboyina, A. (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics*, 48(2):738–762.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Prob. Pered. Inf.*, 23:175–186.

Simar, L. (1976). Maximum likelihood estimation of a compound poisson process. *The Annals of Statistics*, 4(6):1200–1209.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879.

Szegő, G. (1939). *Orthogonal polynomials*, volume 23. American Mathematical Society.

Szpankowski, W. and Weinberger, M. J. (2012). Minimax pointwise redundancy for memoryless models over large alphabets. *IEEE transactions on information theory*, 58(7):4094–4104.

Tatwawadi, K., Jiao, J., and Weissman, T. (2018). Minimax redundancy for Markov chains with large state space. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 216–220. IEEE.

Tian, K., Kong, W., and Valiant, G. (2017). Learning populations of parameters. *Advances in neural information processing systems*, 30.

Trofimov, V. K. (1974). Redundancy of universal coding of arbitrary Markov sources. *Prob. Pered. Inf.*, 10(4):16–24.

Tsybakov, A. B. (2004). Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10).

Valiant, G. (2012). *Algorithmic Approaches to Statistical Questions*. PhD thesis, EECS Department, University of California, Berkeley.

Valiant, G. (2019). Private communication.

Valiant, G. and Valiant, P. (2011). Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 685–694.

Valiant, G. and Valiant, P. (2013). Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2157–2165.

Valiant, G. and Valiant, P. (2016). Instance optimal learning of discrete distributions. In *Proc. 48th Symp. on Th. of Comp. (STOC)*, pages 142–155, Cambridge, MA, USA.

Ver Hoef, J. M. (1996). Parametric empirical bayes methods for ecological applications. *Ecological Applications*, 6(4):1047–1055.

Vinayak, R. K., Kong, W., Valiant, G., and Kakade, S. (2019). Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pages 6448–6457.

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Walter, G. G. and Blum, J. R. (1984). A simple solution to a nonparametric máximum likelihood estimation problem. *The Annals of Statistics*, pages 372–379.

Whittle, P. (1955). Some distribution and moment formulae for the Markov chain. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):235–242.

Wolfer, G. and Kontorovich, A. (2019). Minimax learning of ergodic Markov chains. In *Algorithmic Learning Theory*, pages 904–930. PMLR.

Wolfer, G. and Kontorovich, A. (2020). Minimax testing of identity to a reference ergodic Markov chain. In *International Conference on Artificial Intelligence and Statistics*, pages 191–201. PMLR.

Wolfowitz, J. (1953). Estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, 5(1):9–23.

Wolfowitz, J. (1954). Estimation by the minimum distance method in nonparametric stochastic difference equations. *The Annals of Mathematical Statistics*, 25(2):203–217.

Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, 28(1):75–88.

Wu, Y. and Verdú, S. (2010). Functional properties of mmse. In *2010 IEEE International Symposium on Information Theory*, pages 1453–1457. IEEE.

Wu, Y. and Yang, P. (2018). Sample complexity of the distinct element problem. *Mathematical Statistics and Learning*, 1(1):37–72.

Xie, Q. and Barron, A. R. (1997). Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657.

Yang, Y. and Barron, A. R. (1999). Information-theoretic determination of minimax rates of convergence. 27(5):1564–1599.

Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435.

Zhang, C.-H. (2003). Compound decision theory and empirical bayes methods. *Annals of Statistics*, pages 379–390.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318.