Prediction from compression for models with infinite memory, with applications to hidden Markov and renewal processes

Yanjun Han, Tianze Jiang, Yihong Wu*

April 25, 2024

Abstract

Consider the problem of predicting the next symbol given a sample path of length n, whose joint distribution belongs to a distribution class that may have long-term memory. The goal is to compete with the conditional predictor that knows the true model. For both hidden Markov models (HMMs) and renewal processes, we determine the optimal prediction risk in Kullback-Leibler divergence up to universal constant factors. Extending existing results in finite-order Markov models [HJW23] and drawing ideas from universal compression, the proposed estimator has a prediction risk bounded by redundancy of the distribution class and a memory term that accounts for the long-range dependency of the model. Notably, for HMMs with bounded state and observation spaces, a polynomial-time estimator based on dynamic programming is shown to achieve the optimal prediction risk $\Theta(\frac{\log n}{n})$; prior to this work, the only known result of this type is $O(\frac{1}{\log n})$ obtained using Markov approximation [Sha+18]. Matching minimax lower bounds are obtained by making connections to redundancy and mutual information via a reduction argument.

*Yanjun Han is with the Courant Institute of Mathematical Sciences and the Center for Data Science, New York University. Email: yanjunhan@nyu.edu. Yanjun Han was generously supported by the Norbert Wiener postdoctoral fellowship in statistics at MIT IDSS and a startup grant at New York University. Tianze Jiang is with MIT EECS. Email: tjiang@mit.edu. Yihong Wu is with the Department of Statistics and Data Science, Yale University. Email: yihong.wu@yale.edu.

Contents

1 Introduction

Consider the following "ChatGPT" style of problem: Observing a sample path $X^n \triangleq (X_1, \ldots, X_n)$ of a random process, one is tasked to predict the next (unseen) symbol X_{n+1}. Mathematically, this boils down to estimating the *conditional* distribution $P_{X_{n+1}|X_1^n}$, which informs downstream tasks such as finding the top few most likely realizations in autocomplete or text generation in language models. This is a well-defined but non-standard statistical problem, in that the quantity to be estimated is random and data-dependent, unless the data are i.i.d., in which case the problem is nothing but density estimation and the optimal rate under, say, Kullback-Leibler (KL) divergence loss, is the classical "parametric rate" $\frac{k}{n}$ achieved by smoothed empirical distribution, where k and n refers to the alphabet and sample size respectively. As such, the first non-trivial instance is Markov model and of interest to applications such as natural language processing are large state spaces.

The study of this problem was initiated by [Fal+16] focusing on two-state Markov chains, who showed, via a tour-de-force argument, the surprising result that the optimal KL prediction risk is $\Theta(\frac{\log\log n}{n})$, strictly slower than the parametric rate. Their ad hoc techniques are difficult to extend to larger state space, unless extra conditions are assumed such as a large spectral gap [HOP18]. Although such mixing conditions are necessary for parameter estimation, they are not for prediction. Indeed, a chain that moves at a glacial speed is in fact easy to predict but estimating the transition probabilities is impossible. This is a significant conceptual distinction between estimation and prediction, the latter of which can be studied meaningfully assumption-free without even identifiability conditions.

Departing from conventional approaches based on concentration inequalities of Markov chains which inevitably involves mixing conditions, a strategy based on *universal compression* is proposed in [HJW21; HJW23] for prediction of Markov chains. They showed, by means of information-theoretic arguments, that the optimal prediction risk is within universal constant factors of the so-called *redundancy*, a central quantity in universal compression that measures the KL radius of the model class. Furthermore, this reduction is also *algorithmic*: if there is a computationally efficient probability assignment that achieves the redundancy, one can construct an efficient predictor with guaranteed optimality. However, their method is limited to Markov models with a finite order.

The main goal of this work is to extend these techniques based on universal compression beyond models with finite memory to those with infinite memory, in particular, *hidden Markov models* (HMMs) and *renewal processes*. Along the way, we obtain new theoretical and computational results for prediction HMM that improve the state of the art.

1.1 Main results

Let us begin with the formulation of the *prediction risk* for a general model class. For $n \in \mathbb{N} \triangleq \{1, 2, \ldots\}$, let $\mathcal{P}_{n+1}$ be a collection of joint distributions $P_{X^{n+1}}$ for $X^{n+1} \triangleq (X_1, \ldots, X_{n+1})$, where each observation X_t takes values in some space $\mathcal{X}$. The prediction risk of the next unseen symbol X_{n+1} based on the trajectory $X_1, \ldots, X_n$ is the average KL risk of estimating the (random, data-dependent) distribution $P_{X_{n+1}|X^n}$. Any such estimator can be written as a conditional distribution $Q_{X_{n+1}|X^n}$, whose worst-case prediction risk over the model class is

$$\mathsf{Risk}(Q_{X_{n+1}|X^n}; \mathcal{P}_{n+1}) \triangleq \sup_{P_{X^{n+1}} \in \mathcal{P}_{n+1}} \mathbb{E}_{X^{n+1} \sim P_{X^{n+1}}} \left[\mathrm{KL}(P_{X_{n+1}|X^n} \| Q_{X_{n+1}|X^n})\right] \tag{1}$$

where the KL divergence is $\mathrm{KL}(P\|Q) = \mathbb{E}_P[\log \frac{dP}{dQ}]$ if $P \ll Q$ and ∞ otherwise. The minimax prediction risk is then defined as

$$\mathsf{Risk}(\mathcal{P}_{n+1}) \triangleq \inf_{Q_{X_{n+1}|X^n}} \mathsf{Risk}(Q_{X_{n+1}|X^n}; \mathcal{P}_{n+1}), \tag{2}$$

As exemplary applications, we consider two model classes with infinite memory: HMMs and renewal processes. Relevant notations are deferred to Appendix A.

Hidden Markov Models A hidden Markov process is obtained by passing a Markov process through a memoryless noisy channel. It provides a useful tool for modeling practical data such as natural language and speech signals. Specifically, fix $k, \ell \in \mathbb{N}$. Let $\{Z_t : t \geq 1\}$ be a stationary Markov chain on the state space $[k] \triangleq \{1, \ldots, k\}$ with transition matrix M, which is a $k \times k$ row-stochastic matrix. Let T denote a probability transition kernel from $[k]$ to $[\ell]$, that is, a $k \times \ell$ row-stochastic matrix. Let $\{X_t : t \geq 1\}$ be an $[\ell]$-valued process such that for any n, $P_{X^n|Z^n} = \prod_{t=1}^n T(x_t|z_t)$. We refer to $\{X_t\}$ as a hidden Markov process with *transition probabilities* M and *emission probabilities* T, while $\{Z_t\}$ are called the *hidden (or latent) states*.

Let $\mathcal{P}_n^{\mathsf{HMM}}(k, \ell)$ denote the collection of joint distributions of hidden Markov processes of length $n+1$ with state space $[k]$ and observation space $[\ell]$. This is a finite-dimensional parametric model (by M and T) with a total of $k(k-1) + k(\ell-1)$ parameters. We note that over this class the parameters M and T are *not* identifiable since no further conditions such as full rank of M are assumed cf. [AHL16, Example 1]. Yet, the prediction problem is both well-defined and non-trivial. We define the optimal prediction risk of HMM as follows:

$$\mathsf{Risk}_{\mathsf{HMM}}(n, k, \ell) \triangleq \mathsf{Risk}(\mathcal{P}_n^{\mathsf{HMM}}(k, \ell))$$

Our main results on predicting hidden Markov processes are as follows.

Theorem 1 (Optimal prediction risk for HMM)**.** *The following holds:*

- *There exists a universal constant C such that for all $n \geq Ck(k+\ell)$,*

$$\mathsf{Risk}_{\mathrm{HMM}}(n, k, \ell) \leq C \left(\frac{k\ell}{n} \log \frac{n}{k\ell} + \frac{k^2}{n} \log \frac{n}{k^2} \right), \tag{3}$$

  achieved by an $n^{O(k^2+k\ell)}$-time algorithm.

- *Conversely, if either $\ell \geq k$ and $n \geq k\ell$, or $n \geq k^C$ and $k, \ell \geq 2$, then*

$$\mathsf{Risk}_{\mathrm{HMM}}(n, k, \ell) \geq C^{-1} \left(\frac{k\ell}{n} \log \frac{n}{k\ell} + \frac{k^2}{n} \log \frac{n}{k^2} \right). \tag{4}$$

We note that before this work even for the simplest case of binary-state binary-emission HMMs, the best known result is $O(\frac{1}{\log n})$ by [Sha+18], who considered prediction in HMMs under a somewhat different formulation than (2) (with further averaging over the sample size n and in the weaker total variation loss than KL; see Appendix A.1 for a detailed comparison). In comparison, Theorem 1 shows that for $k, \ell = O(1)$, the optimal rate in KL is $\Theta(\frac{\log n}{n})$ and attainable in polynomial time. Furthermore, we point out that our results, both for the lower bound and the upper bound, can be extended to HMMs with discrete state space but arbitrary observation space (Corollary 1 and 2). For instance, as a side result, we determine optimal prediction risks for HMM with Gaussian emissions in terms of the output dimension (Remark 1).

Since HMM has infinite memory, a natural idea is to first approximate it by a finite-order Markov chain then invoke existing prediction risk bounds for Markov models; this was the key insight in [Sha+18]. However, this approach based on Markov approximation does not achieve the optimal risk bound. Indeed, it was shown in [HJW23] that the optimal prediction risk for order-d Markov chains on $[\ell]$ scales as $\Theta(\frac{\ell^d}{n}\log\frac{n}{\ell^d})$, already much larger than the risk in Theorem 1 for moderate d. Instead, our approach in Section 2 applies ideas from universal compression, in particular, the redundancy to control the complexity of HMMs, while introducing an additional memory term to handle the long-range dependence of the HMM. The overall algorithm is based on dynamic programming that averages state sequences of length n.

On the other hand, for large k or ℓ, the statistically optimal algorithm in Theorem 1 based on dynamic programming is no longer efficient. Next we give a polynomial-time algorithm that achieves a prediction risk vanishing at a suboptimal rate. In contrast to Theorem 1, this efficient algorithm is built upon an order-$O(\log n)$ Markov approximation.

Theorem 2 (Computationally efficient algorithms)**.** *There exists a polynomial-time estimator whose* KL *prediction risk over* $\mathcal{P}_n^{\mathsf{HMM}}(k,\ell)$ *is* $O(\frac{\log k\log\ell}{\log n})$*, provided that* $\log k\log\ell = o(\log n)$*.*

A similar rate was also established in average TV loss using Markov approximation in [Sha+18]. However, their result uses empirical averages to estimate Markov transitions and applies martingale concentration results, making it hard to generalize to e.g. KL. Here, our result applies a much simpler approach via redundancy of the "add-one" code whose KL risk can be controlled.

We also complement the above upper bound with computational lower bounds in HMMs, showing that the prediction risk for any $\mathrm{poly}(n)$-time algorithm is $\Omega(1/\log\log n)$ if $\log(k\ell) = \Omega(\log n)$.

Theorem 3 (Informal: Computational lower bounds)**.** *The following holds under certain cryptographic hardness assumptions:*

1. *For any* $\varepsilon > 0$*,* $k \geq \log^{1+\varepsilon} n$*, no* $\mathrm{poly}(n)$ *algorithm achieves* $o(\frac{\log k}{\log n\log\log n})$ *risk for* $\ell \geq 2$*.*
2. *For every* $\alpha > 0$ *there exists* $k_\alpha \geq 2$*, such that if* $k \geq k_\alpha$ *and* $\ell \geq n^\alpha$*, no* $\mathrm{poly}(n)$ *algorithm can achieve* $o(1)$ *risk.*

Our lower bounds are proven by showing that certain cryptographic structures can be embedded into an HMM with a limited number of states or emission space. Such embedding was studied extensively in prior works (e.g. [MR05; Sha+18]).

Renewal processes As another application of our techniques, we turn to the class of renewal processes. A natural example of predicting a renewal process may be described as follows: Suppose that for a given driver the time (in days) between consecutive traffic accidents are random and i.i.d. Given the driving records (safety or accident) for the past n days, the insurance company seeks to predict the probability of an accident occurring on the next day, where the interarrival distribution is unknown.

To give a formal definition of a renewal process, let $T_0, T_1, T_2, \ldots$ denote a sequence of independent $\mathbb{N}$-valued random variables, where T_i are iid drawn from some distribution μ with a finite mean. A *renewal process* $\{X_t : t \geq 1\}$ is binary valued such that $\{t : X_t = 1\}$ is exactly $\{T_0, T_0+T_1, T_0+T_1+T_2, \ldots\}$. We refer to T_0 and $\{T_i : t \geq 1\}$ as the initial wait time and the interarrival times. It is known ([CS96]) that $\{X_t\}$ is stationary if and only if T_0 is distributed as $\mathbb{P}(T_0 = t) = \frac{1}{\mathbb{E}_\mu[T_1]}\sum_{s\geq t}\mu(s), t \in \mathbb{N}$.

Let $\mathcal{P}_n^{\mathsf{rnwl}}$ denote the collection of joint distributions of a stationary renewal process of length $n+1$ with a finite expected interarrival time. In contrast to the previously considered HMM, this

is a nonparametric (infinite-dimensional) model parameterized by the interarrival time distribution μ. Particularizing (2), define the optimal prediction risk as $\mathsf{Risk}_{\mathsf{rnwl}}(n) \triangleq \mathsf{Risk}(\mathcal{P}_n^{\mathsf{rnwl}})$. The following result determines its sharp rate:

Theorem 4 (Prediction of renewal processes)**.** *There exists an absolute constant C, such that*

$$C^{-1}\sqrt{n^{-1}} \le \mathsf{Risk}_{\mathsf{rnwl}}(n) \le C\sqrt{n^{-1}}.$$

The proof of this result builds upon the redundancy bound $\Theta(\sqrt{n})$ ([CS96; FS02]) for renewal processes. Both a strength and a weakness of our theory, the predictor attaining the optimal rate of $1/\sqrt{n}$ is not computationally efficient (see Section 3.4) and it is unclear how to do so in polynomial time. One idea may be the following. For an oracle who knows the true interarrival distribution μ, it can determine the true $P_{X_{n+1}|X^n}$ by the *hazard rate*:

$$P_{X_{n+1}=1|X^n} = \mu(\tau+1)/\sum_{t>\tau}\mu(t), \tag{5}$$

where τ is the time till the most recent renewal (or the origin of time). Thus a natural idea is to replace μ by its empirical version if sufficiently many renewals are observed, and predict $P_{X_{n+1}=1|X^n}$ by some small probability, e.g. $\frac{1}{\mathrm{poly}(n)}$, otherwise. The analysis of this algorithm, however, appears challenging absent assumptions on the distribution μ.

1.2 Related works

The connections between compression and prediction are long studied dating back to e.g. [Ris84; FMG92; HKW98]. More recently, a line of works ([Fal+16; HJW23]) determined the optimal prediction risk in KL for Markov models up to constants and showed that it is near the "parametric rate" $\frac{K}{n}$, where K is the number of model parameters, but is strictly slower by logarithmic factors. A key assumption of these results is the *finite memory* of the true model, where the next observation may depend on only the most recent few.

Turning to HMMs, the majority of works in the statistical learning literature focus on identifiability ([AHL16; Hua+15]) and parameter estimation, using algorithms include moments or tensor methods ([DGL17; Ana+14; Sha+17; AGN22]) and penalized likelihood ([DGL16] [Leh21]). However, the success of those methods routinely requires extra assumptions on parameters such as spectral properties ([Hua+15; AGN22]) and sparsity ([Sha+17]). For prediction, we need not and do not impose these assumptions. In terms of prediction, the closest work we are aware of is [Sha+18], where the authors focus on algorithms via Markov approximation. Finally, computational barriers (of various forms) are known to exist for both prediction and estimation of HMMs ([MR05; Sha+18]).

2 Prediction risk bound based on universal compression

Having defined the prediction risk (2) for a general model class $\mathcal{P}_{n+1}$, we introduce the closely related redundancy problem which is at the heart of both theory theory and algorithms for universal compression. The *redundancy* of a joint distribution $Q_{X^{n+1}}$ (often referred to as a *probability assignment*) is defined as the worst-case KL risk of fitting the joint distribution of X^n, namely

$$\mathsf{Red}(Q_{X^{n+1}}; \mathcal{P}_{n+1}) \triangleq \sup_{P_{X^{n+1}} \in \mathcal{P}_{n+1}} \mathrm{KL}(P_{X^{n+1}} \| Q_{X^{n+1}}). \tag{6}$$

Optimizing over the probability assignment $Q_{X^{n+1}}$, the minimax redundancy is defined as

$$\mathsf{Red}(\mathcal{P}_{n+1}) \triangleq \inf_{Q_{X^{n+1}}} \mathsf{Red}(Q_{X^{n+1}}; \mathcal{P}_{n+1}), \tag{7}$$

The role of a probability assignment in universal compression is a simultaneous approximation to a class of models. It is known that the Shannon entropy $H(P_{X^{n+1}})$ is within 1 bit of the best average code length for the optimal compressor that knows the source distribution $P_{X^{n+1}}$. The goal of universal compression is to design a compressor that simultaneously approaches the entropy for a class of models. This can be achieved by applying the compressor (e.g. arithmetic coding) designed for a probability assignment $Q_{X^{n+1}}$, whose excess code length over $H(P_{X^{n+1}})$ is at most within 1 bit of $\mathsf{Red}(Q_{X^{n+1}})$ for all $P_{X^{n+1}}$ in the class P_{n+1}. Thanks to this reduction, the design of universal compressor is largely reduced to choosing a good probability assignment and the redundancy is the central quantity in universal compression.

The following result relates the redundancy and the prediction risk for any stationary data-generating process. In the case of i.i.d. models, this type of reduction relating cumulative risks and individual risks is known as *online-to-batch conversion* which, in the present context, dates back at least to [YB99] for density estimation (see e.g. [PW24, Proposition 32.7] for a summary).

Proposition 1 (Upper bound prediction risk by redundancy)**.** *Suppose that each $P_{X^{n+1}} \in \mathcal{P}_{n+1}$ is stationary, that is, $P_{X_{t_1},\ldots,X_{t_k}} = P_{X_{t_1+t},\ldots,X_{t_k+t}}$ for any shift $t \geq 1$ and $1 \leq t_1, \ldots, t_k \leq n+1-t$. Let $Q_{X^{n+1}}$ be an arbitrary joint distribution factorizing as $Q_{X^{n+1}} = \prod_{t=1}^{n+1} Q_{X_t|X^{t-1}}$. Consider an estimator $\widetilde{Q}_{X_{n+1}|X^n}$ defined as*

$$\widetilde{Q}_{X_{n+1}|X^n}(\cdot|x^n) \triangleq \frac{1}{n}\sum_{t=1}^{n} Q_{X_{t+1}|X^t}(\cdot|x_{n-t+1}^n) \tag{8}$$

Then

$$\mathsf{Risk}(\widetilde{Q}_{X_{n+1}|X^n}; \mathcal{P}_{n+1}) \leq \frac{1}{n}\mathsf{Red}(Q_{X^{n+1}}; \mathcal{P}_{n+1}) + \frac{1}{n}\sum_{t=1}^{n} I(X_{n+1}; X^{n-t}|X_{n-t+1}^n). \tag{9}$$

Since the last term in (9) does not depend on the probability assignment Q, taking the supremum over the worst-case P in class then optimizing over Q yields

$$\mathsf{Risk}(\mathcal{P}_{n+1}) \leq \frac{1}{n}\mathsf{Red}(\mathcal{P}_{n+1}) + \mathsf{mem}(\mathcal{P}_{n+1}) \tag{10}$$

where the residual term

$$\mathsf{mem}(\mathcal{P}_{n+1}) \triangleq \sup_{P_{X^{n+1}} \in \mathcal{P}_{n+1}} \frac{1}{n}\sum_{t=1}^{n} I(X_{n+1}; X^{n-t}|X_{n-t+1}^n) \tag{11}$$

measures the memory, in a average sense, of the data-generating process in the model class. Indeed, recall that the conditional mutual information $I(A;B|C)$ measures the conditional dependency between A and B given C, and is zero if they are conditional independent. Thus, for Markov models of order m, $I(X_{n+1}; X^{n-t}|X_{n-t+1}^n) = 0$ for all $t \geq m$ and $\mathsf{mem}(\mathcal{P}_{n+1})$ is at most $O(\frac{mH(X_{n+1})}{n})$. As a result, for bounded m we get $\mathsf{Risk} \lesssim \frac{\mathsf{Red}}{n}$.[1] For models with infinite memory, such as HMMs and renewal processes, applying this redundancy-based risk bound requires bounding the memory term uniformly, which we carry out in the subsequent sections.

[1] In fact, a slightly different argument in [HJW21, Lemma 6] avoids the additive error term and shows $\mathsf{Risk}(\mathcal{P}_{n+1}) \leq \frac{1}{n+1-m}\mathsf{Red}(\mathcal{P}_{n+1})$ for mth-order Markov models.

We end this section with a couple of remarks. First, applying the risk bound in Proposition 1 relies on bounding the redundancy of a model class from above, which is often achieved by further relaxing the redundancy, an approach known as *individual sequences*. Replacing the expectation in $\mathrm{KL}(P_{X^{n+1}}\|Q_{X^{n+1}}) = \mathbb{E}_P\left[\log \frac{P_{X^{n+1}}}{Q_{X^{n+1}}}\right]$ by the maximum, one arrives at the so-called *minimax pointwise redundancy*

$$\mathsf{Red}(\mathcal{P}_{n+1}) \le \widetilde{\mathsf{Red}}(\mathcal{P}_{n+1}) \triangleq \inf_{Q_{X^{n+1}}} \sup_{P_{X^{n+1}}\in\mathcal{P}_{n+1}} \max_{x^{n+1}\in\mathcal{X}^{n+1}} \log \frac{P_{X^{n+1}}(x^{n+1})}{Q_{X^{n+1}}(x^{n+1})}. \tag{12}$$

The optimal probability assignment for (12) is known as Shtarkov's *normalized maximum likelihood* assignment $Q^*_{X^{n+1}}(x^{n+1}) \propto \sup_{P_{X^{n+1}}\in\mathcal{P}_{n+1}} P_{X^{n+1}}(x^{n+1})$, leading to the following formula for the minimax pointwise redundancy as a *Shtarkov sum*

$$\widetilde{\mathsf{Red}}(\mathcal{P}_{n+1}) = \log \sum_{x^{n+1}\in\mathcal{X}^{n+1}} \sup_{P_{X^{n+1}}\in\mathcal{P}_{n+1}} P_{X^{n+1}}(x^{n+1}). \tag{13}$$

Most redundancy bounds, including those that we apply ([CS96]), are obtained by either analyzing the pointwise redundancy or directly bounding the above Shtarkov sum. This combinatorial approach avoids all probabilistic computation and is essentially the reason why one can sidestep mixing conditions in HMMs.

Second, for i.i.d. models, say, distributions over k elements, the upper bound on the prediction risk in Proposition 1 is in fact loose by a logarithmic factor, since we know that $\mathsf{Risk} \asymp \frac{k-1}{n}$ and $\mathsf{Red} \asymp (k-1)\log n$. Interestingly, the compression-prediction method seems particularly effective for models with memory, which is tight up to *constant factors* for finite-order Markov chains ([HJW23]) and, as we show in this paper, HMMs and renewal processes. Complementing Proposition 1, we give a reduction argument that shows the prediction risk of a given class of HMMs is lower bounded by the redundancy of a slightly smaller subclass – see Section 4.1 for details.

3 Proof of the upper bounds

In this section, we make use of Proposition 1 to upper bound the prediction risk for HMMs and renewal processes. This entails upper bounding the minimax redundancy $\mathsf{Red}(\mathcal{P})$ in (7) and the memory term $\mathsf{mem}(\mathcal{P})$ in (11), for both HMMs and renewal processes.

3.1 Bounding the memory term for HMMs

We start with a simple upper bound on the memory term in (10) for HMMs. Similar bounds have appeared previously in the literature, see, e.g., [Bir62, p. 932].

Proposition 2. *Let $\{X_t\}$ be a stationary hidden Markov process. Then*

$$\sum_{t=1}^{n} I(X^{n-t}; X_{n+1}|X_{n-t+1}^{n}) \le I(Z_1; X^{n+1}).$$

Suppose there are at most k latent states. Then $I(Z_1; X^{n+1}) \le H(Z_1) \le \log k$ regardless of the emissions. Applying Proposition 2 to (9) yields

$$\mathsf{Risk}(\widetilde{Q}_{X_{n+1}|X^n}) \le \frac{1}{n}\mathsf{Red}(Q_{X^{n+1}}) + \frac{\log k}{n}, \tag{14}$$

where $Q_{X^{n+1}}$ is any probability assignment and $\widetilde{Q}_{X_{n+1}|X^n}$ is the predictor defined in (8). As we show next, the memory term turns out to be negligible compared with the redundancy.

3.2 Redundancy bound for HMM

Next we upper bound the redundancy $\mathsf{Red}(\mathcal{P}_n^{\mathsf{HMM}}(k,\ell))$ and prove the upper bound in Theorem 1. To this end, it suffices to bound the redundancy of the joint state-emission sequence. Indeed, by definition (7), for any $Q_{X^{n+1},Z^{n+1}}$ and any $P_{X^{n+1},Z^{n+1}}$ in the model class, we have $\mathrm{KL}(P_{X^{n+1},Z^{n+1}}\|Q_{X^{n+1},Z^{n+1}}) \le \mathrm{KL}(P_{X^{n+1}}\|Q_{X^{n+1}})$. Let us define a joint probability assignment by separately approximating the transition and emission probabilities using the probability assignment designed for the Markov and i.i.d. class respectively:

$$\begin{aligned} Q_{X^{n+1},Z^{n+1}}(x^{n+1},z^{n+1}) &= Q_{Z^{n+1}}(z^{n+1}) \cdot Q_{X^{n+1}|Z^{n+1}}(x^{n+1}|z^{n+1}) \\ &= \frac{1}{k}\prod_{t=1}^{n} M_t(z_{t+1}|z_t) \cdot \prod_{t=1}^{n+1} T_t(x_t|z_t), \end{aligned} \tag{15}$$

where M_t and T_t are the add-one estimators ([KT81]) for the transition and emission probabilities, respectively:

$$M_t(z'|z) = \frac{1+\sum_{i=1}^{t-1}\mathbb{1}_{z_{i+1}=z' \text{ and } z_i=z}}{k+\sum_{i=1}^{t-1}\mathbb{1}_{z_i=z}}, \tag{16}$$

$$T_t(x|z) = \frac{1+\sum_{i=1}^{t-1}\mathbb{1}_{z_i=z \text{ and } x_i=x}}{\ell+\sum_{i=1}^{t-1}\mathbb{1}_{z_i=z}}. \tag{17}$$

Finally, let $Q_{X^{n+1}}$ be the marginal of (15). The following result bounds on the pointwise redundancy of $Q_{X^{n+1},Z^{n+1}}$ and thus that of $Q_{X^{n+1}}$. By (12), this also bounds their average-case redundancy.

Proposition 3. *Let $n \ge k(k+\ell)$. For any hidden transition matrix M (with stationary distribution π) on state space $[k]$ and emission matrix T from $[k]$ to $[\ell]$,*

$$\max_{x^{n+1},z^{n+1}} \log \frac{\pi(z_1)\prod_{t=1}^{n} M(z_{t+1}|z_t)\prod_{t=1}^{n+1} T(x_t|z_t)}{Q_{X^{n+1},Z^{n+1}}(x^{n+1},z^{n+1})} \lesssim k^2 \log\frac{n}{k^2} + k\ell\log\frac{n}{k\ell}.$$

Consequently, $\mathsf{Red}(Q_{X^{n+1}};\mathcal{P}_n^{\mathsf{HMM}}(k,\ell)) \lesssim k^2\log(n/k^2) + k\ell\log(n/k\ell)$.

Combining Propositions 1, 2, and 3, we have

$$\mathsf{Risk}(\mathcal{P}_n^{\mathsf{HMM}}(k,\ell)) \lesssim \frac{k^2}{n}\log\frac{n}{k^2} + \frac{k\ell}{n}\log\frac{n}{k\ell} + \frac{\log k}{n},$$

which completes the upper bound proof of Theorem 1. In fact, the same program can be extended to HMMs with general emissions. To this end, let X take value in a general space $\mathcal{X}$, and $\mathcal{Q}$ be a class of probability distributions over $\mathcal{X}$. We use $\mathcal{P}_n^{\mathsf{HMM}}(k,\mathcal{Q})$ to denote the collection of stationary HMMs of length $n+1$, with hidden states in $[k]$ and emissions in $\mathcal{Q}$ (i.e. $P_{X|Z}(\cdot|z) \in \mathcal{Q}$ for all $z \in [k]$). The following corollary, proved in Appendix B.4, bounds the prediction risk:

Corollary 1. *Suppose $\mathsf{Red}(\mathcal{Q}^{\otimes t}) \le R(t)$ for all t for some concave $R(\cdot)$. Then for $n \gtrsim k^2$,*

$$\mathsf{Risk}(\mathcal{P}_n^{\mathsf{HMM}}(k,\mathcal{Q})) \lesssim \frac{k^2}{n}\log\frac{n}{k^2} + \frac{k}{n}R\left(\frac{n+1}{k}\right).$$

Remark 1. *When $\mathcal{Q}$ is the set of Gaussian distributions $\mathcal{N}(w,I_d)$ with $w \in [-1,1]^d$, one has $\mathsf{Red}(\mathcal{Q}^{\otimes n}) \lesssim d\log n$ by Gaussian channel capacity. Hence, Corollary 1 shows that the optimal prediction risk for Gaussian HMM with k hidden states is $O(\frac{k(k+d)}{n}\log n)$. Furthermore, as we will see in the next section (Corollary 2), this bound is tight.*

3.3 An optimal prediction algorithm

We show that the estimator in Theorem 1 can be computed in time $n^{O(k^2+k\ell)}$, and it suffices to prove that the marginal distribution $Q_{X^{n+1}}$ can be efficiently computed based on the joint distribution $Q_{X^{n+1},Z^{n+1}}$ in (15). Our idea relies on an equivalent expression of $Q_{X^{n+1},Z^{n+1}}$ via sufficient statistics: let $M \in \mathbb{R}^{k\times k}, T \in \mathbb{R}^{k\times \ell}$ be the transition and emission count matrices, formally defined as $M_{z,z'} = \sum_{t=1}^{n} \mathbb{1}_{z_t=z,z_{t+1}=z'}$ and $T_{z,x} = \sum_{t=1}^{n+1} \mathbb{1}_{z_t=z,x_t=x}$, then

$$Q_{X^{n+1},Z^{n+1}}(x^{n+1}, z^{n+1}) = \frac{1}{k} \prod_{z\in[k]} \left(\frac{\prod_{z'\in[k]} M_{z,z'}!}{k^{\overline{M_z}}} \cdot \frac{\prod_{x\in[\ell]} T_{z,x}!}{\ell^{\overline{T_z}}} \right) \triangleq F(M,T),$$

where M_z, T_z denotes the row sums of M, T, and $k^{\overline{m}} = k(k+1)\cdots(k+m-1)$ is the rising factorial. Based on the above expression, to compute the marginal distribution $Q_{X^{n+1}}$ it suffices to enumerate over all possible matrices (M,T) and compute the number $\mathcal{A}(M,T;x^{n+1})$ of sequences z^{n+1} that induce a given (M,T). The following lemma, proved in Appendix D.1, shows that for each (M,T) this enumeration can be done in $n^{O(k^2+k\ell)}$ time by dynamic programming.

Lemma 1. *Given any sequence x^{n+1}, the count $\mathcal{A}(M,T;x^{n+1})$ can be computed in time $n^{O(k^2+k\ell)}$.*

Since the entries of (M,T) take values in $\{0,1,\cdots,n+1\}$, the number of all possible matrices is $n^{O(k^2+k\ell)}$. This completes the proof of the computational upper bound in Theorem 1.

3.4 Renewal processes

For the class $\mathcal{P}_n^{\mathsf{rnwl}}$ of renewal processes defined in Section 1.1, we invoke a well-known result on its redundancy:

Lemma 2 ([CS96])**.** $\mathsf{Red}(\mathcal{P}_n^{\mathsf{rnwl}}) = \Theta(\sqrt{n})$.

By Proposition 1 and Lemma 2, it remains to upper bound the memory term $\mathsf{mem}(\mathcal{P}_n^{\mathsf{rnwl}})$. A stationary renewal process $\{X_t\}$ with interarrival distribution μ is represented by a stationary HMM with a countably infinite state space as follows:

1. The hidden states $\{Z_t\}$ takes values in $\mathbb{N}$ represents the "countdown" until the next renewal, where $\mathbb{P}(Z_{t+1} = i-1 | Z_t = i) = 1$ if $i \geq 2$ and $\mathbb{P}(Z_{t+1} = j | Z_t = 1) = \mu(j)$ for $j \geq 1$.
2. The emissions is binary and deterministic: $X_t = \mathbb{1}_{Z_t=1}$.

Furthermore, the stationary state distribution π_μ is precisely the law of the initial wait time in Section 1.1, given by $\pi_\mu(t) = \frac{\sum_{i\geq t}\mu(i)}{\sum_{i\geq 1} i\mu(i)}$. This HMM representation allows us to apply Proposition 2 to bound the memory term

$$\mathsf{mem}(\mathcal{P}_n^{\mathsf{rnwl}}) \leq \frac{1}{n} I(Z_1; X^{n+1}).$$

Although Z_1 takes infinitely many values, we show that the above mutual information is still at most $O(\log n)$: Let $\widetilde{Z}_1 \triangleq \min\{Z_1, n+2\}$. Then $Z_1 \to \widetilde{Z}_1 \to X^{n+1}$ is a Markov chain because $\widetilde{Z}_1 = Z_1$ if $\widetilde{Z}_1 < n+2$, and $X^{n+1} = 0^{n+1}$ if $\widetilde{Z}_1 = n+2$. Therefore, by the data processing inequality:

$$I(Z_1; X^{n+1}) \leq I(\widetilde{Z}_1; X^{n+1}) \leq H(\widetilde{Z}_1) \leq \log(n+2).$$

The upper bound of Theorem 4 then follows from Proposition 1 and Lemma 2.

Note that the redundancy upper bound in Lemma 2 is obtained by analyzing the pointwise redundancy (12) and bounding the Shtarkov sum (13) by the partition number whose asymptotics yields the $\sqrt{n}$ term ([HR18]). Thanks to Proposition 1, averaging of the conditionals of the Shtarkov distribution (normalized maximum likelihood) yields a predictor that attains the optimal rate $\frac{1}{\sqrt{n}}$. As discussed in Section 1.1, finding a computationally efficient optimal predictor is an interesting open question.

4 Proof of the lower bounds

This section proves the lower bounds of the prediction risk for HMMs with further technical results deferred till Appendix C. We first present a generic embedding idea to lower bound the prediction risk using the redundancy of a slightly smaller class of HMMs; this reduction essentially shows the tightness of the compression-prediction program in Section 2 when a (hidden) Markov structure is available. Next we lower bound the redundancy $\mathsf{Red}(\mathcal{P}_n^{\mathsf{HMM}}(k,\ell))$. For renewal processes we use an explicit prior and lower bound the Bayes prediction risk directly (see Appendix E).

4.1 Reduction from redundancy to prediction risk

Complementing the upper bound in Proposition 1, the following result lower bounds the prediction risk by the redundancy of HMMs with one fewer states and observations.

Proposition 4 (Lower bound prediction risk by redundancy)**.** *For $\mathcal{P}_n^{\mathsf{HMM}}(k,\ell)$ with $k \geq 2, \ell \geq 3$,*

$$\mathsf{Risk}_{\mathrm{HMM}}(n,k,\ell) \gtrsim \frac{1}{n}(\mathsf{Red}(\mathcal{P}_{n+1}^{\mathsf{HMM}}(k-1,\ell-1)) - \log \ell).$$

For $k \geq 2, \ell \geq 3$, combining this result with the redundancy lower bound in Theorem 5 proves the lower bound in Theorem 1. (Note that for $k = 2$, $\mathcal{P}_{n+1}^{\mathsf{HMM}}(k-1,\ell-1)$ is in fact an i.i.d. process over $[\ell]$ and has redundancy $\Theta(\ell \log(n/\ell))$ for $n \gtrsim \ell$ [Dav73]).

The proof of Proposition 4 relies on a reduction from redundancy to prediction risk. Given an arbitrary instance Q of the HMM parameters with hidden alphabet $[k-1]$ and emission $[\ell-1]$, we seek to construct another instance P for the HMM parameters with hidden alphabet $[k]$ and emission $[\ell]$. The main idea is to add a "lazy" state k that slows down the chain. This uninformative state has a heavy self loop such that with constant probability, the chain only explores the original state space $[k-1]$ for a period of time that is approximately uniform in $[n]$, effectively reducing the sample size from n to $\mathrm{Unif}([n])$. As such, the prediction risk can then be related to the cumulative risk, that is, the redundancy. Specifically, define

1. Emission probabilities: $P(X = \ell | Z = k) = 1$ and $P_{X|Z=z} = Q_{X|Z=z}$ for all $z \in [k-1]$. In other words, state k always emits ℓ, while the emissions of other states are the same as Q.

2. Transition: let π_Q be the stationary distribution over the state space $[k-1]$ under Q:

$$P(Z_2 = j | Z_1 = i) = \begin{cases} \mathbb{1}_{j=k}(1-1/n) + \mathbb{1}_{j\neq k}\pi_Q(j)/n & \text{if } i = k, \\ 1/n & \text{if } i \neq k, j = k, \\ (n-1)Q(Z_2 = j | Z_1 = i)/n & \text{if } i \neq k, j \neq k. \end{cases} \tag{18}$$

One can verify that the stationary state distribution of the HMM P is $\pi_P(k) = 1/2$ and $\pi_P(i) = \pi_Q(i)/2$. For $0 \leq t \leq n-1$, define the event $E_t = \{x^n : x^t = \ell^t, x_{t+1}^n \in [\ell-1]^{n-t}\}$. A simple computation shows that $\mathbb{P}(E_t) = \Theta\left(\frac{1}{n}\right)$ for all $1 \leq t \leq n-1$, and $\mathbb{P}(E_0) = \Theta(1)$.

Next we consider a general prior distribution of Q, which induces a prior of P. Note that conditioned on the event E_t, the Bayes prediction risk of $P_{X_{n+1}|X^n}$ (or equivalently $I(P; X_{n+1}|X^n, E_t)$) equals to the Bayes prediction risk of $Q_{X_{n+1}|X_{t+1}^n}$ (or equivalently $I(Q; X_{n+1}|X_{t+1}^n)$) times $1 - 1/n$, the scaling factor between the transition probabilities under Q and P on $[k-1]$.

Therefore, the overall Bayes prediction risk of $P_{X_{n+1}|X^n}$ is lower bounded by

$$\begin{aligned}\sum_{t=0}^{n-1} \mathbb{P}(E_t) I(P; X_{n+1}|X^n, E_t) &= \sum_{t=0}^{n-1} \mathbb{P}(E_t) \cdot \left(1 - \frac{1}{n}\right) I(Q; X_{n+1}|X_{t+1}^n) \\ &\gtrsim \frac{1}{n} \sum_{t=0}^{n-1} I(Q; X_{n-t+1}|X^{n-t}) \\ &= \frac{1}{n} \left(I(Q; X^{n+1}) - I(Q; X_1)\right) \geq \frac{1}{n} \left(I(Q; X^{n+1}) - \log \ell\right).\end{aligned}$$

Maximizing over the prior distributions of Q leads to redundancy and proves Proposition 4.

The simple embedding above is a bit wasteful as it designates a special emission symbol to signify the lazy state. As such, the case of $\ell = 2$ is out of reach. Applying more delicate reductions, the next result (proved in Appendix C.1) gives a risk lower bound based on the redundancy of HMMs with the same observation space, with the additional constraint that the stationary state distribution is uniform, which, for large k, has the same redundancy within constant factors. Thus this result is applicable to the case of binary and even continuous emissions such as Gaussians (Remark 1).

Proposition 5. *In the context of Corollary 1, for all $k \geq 2$ it holds that*

$$\mathsf{Risk}(\mathcal{P}_n^{\mathsf{HMM}}(k, \mathcal{Q})) \gtrsim \frac{1}{n}\mathsf{Red}(\mathcal{P}_{n+1,\mathsf{U}}^{\mathsf{HMM}}(k-1, \mathcal{Q})) - \frac{\log(nk) + \mathsf{Red}(\mathcal{Q})}{n},$$

where $\mathcal{P}_{n,\mathsf{U}}^{\mathsf{HMM}}(k, \mathcal{Q})$ is the set of all stationary HMMs with hidden states in $[k]$, emission distributions in $\mathcal{Q}$, and a uniform stationary distribution for the hidden states.

Corollary 2. *Suppose that there are constants $0 \leq c_1 < c_2 \leq 1$ and a map $f : \mathcal{X} \to \{0, 1\}$ such that for all $c \in [c_1, c_2]$, there exists $Q_c \in \mathcal{Q}$ such that $f_\# Q_c = \mathsf{Bern}(c)$. Then*

$$\mathsf{Risk}(\mathcal{P}_n^{\mathsf{HMM}}(k, \mathcal{Q})) \gtrsim \frac{k^2 \log n}{n} + \frac{k}{n}\mathsf{Red}(\mathcal{Q}^{\otimes \lfloor n/k \rfloor})$$

as long as $k \geq C$ and $n \geq k^D$, where (C, D) are absolute constants.

4.2 Lower bounding the redundancy of HMM

The general idea of lower bounding the redundancy is via the variational representation $\mathsf{Red}(\mathcal{P}) = \sup_\mu I(\theta; X)$, where $\theta \sim \mu$ is a random element of $\mathcal{P}$ according to the distribution (prior) μ, and conditioned on this element, the random variable X follows the distribution θ. Following [Dav+81], lower bounding the mutual information $I(\theta; X)$ requires us to construct an estimator of θ that achieves small error on a sufficiently rich sub-model class (that we also need to construct), leading to high mutual information. For the challenging case of overcomplete HMM (e.g. $\ell = 2$), this estimator is based on tensor decomposition.

Large ℓ. We start with the easy case of $\ell \geq k$, where the redundancy of the HMM mainly comes from the emission probabilities. The prior distribution μ is constructed as follows: the transition of the hidden states is a deterministic cycle C_k (i.e. $1 \to 2 \to \cdots \to k \to 1$), and the emission distributions are drawn independently: $(\theta_z)_{z\in[k]} \triangleq (P_{X|Z}(\cdot|z))_{z\in[k]} \sim \mu_0^{\otimes k}$, where μ_0 is some prior distribution over $\mathcal{P}([\ell])$, the collection of all probability measures on $[\ell]$. Therefore,

$$\begin{aligned} I(\theta; X^n) = I(\theta; Z^n, X^n) - I(\theta; Z^n|X^n) &\overset{\text{(a)}}{\geq} I(\theta; X^n|Z^n) - H(Z^n) \\ &\overset{\text{(b)}}{\geq} \sum_{z\in[k]} I(\theta_z; (X_i : Z_i = z)|Z^n) - \log k \\ &\overset{\text{(c)}}{\geq} \sum_{z\in[k]} I(\theta_z; Y_z^{\lfloor n/k \rfloor}) - \log k = kI(\theta_1; Y_1^{\lfloor n/k \rfloor}) - \log k, \end{aligned}$$

where (a) is due to $I(\theta; Z^n) = 0$, (b) follows from the mutual independence of $(\theta_z, (X_i : Z_i = z))_{z\in[k]}$ given Z^n, (c) introduces an auxiliary sequence of i.i.d. random variables $Y_{z,1}, Y_{z,2}, \cdots \sim \theta_z$, and uses that the sample size of Y_z is at least $\lfloor n/k \rfloor$ for each $z \in [k]$. Consequently,

$$\mathsf{Red}(\mathcal{P}_n^{\mathsf{HMM}}(k,\ell)) = \sup_{\mu} I(\theta; X^n) \geq k \cdot \sup_{\mu_0} I(\theta_1; Y_1^{\lfloor n/k \rfloor}) - \log k \gtrsim k\ell \log \frac{n}{k\ell} - \log k$$

for $n \gtrsim k\ell$, where we use the classical redundancy bound of i.i.d. model $\mathsf{Red}(\mathcal{P}([\ell])^{\otimes m}) = \Omega(\ell \log(m/\ell))$ if $m \gtrsim \ell$ ([Dav73]).

Large k. The analysis for the overcomplete case of large k is far more challenging. Without loss of generality consider $\ell = 2$. The lower bound on mutual information crucially relies on the following lemma, which shows that an estimator on tensor decomposition succeeds provided that the transition matrices are close to a deterministic cycle.

Lemma 3 (Estimating the transition)**.** *Let $0 \leq t_1 < t_2 \leq 1$ be fixed constants. There exist positive constants c_0, c_1, c_2, c_3 and fixed $p_1, \cdots, p_k \in (t_1, t_2)$ such that when $k \geq c_0, n \geq k^{c_1}$ and:*

1. *the transition matrix Q of the hidden states is doubly stochastic and $\|Q - \mathrm{C}_k\|_{\max} \leq k^{-c_2}$;*
2. *the emission probabilities are fixed as $\mathbb{P}(X = 1|Z = i) = p_i$ for all $i = 1, 2, \cdots, k$,*

then there exists an estimator $\widehat{Q}_k(X^n)$ such that

$$\mathbb{P}_{X^n|Q}\left[\left\|\widehat{Q}_k(X^n) - Q\right\|_{\mathrm{F}}^2 \leq n^{-c_3}\right] \geq 0.99. \tag{19}$$

The constraints imposed on Q in Lemma 3 still result in a sufficiently rich model: one can show that there is a prior distribution μ supported on this set such that $h(Q) \gtrsim -k^2 \log k$ for $Q \sim \mu$, where $h(\cdot)$ is the differential entropy (Lemma 14). For $Q \sim \mu$, a direct consequence of Lemma 3 is $h(Q|\widetilde{Q}_k(X^n)) \lesssim -k^2 \log n$ (Lemma 4). Therefore, under this prior μ we have

$$I(Q; X^n) = h(Q) - h(Q|X^n) \geq h(Q) - h(Q|\widetilde{Q}_k(X^n)) \gtrsim k^2 \log n \tag{20}$$

as long as $n \geq k^D$ for a large constant $D > 0$. The above two cases lead to the following theorem.

Theorem 5 (Lower bound on the redundancy)**.** *There exist universal constants $c_0, c_1, c_2, D > 0$ such that*

$$\mathsf{Red}(\mathcal{P}_n^{\mathsf{HMM}}(k,\ell)) \geq c_0 \left(k^2 \log \frac{n}{k^2} + k\ell \log \frac{n}{k\ell}\right),$$

if either $k \geq \max\{\ell, c_1\}$ and $n > k^D$, or $n \geq c_2 k\ell$ and $\ell \geq k$.

Combined with the reduction in Section 4.1, we conclude the proof of the lower bounds.

References

[AGN22] Kweku Abraham, Elisabeth Gassiat, and Zacharie Naulet. "Fundamental limits for learning hidden Markov model parameters". In: *IEEE Transactions on Information Theory* 69.3 (2022), pp. 1777–1794.

[AHL16] Grigory Alexandrovich, Hajo Holzmann, and Anna Leister. "Nonparametric identification and maximum likelihood estimation for hidden Markov models". In: *Biometrika* 103.2 (2016), pp. 423–434.

[Ana+14] Anima Anandkumar et al. *Tensor decompositions for learning latent variable models*. 2014. arXiv: 1210.7559 [cs.LG].

[Bha+13] Aditya Bhaskara et al. "Smoothed Analysis of Tensor Decompositions". In: *CoRR* abs/1311.3651 (2013). arXiv: 1311.3651. URL: http://arxiv.org/abs/1311.3651.

[Bir62] John J. Birch. "Approximations for the Entropy for Functions of Markov Chains". In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 930–938.

[BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. "Noise-tolerant learning, the parity problem, and the statistical query model". In: *Journal of the ACM (JACM)* 50.4 (2003), pp. 506–519.

[CS96] I. Csiszár and P.C. Shields. "Redundancy rates for renewal and other processes". In: *IEEE Transactions on Information Theory* 42.6 (1996), pp. 2065–2072. DOI: 10.1109/18.556596.

[Dav+81] L. Davisson et al. "Efficient universal noiseless source codes". In: *IEEE Transactions on Information Theory* 27.3 (1981), pp. 269–279. DOI: 10.1109/TIT.1981.1056355.

[Dav73] L. Davisson. "Universal noiseless coding". In: *IEEE Transactions on Information Theory* 19.6 (1973), pp. 783–795. DOI: 10.1109/TIT.1973.1055092.

[DGL16] Yohann De Castro, Elisabeth Gassiat, and Claire Lacour. "Minimax Adaptive Estimation of Nonparametric Hidden Markov Models". In: *Journal of Machine Learning Research* 17.111 (2016), pp. 1–43. URL: http://jmlr.org/papers/v17/15-381.html.

[DGL17] Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. "Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models". In: *IEEE Transactions on Information Theory* 63.8 (2017), pp. 4758–4777.

[Fal+16] Moein Falahatgar et al. "Learning Markov distributions: Does estimation trump compression?" In: *2016 IEEE International Symposium on Information Theory (ISIT)*. 2016, pp. 2689–2693. DOI: 10.1109/ISIT.2016.7541787.

[FMG92] Meir Feder, Neri Merhav, and Michael Gutman. "Universal prediction of individual sequences". In: *IEEE transactions on Information Theory* 38.4 (1992), pp. 1258–1270.

[FPV15] Vitaly Feldman, Will Perkins, and Santosh Vempala. "On the complexity of random satisfiability problems with planted solutions". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*. 2015, pp. 77–86.

[FS02] P. Flajolet and W. Szpankowski. "Analytic variations on redundancy rates of renewal processes". In: *IEEE Transactions on Information Theory* 48.11 (2002), pp. 2911–2921. DOI: 10.1109/TIT.2002.804115.

[Gas18] Élisabeth Gassiat. *Universal Coding and Order Identification by Model Selection Methods*. Springer, 2018.

[HJW21] Yanjun Han, Soham Jana, and Yihong Wu. "Optimal prediction of Markov chains with and without spectral gap". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11233–11246.

[HJW23] Yanjun Han, Soham Jana, and Yihong Wu. "Optimal prediction of Markov chains with and without spectral gap". In: *IEEE Transactions on Information Theory* 69.6 (2023), pp. 3920–3959.

[HKW98] David Haussler, Jyrki Kivinen, and Manfred K Warmuth. "Sequential prediction of individual sequences under general loss functions". In: *IEEE Transactions on Information Theory* 44.5 (1998), pp. 1906–1925.

[HOP18] Yi Hao, A. Orlitsky, and V. Pichapati. "On learning Markov chains". In: *In Advances in Neural Information Processing Systems* (2018), pp. 648–657.

[HR18] Godfrey H Hardy and Srinivasa Ramanujan. "Asymptotic formulaæ in combinatory analysis". In: *Proceedings of the London Mathematical Society* 2.1 (1918), pp. 75–115.

[Hua+15] Qingqing Huang et al. "Minimal realization problems for hidden markov models". In: *IEEE Transactions on Signal Processing* 64.7 (2015), pp. 1896–1904.

[Kot+17] Pravesh K Kothari et al. "Sum of squares lower bounds for refuting any CSP". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 132–145.

[KT81] R. Krichevsky and V. Trofimov. "The performance of universal encoding". In: *IEEE Transactions on Information Theory* 27.2 (1981), pp. 199–207. DOI: 10.1109/TIT.1981.1056331.

[Leh21] Luc Lehéricy. "Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models". In: *Electronic Journal of Statistics* 15.2 (2021), pp. 4916–4965.

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[Mir60] L. Mirsky. "SYMMETRIC GAUGE FUNCTIONS AND UNITARILY INVARIANT NORMS". In: *Quarterly Journal of Mathematics* 11 (1960), pp. 50–59. URL: https://api.semanticsch

[MR05] Elchanan Mossel and Sébastien Roch. "Learning nonsingular phylogenies and hidden Markov models". In: *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. 2005, pp. 366–375.

[PW24] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. http://www.stat.ya Cambridge University Press, 2024.

[Ris84] J. Rissanen. "Universal coding, information, prediction, and estimation". In: *IEEE Transactions on Information Theory* 30.4 (1984), pp. 629–636. DOI: 10.1109/TIT.1984.1056936.

[Sha+17] Vatsal Sharan et al. "Learning Overcomplete HMMs". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[Sha+18] Vatsal Sharan et al. "Prediction with a short memory". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1074–1087.

[Ste69] G. W. Stewart. "On the Continuity of the Generalized Inverse". In: *SIAM Journal on Applied Mathematics* 17.1 (1969), pp. 33–45. DOI: 10.1137/0117004. eprint: https://doi.org/10.113 URL: https://doi.org/10.1137/0117004.

[WS21] Thom Wiggers and Simona Samardjiska. "Practically Solving LPN". In: *2021 IEEE International Symposium on Information Theory (ISIT)*. 2021, pp. 2399–2404. DOI: 10.1109/ISIT45174.2021.9518109.

[XB00] Qun Xie and Andrew R Barron. "Asymptotic minimax regret for data compression, gambling, and prediction". In: *IEEE Transactions on Information Theory* 46.2 (2000), pp. 431–445.

[YB99] Yuhong Yang and Andrew Barron. "Information-theoretic determination of minimax rates of convergence". In: *Annals of Statistics* (1999), pp. 1564–1599.

A Preliminaries and technical lemmas

Recall the following information-theoretic quantities. For probability distributions P_X and Q_X on the space $\mathcal{X}$, the KL divergence is $\mathrm{KL}(P_X\|Q_X) = \mathbb{E}_P\left[\log\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}\right]$ if $P_X \ll Q_X$ and ∞ otherwise. The conditional KL divergence is $\mathrm{KL}(P_{X|Y}\|Q_{X|Y}|P_Y) = \mathbb{E}_{y\sim P_Y}[\mathrm{KL}(P_{X|Y=y}\|Q_{X|Y=y})]$. The mutual information between random variables X and Y is defined as $I(X;Y) = \mathrm{KL}(P_{XY}\|P_X \otimes P_Y) = \mathrm{KL}(P_{X|Y}\|P_X|P_Y)$ and the conditional mutual information is defined similarly $I(X;Y|Z) = \mathrm{KL}(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})$.

We use $o, O, \omega, \Omega, \Theta$ following the common big-O notations, where an added $\widetilde{(\cdot)}$ denotes ignoring log factors. We also use $\lesssim, \gtrsim, \asymp$ to denote comparison ignoring universal constants. We shorthand $\mathbb{N} = \{1, 2, \dots\}$ and $[t] = \{1, 2, \dots, t\}$.

A.1 Comparison with the formulation in [Sha+18]

Recall that our prediction risk with respect to a true model $P_{X^{n+1}}$ is defined as

$$\mathsf{Risk}(Q_{X_{n+1}|X^n}; P_{X^{n+1}}) = \mathbb{E}_{P_{X^{n+1}}}[\mathrm{KL}(P_{X_{n+1}|X^n}\|Q_{X_{n+1}|X^n})] \tag{21}$$

which compares a prediction algorithm $Q_{X_{n+1}|X^n}$ to the oracle prediction $P_{X^{n+1}|X^n}$. Maximizing $P_{X^{n+1}}$ in a given model class, e.g., HMM, leads to the worst-case risk in (2).

In [Sha+18], the authors formulated the prediction problem differently as follows. First, it is assumed that observed sample path can be extended to a double-sided process $(X_t)_{t\in\mathbb{Z}}$. Then, for a sequence of predictors $Q_{X_{t+1}|X^t}$ indexed by the sample size $t = 1, \dots, n$ they consider the prediction loss in TV with respect to the respective oracle $P_{X^{t+1}|X_{-\infty}^t}$ and define:

$$\mathsf{Risk}'(\{Q_{X_{t+1}|X^t}\}_{t=1}^n; P_{X^{n+1}}) = \mathbb{E}_{X_{-\infty}^{n+1}}\left[\frac{1}{n}\sum_{t=1}^n \mathrm{TV}(P_{X_{t+1}|X_{-\infty}^t}, Q_{X_{t+1}|X^t})\right]. \tag{22}$$

It is not straightforward to compare results under this formulation to our results partly due to this averaging over t, which means the prediction guarantee is not made for a given sample size n by on average for a random sample size drawn uniformly from 1 to n. Nevertheless, a firm comparison one can make is the following. In the spirit of (22), consider the following variant of (21):

$$\mathsf{Risk}''(Q_{X_{n+1}|X^n}; P_{X^{n+1}}) = \mathbb{E}_{P_{X^{n+1}}}[\mathrm{KL}(P_{X_{n+1}|X_{-\infty}^n}\|Q_{X_{n+1}|X^n})]. \tag{23}$$

In other words, the goal is to compete with an oracle who only knows the true model parameters but also has access to infinite historical data. While this appears to be a more difficult task, for HMM the difference of these two risks is in fact negligible. Indeed, by the chain rule, we have for any predictor Q and any model P,

$$\mathsf{Risk}'' - \mathsf{Risk} = I(X_{-\infty}^0; X_{n+1}|X^n) \le I(Z_1; X_{n+1}|X^n)$$

which, for HMM with k hidden states, is at most $\log k/n$, thanks to (24). As a result, all of our results proved in Risk (which is more natural in our settings) apply immediately to Risk''.

A.2 Technical lemmas

Lemma 4. *Let $U \to X \to \widehat{U}$ be a Markov chain with U being a continuous random variable with a density function f_U taking values in $[0,t]^d$, and $\|\widehat{U}-U\|_2^2 \le d\varepsilon^2$ with probability at least 0.99. Let $h(U) \triangleq \int f_U(u)du \log \frac{1}{f_U(u)}$ denote the differential entropy of U. Then*

$$I(U;X) \ge h(U) + d\log\frac{1}{\varepsilon\sqrt{2\pi e}} - 0.01d\log\frac{1}{t} - \log 2.$$

Proof. Let E be the event that $\|\widehat{U}-U\|_2^2 \le d\varepsilon^2$. Then

$$\begin{aligned}
I(U;X) \ge I(U;\widehat{U}) &= h(U) - h(U|\widehat{U}) \\
&= h(U) - h(U|\widehat{U},\mathbb{1}_E) - I(U;\mathbb{1}_E|\widehat{U}) \\
&\ge h(U) - h(U|\widehat{U},E)\mathbb{P}(E) - h(U|\widehat{U},E^c)\mathbb{P}(E^c) - \log 2 \\
&\overset{\text{(a)}}{\ge} h(U) - h(U-\widehat{U}|\widehat{U},E)\mathbb{P}(E) - 0.01d\log\frac{1}{t} - \log 2 \\
&\ge h(U) - h(U-\widehat{U}|E) - 0.01d\log\frac{1}{t} - \log 2 \\
&\overset{\text{(b)}}{\ge} h(U) + d\log\frac{1}{\varepsilon\sqrt{2\pi e}} - 0.01d\log\frac{1}{t} - \log 2,
\end{aligned}$$

where (a) and (b) apply the fact that the differential entropy is maximized by the uniform (resp. Gaussian) distribution subject to a support (resp. second moment) constraint. □

The following lemma bounds the change of the prediction risk when certain auxiliary information is observed.

Lemma 5. *For a generic prior on θ, the model parameters, and an auxiliary random variable U, it holds that*

$$\begin{aligned}
&\inf_{\widehat{P}_{X_{n+1}|X^n,U}} \mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\theta,U}\|\widehat{P}_{X_{n+1}|X^n,U})] \\
&\le \inf_{\widehat{P}_{X_{n+1}|X^n}} \mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\theta}\|\widehat{P}_{X_{n+1}|X^n})] + I(U;X_{n+1}|X^n,\theta).
\end{aligned}$$

where the expectation is taken with respect to both the model parameters θ according to the prior and the observations X^n, U

Proof. By the mutual information representation of the prediction risk (cf. [HJW23, Appendix A]), the statement is equivalent to

$$I(\theta;X_{n+1}|X^n,U) \le I(\theta;X_{n+1}|X^n) + I(U;X_{n+1}|X^n,\theta).$$

This is obvious since

$$I(\theta;X_{n+1}|X^n,U) = I(\theta;X_{n+1}|X^n) + I(U;X_{n+1}|X^n,\theta) - I(U;X_{n+1}|X^n)$$

by the chain rule of mutual information. □

Lemma 6 (Mirsky's theorem, [Mir60])**.** *For matrices $A, E \in \mathbb{R}^{m\times k}$ with $m \ge k$, it holds that*

$$\sum_{i=1}^{k} (\sigma_i(A+E) - \sigma_i(A))^2 \le \|E\|_{\mathrm{F}}^2,$$

where $\sigma_i (i \in [k])$ are the sorted singular values.

B Deferred proofs of upper bounds

B.1 Proof of Proposition 1

It holds that

$$
\begin{aligned}
&\mathrm{KL}\left(P_{X_{n+1}|X^n}\|\widetilde{Q}_{X_{n+1}|X^n}|P_{X^n}\right)\\
&\overset{\text{(a)}}{=}\mathbb{E}\left[\mathrm{KL}\left(P_{X_{n+1}|X^n}\left(\cdot|X^n\right)\middle\|\frac{1}{n}\sum_{t=1}^{n}Q_{X_{t+1}|X^t}\left(\cdot|X_{n-t+1}^n\right)\right)\right]\\
&\overset{\text{(b)}}{\le}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\log\frac{P_{X_{n+1}|X^n}\left(X_{n+1}|X^n\right)}{Q_{X_{t+1}|X^t}\left(X_{n+1}|X_{n-t+1}^n\right)}\right]\\
&=\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\log\frac{P_{X_{n+1}|X_{n-t+1}^n}\left(X_{n+1}|X_{n-t+1}^n\right)}{Q_{X_{t+1}|X^t}\left(X_{n+1}|X_{n-t+1}^n\right)}\right]+\mathbb{E}\left[\log\frac{P_{X_{n+1}|X^n}\left(X_{n+1}|X^n\right)}{P_{X_{n+1}|X_{n-t+1}^n}\left(X_{n+1}|X_{n-t+1}^n\right)}\right]\\
&\overset{\text{(c)}}{=}\frac{1}{n}\sum_{t=1}^{n}\underbrace{\mathbb{E}\left[\log\frac{P_{X_{t+1}|X^t}\left(X_{t+1}|X^t\right)}{Q_{X_{t+1}|X^t}\left(X_{t+1}|X^t\right)}\right]}_{\mathrm{KL}\left(P_{X_{t+1}|X^t}\|Q_{X_{t+1}|X^t}|P_{X^t}\right)}+I\left(X_{n+1};X^{n-t}|X_{n-t+1}^n\right)\\
&\overset{\text{(d)}}{=}\frac{1}{n}\left(\mathrm{KL}\left(P_{X^{n+1}}\|Q_{X^{n+1}}\right)-\mathrm{KL}\left(P_{X_1}\|Q_{X_1}\right)+\sum_{t=1}^{n}I\left(X_{n+1};X^{n-t}|X_{n-t+1}^n\right)\right)
\end{aligned}
$$

where (a) applies the definition (8); (b) is due to the convexity of the KL divergence; (c) uses the crucial fact that due to stationarity, $(X_{n-t+1},\ldots,X_{n+1})\overset{\text{law}}{=}(X_1,\ldots,X_{t+1})$ and $P_{X_{n+1}|X_{n-t+1}^n}=P_{X_{t+1}|X^t}$ for all t; (d) applies the chain rule of KL divergence. Using $\mathrm{KL}\left(P_{X_1}\|Q_{X_1}\right)\ge 0$ and taking the supremum over $P_{X^{n+1}}$, the proposition follows. □

B.2 Proof of Proposition 2

By the hidden Markov structure, $X^{n-t}\to(X_{n-t+1}^n,Z_{n-t+1})\to X_{n+1}$ forms a Markov chain, or equivalently, $X^{n-t}\to Z_{n-t+1}\to X_{n+1}$ conditioned on X_{n-t+1}^n. Thus data processing inequality yields

$$I(X^{n-t};X_{n+1}|X_{n-t+1}^n)\le I(Z_{n-t+1};X_{n+1}|X_{n-t+1}^n).$$

By stationarity, we have $I(Z_{n-t+1};X_{n+1}|X_{n-t+1}^n)=I(Z_1;X_{t+1}|X_1^t)$ and, furthermore

$$\sum_{t=1}^{n}I(X^{n-t};X_{n+1}|X_{n-t+1}^n)\le\sum_{t=1}^{n}I(Z_1;X_{t+1}|X_1^t)=I(Z_1;X^{n+1}),$$

by the chain rule of mutual information. □

Furthermore, since $(Z_1,X_2)\to Z_2\to X_3^{n+1}$ is a Markov chain, one has that:

$$H(X_{n+1}|Z_1,X_2^n)\ge H(X_{n+1}|Z_2,X_3^n)=H(X_n|Z_1,X_2^{n-1})$$

where the last equality follows from stationarity. Moreover, clearly

$$H(X_{n+1}|X_1^n)\le H(X_{n+1}|X_2^n)=H(X_n|X_1^{n-1}).$$

Taking the difference of the two equations, one has that: $I(Z_1;X_{n+1}|X_1^n)\le I(Z_1;X_n|X_1^{n-1})$. Therefore applying the chain rule $\sum_{t=1}^{n}I(Z_1;X_{t+1}|X_1^t)=I(Z_1;X^{n+1})$ and $I(Z_1;X^{n+1})\le H(Z_1)\le\log k$, we obtain

$$I(Z_1;X_{n+1}|X_1^n)\le\frac{1}{n}I(Z_1;X^{n+1})\le\frac{1}{n}\log k. \tag{24}$$

B.3 Proof of Proposition 3

The following result on Markov estimators was proven in [HJW21, Lemma 7]:

Lemma 7 (Markov redundancy). *For $n \geq 2k^2$, any initial distribution π, and Markov transition matrix M, the marginal distribution $Q_{Z^{n+1}}$ defined in (15) and (16) satisfies*

$$\max_{z^{n+1}} \log \frac{\pi(z_1)\prod_{t=1}^{n} M(z_{t+1}|z_t)}{Q_{Z^{n+1}}(z^{n+1})} \lesssim k^2 \log \frac{n}{k^2}.$$

Now for the joint distribution $Q_{X^{n+1},Z^{n+1}}$,

$$\log \frac{\pi(z_1)\prod_{t=1}^{n} M(z_{t+1}|z_t)\prod_{t=1}^{n+1} T(x_t|z_t)}{Q_{X^{n+1},Z^{n+1}}(x^{n+1},z^{n+1})} = \log \frac{\pi(z_1)\prod_{t=1}^{n} M(z_{t+1}|z_t)}{Q_{Z^{n+1}}(z^{n+1})} + \log \frac{\prod_{t=1}^{n+1} T(x_t|z_t)}{Q_{X^{n+1}|Z^{n+1}}(x^{n+1}|z^{n+1})}.$$

The upper bound of the first term is stated in Lemma 7. For the second term, let N_i be the number of appearances of i in z^{n+1}, and N_{ij} be the number of appearances of (i,j) in the state-emission pairs $(z_t,x_t)_{t=1}^{n+1}$. Then (15) and (17) imply that

$$Q_{X^{n+1}|Z^{n+1}}(x^{n+1}|z^{n+1}) = \prod_{i=1}^{k} \frac{\prod_{j=1}^{\ell} N_{ij}!}{\ell^{\overline{N_i}}},$$

where $a^{\overline{m}} = a(a+1)\cdots(a+m-1)$ is the rising factorial. Therefore,

$$\begin{aligned}
\log \frac{\prod_{t=1}^{n+1} T(x_t|z_t)}{Q_{X^{n+1}|Z^{n+1}}(x^{n+1}|z^{n+1})} &= \log \prod_{i=1}^{k} \ell^{\overline{N_i}} \prod_{j=1}^{\ell} \frac{T(j|i)^{N_{ij}}}{N_{ij}!} \\
&\overset{(a)}{\leq} \sum_{i=1}^{k} \log \frac{\ell^{\overline{N_i}}}{N_i!} \\
&\overset{(b)}{\leq} \sum_{i=1}^{k} \left((\ell-1)\log\left(1+\frac{N_i}{\ell-1}\right) + N_i \log\left(1+\frac{\ell-1}{N_i}\right)\right) \\
&\overset{(c)}{\lesssim} k\ell \log \frac{n}{k\ell},
\end{aligned}$$

where (a) follows from the multinomial theorem $\binom{N_i}{N_{i1}\ \cdots\ N_{ik}} \prod_{j=1}^{\ell} T(j|i)^{N_{ij}} \leq 1$, (b) is due to

$$\begin{aligned}
\log \frac{\ell^{\overline{m}}}{m!} = \sum_{i=1}^{m} \log\left(1+\frac{\ell-1}{i}\right) &\leq \int_0^m \log\left(1+\frac{\ell-1}{x}\right) \mathrm{d}x \\
&= (\ell-1)\log\left(1+\frac{m}{\ell-1}\right) + m\log\left(1+\frac{\ell-1}{m}\right),
\end{aligned}$$

and (c) uses the concavity of $x \to \log x$, $\log(1+x) \leq x$, and $\sum_{i=1}^{k} N_i = n+1$. The above two terms then complete the proof of the first statement. The second statement (marginalization) simply follows from

$$\begin{aligned}
\mathsf{Red}(Q_{X^{n+1}}) &\leq \max_{x^{n+1}} \log \frac{\sum_{z^{n+1}} \pi(z_1)\prod_{t=1}^{n} M(z_{t+1}|z_t)\prod_{t=1}^{n+1} T(x_t|z_t)}{\sum_{z^{n+1}} Q_{X^{n+1},Z^{n+1}}(x^{n+1},z^{n+1})} \\
&\leq \max_{x^{n+1},z^{n+1}} \log \frac{\pi(z_1)\prod_{t=1}^{n} M(z_{t+1}|z_t)\prod_{t=1}^{n+1} T(x_t|z_t)}{Q_{X^{n+1},Z^{n+1}}(x^{n+1},z^{n+1})} \\
&\lesssim k^2 \log \frac{n}{k^2} + k\ell \log \frac{n}{k\ell}.
\end{aligned}$$

□

B.4 Proof of Corollary 1

By Propositions 1 and 2, it suffices to construct $Q_{X^{n+1}}$ such that

$$\mathsf{Red}(Q_{X^{n+1}}; \mathcal{P}_n^{\mathsf{HMM}}(k, \mathcal{Q})) \lesssim k^2 \log \frac{n}{k^2} + kR((n+1)/k).$$

To this end, let $Q_{X^{n+1}}$ be the marginal distribution of $Q_{X^{n+1},Z^{n+1}}$, where $Q_{Z^{n+1}}(z^{n+1})$ is again given by (15) and (16). For the conditional distribution $Q_{X^{n+1}|Z^{n+1}}(x^{n+1}|z^{n+1})$, let $I_z(z^{n+1}) = \{t \in [n+1] : z_t = z\}$ be the time indices where $z_t = z$, we construct

$$Q_{X^{n+1}|Z^{n+1}}(x^{n+1}|z^{n+1}) = \prod_{z\in[k]} Q^\star_{|I_z(z^{n+1})|}(x_{I_z(z^{n+1})}),$$

where $Q^\star_m$ is the joint distribution such that $\mathsf{Red}(Q^\star_m; \mathcal{Q}^{\otimes m}) \le R(m)$. Then for any HMM in $\mathcal{P}_n^{\mathsf{HMM}}(k, \mathcal{Q})$ with transition matrix M, stationary distribution π, and emission distributions $(Q_z)_{z\in[k]} \subseteq \mathcal{Q}$, one has

$$\begin{aligned}
&\log \frac{\pi(z_1)\prod_{t=1}^n M(z_{t+1}|z_t)\prod_{t=1}^{n+1} Q_{z_t}(x_t)}{Q_{X^{n+1},Z^{n+1}}(x^{n+1}, z^{n+1})} \\
&= \log \frac{\pi(z_1)\prod_{t=1}^n M(z_{t+1}|z_t)}{Q_{Z^{n+1}}(z^{n+1})} + \sum_{z\in[k]} \log \frac{\prod_{t\in I_z(z^{n+1})} Q_z(x_t)}{Q^\star_{|I_z(z^{n+1})|}(x_{I_z(z^{n+1})})}.
\end{aligned}$$

Taking expectation over x^{n+1} conditioned on z^{n+1} gives that

$$\mathbb{E}\left[\log \frac{\prod_{t\in I_z(z^{n+1})} Q_z(x_t)}{Q^\star_{|I_z(z^{n+1})|}(x_{I_z(z^{n+1})})}\right] \le \mathsf{Red}(Q^\star_{|I_z(z^{n+1})|}; \mathcal{Q}^{\otimes|I_z(z^{n+1})|}) \le R(|I_z(z^{n+1})|),$$

so the concavity of R leads to

$$\mathbb{E}\left[\sum_{z\in[k]} \log \frac{\prod_{t\in I_z(z^{n+1})} Q_z(x_t)}{Q_{|I_z(z^{n+1})|}(x_{I_z(z^{n+1})})}\right] \le \sum_{z\in[k]} R(|I_z(z^{n+1})|) \le kR((n+1)/k).$$

A combination of the above inequality and Lemma 7 completes the proof of the desired redundancy upper bound.

Remark 2. *As discussed after Corollary 1, for Gaussian location model $\mathcal{Q} = \{\mathcal{N}(w, I_d) : w \in [-1,1]^d\}$, one has*[2] $\mathsf{Red}(\mathcal{Q}^{\otimes n}) \lesssim d\log n$ *and thus* $\mathsf{Red}(\mathcal{P}_n^{\mathsf{HMM}}(k, \mathcal{Q})) \lesssim k^2 \log \frac{n}{k^2} + d\log n$. *This is also implied by the pointwise redundancy bound of HMM with Gaussian emissions in [Gas18, Sec. 4.2.3].*

C Deferred proofs of lower bounds

C.1 Improved redundancy-based risk lower bound

In this section we prove Proposition 5 which improves over Proposition 4. Let $\mathcal{Q}$ be an arbitrary collection of distributions on the observation space $\mathcal{X}$. Let $\mathcal{P}_{n,\mathsf{U}}^{\mathsf{HMM}}(k, \mathcal{Q})$ denote the class of all stationary HMMs of length n with hidden states in $[k]$, emissions in $\mathcal{X}$ with emission probabilities

[2] For sharp asymptotic bounds of both average and pointwise redundancy in Gaussian models, see [XB00].

chosen in $\mathcal{Q}$. In addition, let $\mathcal{P}_{n,\mathsf{U}}^{\mathsf{HMM}}(k,\mathcal{Q})$ denote the subclass of HMMs whose stationary distribution over the states is uniform. Let $\mathsf{Red}(\mathcal{Q})$ denote the redundancy of the distribution class $\mathcal{Q}$ in the same sense as (7), namely,

$$\mathsf{Red}(\mathcal{Q}) \triangleq \inf_{Q'} \sup_{Q\in\mathcal{Q}} \mathrm{KL}(Q\|Q')$$

which, by the capacity-redundancy theorem (see, e.g., [PW24, (13.10)]), equals

$$\mathsf{Red}(\mathcal{Q}) = \sup I(\theta; X) \tag{25}$$

where $P_{X|\theta} \in \mathcal{Q}$ and the maximization is taken over the distribution (prior) of θ. Proposition 5 lower bounds the prediction risk of $\mathcal{P}_n^{\mathsf{HMM}}(k,\mathcal{Q})$ using the redundancy of the (slightly smaller) class $\mathcal{P}_{n,\mathsf{U}}^{\mathsf{HMM}}(k,\mathcal{Q})$, which is the set of all stationary HMMs in $\mathcal{P}_n^{\mathsf{HMM}}(k,\mathcal{Q})$ with uniform stationary distribution for the hidden states.

Proof of Proposition 5. Given a HMM configuration Q (comprising the transition probability $Q_{X_2|X_1}$ and emission probability $Q_{X_1|Z_1}$) in $\mathcal{P}_{n+1,\mathsf{U}}^{\mathsf{HMM}}(k-1,\mathcal{Q})$, we construct an HMM P in $\mathcal{P}_n^{\mathsf{HMM}}(k,\mathcal{Q})$ as follows. in a similar way as that in Section 4.1. The emission probabilities for states $i \in [k-1]$ are identical, namely $P_{X_1|Z_1=i} = Q_{X_1|Z_1=i}$ and $P_{X_1|Z_1=k}$ is any fixed distribution in $\mathcal{Q}$. The transition probabilities $P_{X_2|X_1}$ is defined based on $Q_{X_2|X_1}$ using (18). Hence, both HMM configurations P and Q are parameterized by the transition matrix $Q_{X_2|X_1}$ and the emission probability matrix $Q_{X_1|Z_1}$. Let θ denote the collection of these parameters.

Let $T \in \{1, 2, \cdots, n, \perp\}$ be the smallest $t \in [n]$ such that $Z_t \neq k$ (if $Z^n = k^n$ then $T = \perp$). In the later proof we will condition on T; a subtlety here is that T is determined by the states Z^n but it may not be measurable with respect to X^n. Nevertheless, we will consider the setting where both the estimand $P_{X_{n+1}|X^n}$ and the estimator $\widehat{P}_{X_{n+1}|X^n}$ have access to the extra information T. While replacing $\widehat{P}_{X_{n+1}|X^n}$ by $\widehat{P}_{X_{n+1}|X^n,T}$ is valid for the sake of lower bound, replacing $P_{X_{n+1}|X^n}$ by $P_{X_{n+1}|X^n,T}$ requires further justification. By Lemma 5, this increases the prediction risk by at most $I(T; X_{n+1}|X^n)$ which we bound below.

Lemma 8. *For any fixed θ, we have*

$$I(T; X_{n+1}|X^n) \lesssim \frac{\log(kn)}{n}.$$

Proof. Let $U_t = \mathbb{1}_{T=t}$ for $t \in [n]$, then T is determined by U^n. Then

$$\begin{aligned}
I(T; X_{n+1}|X^n) &\le I(U^n; X_{n+1}|X^n) \\
&= \sum_{t=1}^{n} I(U_t; X_{n+1}|X^n, U^{t-1}) \\
&\overset{\text{(a)}}{\le} I(U_1; X_{n+1}|X^n) + \sum_{t=2}^{n} I(U_t; X_{n+1}|X^n, U^{t-1} = 0^{t-1}) \\
&= I(U_1; X_{n+1}|X^n) + \sum_{t=2}^{n} I(U_t; X_{n+1}|X^n, Z^{t-1} = k^{t-1}) \\
&\overset{\text{(b)}}{=} I(U_1; X_{n+1}|X^n) + \sum_{t=2}^{n} I(U_t; X_{n+1}|X_t^n, Z_{t-1} = k) \\
&\overset{\text{(c)}}{=} I(U_1; X_{n+1}|X^n) + \sum_{t=2}^{n} I(U_2; X_{n+3-t}|X_2^{n+2-t}, Z_1 = k) \\
&= I(U_1; X_{n+1}|X^n) + I(U_2; X_3^{n+1}|X_2, Z_1 = k) \\
&\overset{\text{(d)}}{\le} \frac{H(Z_1)}{n} + H(U_2|Z_1 = k) \overset{\text{(e)}}{\lesssim} \frac{\log(nk)}{n},
\end{aligned}$$

where (a) uses the fact that $U_t = 0$ deterministically if any entry of U^{t-1} is one so that

$$I(U_t; X_{n+1}|X^n, U^{t-1}) = I(U_t; X_{n+1}|X^n, U^{t-1} = 0^{t-1})\mathbb{P}(U^{t-1} = 0^{t-1}|X^n);$$

(b) is due to the Markov structure $(X^{t-1}, Z^{t-2}) \to (X_t^m, Z_{t-1}) \to U_t$ given $Z^{t-1} = k^{t-1}$ for $m \in \{n, n+1\}$; (c) is due to stationarity, and (d) follows from $I(U_1; X_{n+1}|X^n) \le I(Z_1; X_{n+1}|X^n) \le_{(24)} H(Z_1)/n$; (e) is because by definition (18), conditioned on $Z_1 = k$, $U_2 \sim \mathrm{Bern}(\frac{1}{n})$. □

Next we assume an arbitrary prior on θ and use this Bayesian setting to lower bound the prediction risk. Similar to the analysis in Section 4.1, it is clear that for every $t \in [n]$, it holds that $\mathbb{P}(T = t) = \Omega(1/n)$. Conditioned on the event $T = t$, the random distribution $P_{X_{n+1}|X^n,\theta,T=t}$ shares the same law as $P_{Y_{n-t+2}|Y^{n-t+1},\theta}$, where Y^m is a sample path of length m from the same k-state HMM, but with the initial hidden state Z_1 drawn uniformly from $[k-1]$. Similarly, conditioned on $T = t$, the posterior distribution of θ given X^n has the same law as the posterior distribution of θ given Y^{n-t+1}. Consequently, the Bayes prediction risk satisfies

$$\begin{aligned}
&\inf_{\widehat{P}_{X_{n+1}|X^n,T}} \mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\theta,T} \| \widehat{P}_{X_{n+1}|X^n,T})] \\
&\ge \sum_{t=1}^{n} \mathbb{P}(T = t) \cdot \inf_{\widehat{P}_{X_{n+1}|X^n,T}} \mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\theta,T} \| \widehat{P}_{X_{n+1}|X^n,T})|T = t] \\
&= \sum_{t=1}^{n} \mathbb{P}(T = t) \cdot \inf_{\widehat{P}_{Y_{n+2-t}|Y^{n+1-t}}} \mathbb{E}[\mathrm{KL}(P_{Y_{n+2-t}|Y^{n+1-t},\theta} \| \widehat{P}_{Y_{n+2-t}|Y^{n+1-t}})] \\
&\gtrsim \frac{1}{n} \sum_{t=1}^{n} I(\theta; Y_{n+2-t}|Y^{n+1-t}) = \frac{I(\theta; Y^{n+1}) - I(\theta; Y_1)}{n}.
\end{aligned}$$

To deal with the above terms, for the second mutual information we have

$$I(\theta; Y_1) \le I(\theta; Y_1, Z_1) = I(\theta; Y_1|Z_1) + I(\theta; Z_1) \overset{\text{(a)}}{=} I(\theta; Y_1|Z_1) \overset{\text{(b)}}{\le} \mathsf{Red}(\mathcal{Q}),$$

where (a) uses that the distribution of Z_1 is uniform regardless of θ by definition of the model class $\mathcal{P}^{\mathsf{HMM}}_{n+1,\mathsf{U}}$, and (b) is because for any state $i \in [k]$, $I(\theta; Y_1|Z_1 = i) \le \mathsf{Red}(\mathcal{Q})$ by the capacity-redundancy representation (25), since the emission probabilities $P_{Y_1|Z_1=i,\theta}$ are chosen from $\mathcal{Q}$.

For the first mutual information $I(\theta; Y^{n+1})$, let (Y^{n+1}, Z^{n+1}) be the observations and hidden states in the HMM with k states (starting from $Z_1 \sim \mathsf{Unif}([k-1])$), and X_Q^{n+1} be the observations in the reduced HMM (with $k-1$ states and model parameters Q). Let E be the event that $Z_t \neq k$ for all $t \in [n+1]$, by chain rule $I(\theta; Y^{n+1}) = I(\theta, \mathbb{1}_E; Y^{n+1}) - I(\mathbb{1}_E; Y^{n+1}|\theta)$, we have

$$I(\theta; Y^{n+1}) \ge I(\theta; Y^{n+1}|\mathbb{1}_E) - \log 2 \ge \mathbb{P}(E) I(\theta; Y^{n+1}|E) - \log 2.$$

We note that $\mathbb{P}(E) = (1-1/n)^n = \Omega(1)$, and the joint distribution $P_{\theta, Y^{n+1}|E}$ is the same as $P_{\theta, X_Q^{n+1}}$. Therefore,

$$I(\theta; Y^{n+1}) \gtrsim I(\theta; X_Q^{n+1}) - \log 2.$$

Using Lemmas 5, 8, as well as the above inequalities yields a lower bound on the Bayes (and hence minimax) prediction risk:

$$\mathsf{Risk}(\mathcal{P}_n^{\mathsf{HMM}}(k, \mathcal{Q})) \gtrsim \frac{1}{n} I(\theta; X_Q^{n+1}) - \frac{\log(nk) + \mathsf{Red}(\mathcal{Q})}{n}.$$

Finally, taking the supremum over the prior of the model parameter θ and invoking the capacity-redundancy identity (25) complete the proof of Proposition 5. □

C.2 The case of $\ell = 2$ and proof of Corollary 2

When $\ell = 2$, we distinguish into two cases: $k = 2$, and $k \ge C$ for a large absolute constant $C > 0$. In the first case we establish an $\Omega(\log n/n)$ lower bound for the prediction risk of $\mathcal{P}_n^{\mathsf{HMM}}(2, 2)$, and in the second case we prove Corollary 2 when $n \ge k^D$, which implies the lower bound in Theorem 1 for $\mathsf{Risk}_{\mathsf{HMM}}(n, k, 2)$.

C.2.1 $k = 2$

Let us first consider the case $k = \ell = 2$. Note that we cannot directly apply Proposition 5 as the remainder term $\log(nk)/n$ is too large. Instead, similar to the general reduction in Section 4.1, we consider a specific transition matrix for the hidden states:

$$M = \begin{bmatrix} \frac{n-1}{n} & \frac{1}{n} \\ \frac{1}{n} & \frac{n-1}{n} \end{bmatrix}.$$

For the emission probabilities, we set $\mathbb{P}(X = 1|Z = 0) = 1$ and $\mathbb{P}(X = 1|Z = 1) = \theta$, where the prior distribution of θ is $\mathsf{Unif}([0.1, 0.9])$. We show that the Bayes prediction risk is $\Omega(\log n/n)$.

For $t \in [n-1]$, let E_t be the event $Z^t = 0^t, Z_{t+1}^n = 1$ and $X_n = 0$. Clearly

$$\mathbb{P}(E_t) \ge \frac{1}{2} \cdot \left(1 - \frac{1}{n}\right)^{t-1} \cdot \frac{1}{n} \cdot \left(1 - \frac{1}{n}\right)^{n-t-1} \cdot (1 - 0.9) \ge \frac{1}{60n}.$$

We investigate the Bayes prediction risk conditioned on E_t. For the true distribution $P_{X_{n+1}|X^n,\theta}$, the event E_t implies $X_n = 0$ (observable from X^n) and consequently $Z_n = 1$, so that $P_{X_{n+1}|X^n,\theta} =$

$\mathsf{Bern}(\frac{n-1}{n}\theta + \frac{1}{n})$. For the estimator $Q_{X_{n+1}|X^n}$, the posterior distribution of θ given (X^n, E_t) is the same as the posterior distribution of θ given $Y^{n-t} \sim \mathsf{Bern}(\theta)^{\otimes(n-t)}$. Consequently,

$$\begin{aligned}
&\inf_{Q(\cdot|X^n)} \mathbb{E}\left[\log \frac{P(X_{n+1}|X^n,\theta)}{Q(X_{n+1}|X^n)}\middle| E_t\right] \\
&\geq \inf_{Q(\cdot|X^n,E_t)} \mathbb{E}\left[\log \frac{P(X_{n+1}|X^n,\theta)}{Q(X_{n+1}|X^n,E_t)}\middle| E_t\right] \\
&\geq \inf_{\widehat{p}(X^n,E_t)} \mathbb{E}\left[\mathrm{KL}\left(\mathsf{Bern}\left(\frac{n-1}{n}\theta+\frac{1}{n}\right)\|\mathsf{Bern}(\widehat{p}(X^n,E_t))\right)\middle| E_t\right] \\
&= \inf_{\widehat{p}(Y^{n-t})} \mathbb{E}\left[\mathrm{KL}\left(\mathsf{Bern}\left(\frac{n-1}{n}\theta+\frac{1}{n}\right)\|\mathsf{Bern}(\widehat{p}(Y^{n-t}))\right)\right] \\
&\geq 2\left(\frac{n-1}{n}\right)^2 \inf_{\widehat{\theta}(Y^{n-t})} \mathbb{E}\left[(\theta-\widehat{\theta}(Y^{n-t}))^2\right] \gtrsim \frac{1}{n-t},
\end{aligned}$$

where we have used Pinsker's inequality and the $\Omega(1/(n-t))$ Bayes mean squared error of estimating θ under $Y^{n-t} \sim \mathsf{Bern}(\theta)^{\otimes(n-t)}$. Therefore, the overall Bayes prediction risk is

$$\begin{aligned}
\inf_{Q(\cdot|X^n)} \mathbb{E}\left[\log \frac{P(X_{n+1}|X^n,\theta)}{Q(X_{n+1}|X^n)}\right] &\geq \sum_{t=1}^{n-1} \mathbb{P}(E_t) \cdot \inf_{Q(\cdot|X^n)} \mathbb{E}\left[\log \frac{P(X_{n+1}|X^n,\theta)}{Q(X_{n+1}|X^n)}\middle| E_t\right] \\
&\gtrsim \sum_{t=1}^{n-1} \frac{1}{n}\cdot\frac{1}{n-t} \gtrsim \frac{\log n}{n},
\end{aligned}$$

completing the proof of $\mathsf{Risk}_{\mathsf{HMM}}(n,2,2) = \Omega(\log n/n)$.

C.2.2 Corollary 2: $k \geq C$ and $n \geq k^D$

Since Corollary 2 implies the lower bound of Theorem 1 in this case, it suffices to prove Corollary 2. In fact, by Proposition 5, the final step in proving Corollary 2 is to lower bound the redundancy. This can be divided into two steps:

1. That $\mathsf{Red}(\mathcal{P}^{\mathsf{HMM}}_{n,\mathsf{U}}(k,\mathcal{Q})) \gtrsim k^2 \log n$. This follows immediately from Lemma 3 and the construction therein (Lemma 14). Indeed, let Q the transition matrix be sampled according to Lemma 14 and plug in $t_1 < t_2$ in Lemma 3 be according to $c_1 < c_2$ in Corollary 2. Let f be the map pushing $Q \in \mathcal{Q}$ to Bernoulli's in Corollary 2. Then one has that:

$$I(X^n_Q;\theta) \geq I(f(X^n_Q);\theta) \gtrsim k^2 \log n.$$

2. That $\mathsf{Red}(\mathcal{P}^{\mathsf{HMM}}_{n,\mathsf{U}}(k,\mathcal{Q})) \gtrsim k\mathsf{Red}(\mathcal{Q}^{\otimes\lfloor n/k\rfloor}) - \log k$, where we ignore the $k \to k-1$ issue since $\mathsf{Red}(\mathcal{Q}^t)$ grows at most linearly in t. To show this, consider the HMM whose latent states evolve according to the cycle C_k, and each emission $P_i = \mathbb{P}(X|Z=i)$ is sampled independently from some distribution μ supported on $\mathcal{Q}$. Then:

$$\sup_{\mu} I(P^k;X^n) \geq \sup_{\mu} I(P^k;X^n,Z_1) - \log k = k\mathsf{Red}(\mathcal{Q}^{\otimes\lfloor n/k\rfloor}) - \log k$$

by optimizing the prior μ. Therefore Corollary 2 is proven.

Remark 3. *In fact, the risk from emission $\frac{k}{n}\mathsf{Red}(\mathcal{Q}^{\otimes\lfloor n/k\rfloor})$ holds even for $k=2$ and $n \in O(1)$. The constraint on $k \geq C, n \geq k^D$ is only used to show $\mathsf{Risk} \gtrsim \frac{k^2}{n}\log n$ via Lemma 3.*

C.3 Proof of Lemma 3

The proof of Lemma 3 relies critically on the following Markov property:

$$\begin{aligned} P_{X_{-L}^{L}}(x_{-L}^{L}) &= \sum_{z_0 \in [k]} P_{Z_0}(z_0) P_{X_{-L}^{L}|Z_0}(x_{-L}^{L}|z_0) \\ &= \sum_{z_0 \in [k]} P_{Z_0}(z_0) P_{X_0|Z_0}(x_0|z_0) P_{X_1^{L}|Z_0}(x_1^{L}|z_0) P_{X_{-L}^{-1}|Z_0}(x_{-L}^{-1}|z_0) \\ &= \frac{1}{k} \sum_{z_0 \in [k]} P_{X_0|Z_0}(x_0|z_0) P_{X_1^{L}|Z_0}(x_1^{L}|z_0) P_{X_{-L}^{-1}|Z_0}(x_{-L}^{-1}|z_0), \end{aligned}$$

where the last step uses the fact that the stationary distribution is uniform when the transition matrix is doubly stochastic. In a tensor form, we write $P_{X_{-L}^{L}}$ as a $2^L \times 2^L \times 2$ tensor M, and express $P_{X_0|Z_0=z_0}, P_{X_1^{L}|Z_0=z_0}, P_{X_{-L}^{-1}|Z_0=z_0}$ as vectors $o_{z_0} \in \mathbb{R}^2, e_{z_0} \in \mathbb{R}^{2^L}, f_{z_0} \in \mathbb{R}^{2^L}$, respectively. Then we have the following tensor decomposition of the moment matrix ([Hua+15]):

$$M = \frac{1}{k} \sum_{z_0 \in [k]} e_{z_0} \otimes f_{z_0} \otimes o_{z_0}. \tag{26}$$

We also let $E, F \in \mathbb{R}^{2^L \times k}$ be the matrices with column vectors e_{z_0} and f_{z_0}, respectively.

Based on the above tensor decomposition, the proof consists of several steps:

1. In the first step, we show that there exist constants $C_1, C_2, d_1, d_2, d_3 > 0$ such that for $L = \lceil d_1 \log k \rceil$, based on the HMM trajectory X^n one may construct an estimator $\widehat{M}_n$ such that
$$\mathbb{P}(\|\widehat{M}_n - M\|_{\mathrm{F}} \geq n^{-d_2}) \leq C_1 \exp(-C_2 n^{d_3}). \tag{27}$$
2. In the second step, we show that based on the estimate $\widehat{M}_n$ and tensor decomposition, one may construct estimators $\widehat{E}_n, \widehat{F}_n$ for matrices E, F such that (up to permutations of columns)
$$\mathbb{P}(\|\widehat{E}_n - E\|_{\mathrm{F}} + \|\widehat{F}_n - F\|_{\mathrm{F}} \geq n^{-d_4}) \leq 0.01, \tag{28}$$
where $d_4 > 0$ is an absolute constant;
3. In the last step, we conclude the statement of Lemma 3 based on (28).

We break these steps into several subsections.

C.3.1 First step: proof of (27)

Note that if the transition matrix is precisely the cycle C_k, then X^{2L+1} and X_T^{T+2L} are independent if we choose a random time index $T \sim \mathsf{Unif}(\{2L+2, \cdots, 2L+k+1\})$. The following lemma states that this is essentially the case whenever the transition matrix Q is close to C_k.

Lemma 9. *Consider a Markov chain (Z_t) with transition matrix $Q \in \mathbb{R}^{k \times k}$ such that $\|Q - \mathrm{C}_k\|_{\max} \leq \varepsilon$, and that the stationary distribution is uniform. Then for any $m \in \mathbb{N}$, there exists a distribution P_m supported on $[mk]$ such that if $T \sim P_m$ (independent of the chain), then*

$$\max_{z \in [k]} \|\mathsf{Unif}([k]) - P_{Z_T|Z_0=z}\|_{\mathrm{TV}} \leq (2k\varepsilon)^m.$$

Proof. Let $T_1, T_2, \cdots$ be i.i.d. from $\mathsf{Unif}([k])$, and $T = \sum_{i=1}^m T_i$. Clearly T is supported on $[mk]$. Note that for any $z, z' \in [k]$, one has

$$\mathbb{P}(Z_{T_1} = z' | Z_0 = z) \geq \frac{1}{k}(1-\varepsilon)^k,$$

where $1/k$ is the probability that the T_1 equals to the time z travels to z' under C_k, and $(1-\varepsilon)^k$ lower bounds the probability of following the path in C_k. This implies that

$$\max_{z \in [k]} \|\mathsf{Unif}([k]) - P_{Z_{T_1}|Z_0=z}\|_{\mathrm{TV}} \leq 1 - (1-\varepsilon)^k \leq k\varepsilon.$$

Now the result follows from the standard mixing bound for the new chain $(Z_0, Z_{T_1}, Z_{T_1+T_2}, \cdots)$ (cf. [LP17, Lemma 4.12]). □

Now we choose $m = \lceil \sqrt{n} \rceil$, and consider $J = \lceil n^{1/3} \rceil$ disjoint intervals $I_j = [(j-1)(mk+2L) + T_j, (j-1)(mk+2L) + T_j + 2L] \subseteq [(j-1)(mk+2L)+1, j(mk+2L)]$, where $T_1, \cdots, T_J$ are i.i.d. according to the distribution P_m in Lemma 9. Note that $I_1, \cdots, I_J \subseteq [n]$ for $n \geq k^6$. By the HMM structure, for any $j \in [J]$ it holds that

$$\begin{aligned}
\mathbb{E}\|P_{X^{2L+1}} - P_{X_{I_j}|(X_{I_i})_{i<j}}\|_{\mathrm{TV}} &\overset{\text{(a)}}{\leq} \mathbb{E}\|P_{Z_1} - P_{Z_{t_{j-1}+T_j}|(X_{I_i})_{i<j}}\|_{\mathrm{TV}} \\
&\overset{\text{(b)}}{\leq} \mathbb{E}\|\mathsf{Unif}([k]) - P_{Z_{t_{j-1}+T_j}|Z_{t_{j-1}}}\|_{\mathrm{TV}} \\
&\overset{\text{(c)}}{\leq} (2k\varepsilon)^m,
\end{aligned}$$

where (a) follows from the data processing inequality for the TV distance (where t_j is the end time of I_j), (b) is due to the Markov structure $(X_{I_i})_{i<j} \to Z_{t_{j-1}} \to Z_{t_{j-1}+T_j}$, and (c) follows from Lemma 9. Therefore, by the subadditivity of TV distance,

$$\|P_{X^{2L+1}}^{\otimes J} - P_{(X_{I_j})_{j\in J}}\|_{\mathrm{TV}} \leq \sum_{j=1}^{J} \mathbb{E}\|P_{X^{2L+1}} - P_{X_{I_j}|(X_{I_i})_{i<j}}\|_{\mathrm{TV}} \leq J(2k\varepsilon)^m = e^{-\widetilde{\Omega}(n)}, \tag{29}$$

where the last inequality follows from the assumption $\varepsilon \leq k^{-c_2}$ in Lemma 3, for $c_2 \geq 2$.

Using the near independence, we proceed to estimate the joint distribution $P_{X^{2L+1}}$ (or equivalently the tensor M) based on the empirical distribution $\widehat{M}_n$ of $\{X_{I_j}\}_{j\in J}$. If $\{X_{I_j}\}_{j\in J}$ were indeed i.i.d., by Hoeffding's inequality and union bound we would have

$$\begin{aligned}
\mathbb{P}(\|\widehat{M}_n - M\|_{\mathrm{F}} \geq n^{-d_2}) &\leq \mathbb{P}(\|\widehat{M}_n - M\|_{\max} \geq n^{-d_2}) \\
&\leq 2^{2L+1} \cdot 2\exp(-2J(n^{-d_2})^2) \\
&= \exp(-\widetilde{\Omega}(n^{1/2-2d_2})).
\end{aligned}$$

For weakly dependent $\{X_{I_j}\}_{j\in J}$, we invoke (29) to conclude that

$$\mathbb{P}(\|\widehat{M}_n - M\|_{\mathrm{F}} \geq n^{-d_2}) \leq \exp(-\widetilde{\Omega}(n^{1/2-2d_2})) + e^{-\widetilde{\Omega}(n)}.$$

Consequently, (27) holds with $d_2 = d_3 = 1/8$.

C.3.2 Second step: proof of (28)

Given the estimate $\widehat{M}_n$ of the tensor M, we aim to recover the matrices E and F (up to permutations of columns). To this end we recall the following result in tensor decomposition.

Lemma 10 (Stability of tensor decomposition, Theorem 2.3 in [Bha+13]). *Let $T = \sum_{i=1}^k u_i \otimes v_i \otimes w_i$ be a tensor satisfying the following conditions:*

1. *The condition numbers $\kappa(U), \kappa(V) \le \kappa$, where U, V are $m \times k$ matrices with column vectors u_i and v_i, respectively;*
2. *The vectors $w_i \in \mathbb{R}^2$ are not close to parallel: $\min_{i \neq j} \left\| \frac{w_i}{\|w_i\|} - \frac{w_j}{\|w_j\|} \right\|_2 \ge \delta > 0$;*
3. *The decompositions are bounded: for all i, $\|u_i\|_2, \|v_i\|_2, \|w_i\|_2 \le 1$.*

Now given the noisy tensor $T + E \in \mathbb{R}^{m \times m \times 2}$ with the entries of E bounded by ε, there exists an efficient algorithm that returns each rank one term in the decomposition of T (up to renaming), within an additive error of $\varepsilon \cdot \mathrm{poly}(m, k, \kappa, 1/\delta)$.

Note that in our tensor decomposition (26), we have $m = 2^L = \mathrm{poly}(k)$, and $\varepsilon \le kn^{-d_2}$ by (27) with high probability. Consequently, if the conditions of Lemma 10 hold with $\kappa, 1/\delta = \mathrm{poly}(k)$, then by choosing $c_1 > 0$ large enough in the condition $n \ge k^{c_1}$ in Lemma 3, Lemma 10 will imply (28). Hence it remains to verify the conditions of Lemma 10.

The third condition is straightforward: a probability vector e must satisfy $\|e\|_2 \le \|e\|_1 = 1$. To verify the first two conditions, we use a probabilistic argument and choose the vectors o_i (i.e. emission probabilities $P_{X|Z}$) randomly. Specifically, for fixed constants a_1, a_2 with $0 < a_1 < a_2 < \pi/2$, we generate i.i.d. angles $\theta_1, \cdots, \theta_k \sim \mathsf{Unif}([a_1, a_2])$, and set

$$p_i = \mathbb{P}(X = 1 | Z = i) = \frac{\cos\theta_i}{\cos\theta_i + \sin\theta_i}, \quad i \in [k].$$

Note that for appropriately chosen constants a_1, a_2, the condition $p_i \in (t_1, t_2)$ required in Lemma 3 holds almost surely. The reason behind the choice of p_i is summarized by the following lemma.

Lemma 11. *If $Q = \mathrm{C}_k$ and $k \ge L$, then for a large enough constant $d_1 > 0$, there exist absolute constants $d_5, d_6 > 0$ such that*

$$\mathbb{P}(\max\{\kappa(E), \kappa(F)\} \le k^{d_5}) \ge 1 - 2k^{-d_6},$$

where $\kappa(E)$ denotes the condition number of E, i.e. the ratio between the largest and the kth singular values of E.

Proof. By symmetry we only prove the claim for E. If $Q = \mathrm{C}_k$ and $k \ge L$, we have

$$E_{x^L, i} = P_{X_1^L | Z_0}(x^L | i) = \prod_{t=1}^{L} \frac{(\cos\theta_{t+i})^{1-x_t} (\sin\theta_{t+i})^{x_t}}{\cos\theta_{t+i} + \sin\theta_{t+i}},$$

where the indices of θ are understood modulo k. Consequently, for $i, j \in [k]$,

$$(E^\top E)_{i,j} = \prod_{t=1}^{L} \frac{\cos(\theta_{t+i} - \theta_{t+j})}{(\cos\theta_{t+i} + \sin\theta_{t+i})(\cos\theta_{t+j} + \sin\theta_{t+j})}.$$

In matrix forms, $E^\top E = DAD$, where $D \in \mathbb{R}^{k\times k}$ is a diagonal matrix with $D_{ii} = \prod_{t=1}^{L}(\cos\theta_{t+i} + \sin\theta_{t+i})^{-1} \in [2^{-L/2}, 1]$, and $A_{ij} = \prod_{t=1}^{L}\cos(\theta_{t+i} - \theta_{t+j})$. Consequently,

$$\lambda_1(E^\top E) \le \lambda_1(A), \qquad \lambda_k(E^\top E) \ge (2^{-L/2})^2\lambda_k(A) \ge 2^{-L}\lambda_k(A). \tag{30}$$

Next we analyze the matrix A. Clearly the diagonal entries of A are all 1. For off-diagonal entries A_{ij}, as $k \ge L$, we may pick $N \ge L/2$ elements $t_1, \cdots, t_N \in [L]$ such that the set $\{i + t_r, j + t_r : r \in [N]\}$ contains $2N$ distinct elements. Then

$$\log A_{ij} \le \sum_{r=1}^{N} \log\cos(\theta_{i+t_r} - \theta_{j+t_r}),$$

where the RHS is the sum of N i.i.d. random variables. Since $\mathbb{E}[\log\cos(\theta_1 - \theta_2)] \le -c(a_1, a_2)$ for some constant $c(a_1, a_2) > 0$, and $|\log\cos(\theta_1 - \theta_2)| \le |\log\cos(a_1 - a_2)|$ almost surely, Hoeffding's inequality implies that $\log A_{ij} = -\Omega(N)$ with probability at least $1 - \exp(-\Omega(N)) = 1 - k^{-\Omega(d_1)}$. By choosing d_1 large enough, a union bound implies that with probability at least $1 - k^{-d_6}$, all off-diagonal entries of A have magnitude at most $1/(2k)$. Then by Gershgorin circle theorem,

$$\lambda_1(A) \le 1 + \max_{i\in[k]} \sum_{j\ne i} |A_{ij}| \le 1 + \frac{1}{2k}\cdot(k-1) \le \frac{3}{2},$$
$$\lambda_k(A) \ge 1 - \max_{i\in[k]} \sum_{j\ne i} |A_{ij}| \ge 1 - \frac{1}{2k}\cdot(k-1) \ge \frac{1}{2}$$

hold with high probability. A combination of (30) and the above result completes the proof. □

The result of Lemma 11 assumes $Q = \mathrm{C}_k$, and we still need to generalize it to the case where $\|Q - \mathrm{C}_k\|_{\max} \le k^{-c_2}$. To this end we apply a matrix perturbation analysis. Let $E^\star$ be the matrix E with columns $e_i^\star$ under $Q = \mathrm{C}_k$, then

$$\begin{aligned}
\|e_i^\star - e_i\|_1 &= 2\|P_{X^L|Z_0=i,Q=\mathrm{C}_k} - P_{X^L|Z_0=i,Q}\|_{\mathrm{TV}} \\
&\overset{\text{(a)}}{\le} 2\|P_{Z^L|Z_0=i,Q=\mathrm{C}_k} - P_{Z^L|Z_0=i,Q}\|_{\mathrm{TV}} \\
&\overset{\text{(b)}}{=} 2\left(1 - \prod_{t=1}^{L} Q_{i+t-1,i+t}\right) \\
&\le 2\left(1 - (1-k^{-c_2})^L\right) \le 2Lk^{-c_2},
\end{aligned} \tag{31}$$

where (a) follows from the data processing equality for the TV distance and that $P_{X^L|Z^L}$ does not depend on Q, and (b) observes that $P_{Z^L|Z_0=i,Q=\mathrm{C}_k}$ is supported on a single path $z^L = (i+1, \cdots, i+L-1)$ (all indices are understood modulo k), and $\|P-Q\|_{\mathrm{TV}} = \sum_{x:P(x)>Q(x)}(P(x)-Q(x))$. Consequently,

$$\|E^\star - E\|_{\mathrm{F}}^2 = \sum_{i=1}^{k}\|e_i^\star - e_i\|_2^2 \le \sum_{i=1}^{k}\|e_i^\star - e_i\|_1^2 = \widetilde{O}(k^{1-2c_2}), \tag{32}$$

so Mirsky's theorem (cf. Lemma 6) shows that for a large enough constant $c_2 > 0$, the condition number of E is close to the condition number of $E^\star$. This shows that $\kappa = \mathrm{poly}(k)$ with probability at least $1 - 2k^{-d_6}$ for the first condition in Lemma 10.

Finally we check the second condition in Lemma 10. Since $\theta_1, \cdots, \theta_k$ are i.i.d. and uniformly distributed on $[a_1, a_2]$, it holds that

$$\mathbb{P}(|\theta_1 - \theta_2| \le k^{-3}) \lesssim k^{-3}.$$

By a union bound, with probability at least $1 - 1/k$, we have $|\theta_i - \theta_j| \ge 1/k^3$ for all $i \neq j$. By the definition of p_i, this implies that for $i \neq j$,

$$\left\| \frac{(p_i, 1-p_i)}{\sqrt{p_i^2 + (1-p_i)^2}} - \frac{(p_j, 1-p_j)}{\sqrt{p_j^2 + (1-p_j)^2}} \right\|_2 \gtrsim |p_i - p_j| \gtrsim |\theta_i - \theta_j| \ge k^{-3},$$

so that $\delta = \Omega(1/k^3)$ in Lemma 10 with probability at least $1 - 1/k$.

In summary, for $k \ge c_0$ with a large enough constant $c_0 > 0$, all conditions in Lemma 10 are satisfied for $\kappa, 1/\delta = \mathrm{poly}(k)$ with probability at least 0.99. By the arguments under Lemma 10 we arrive at (28), as desired.

C.3.3 Third step: proof of Lemma 3

Given accurate estimates of $\widehat{E}_n, \widehat{F}_n$, we now seek to recover the transition matrix $\widehat{Q}_n$ with a small error. To this end, we note the following lemma:

Lemma 12 (Tensor to Transition, Theorem 4 in [Hua+15]). *Given matrix $E \in \mathbb{R}^{2^L \times k}$ such that $E_{x^L, z_0} = \mathbb{P}\left(X^L = x^L | Z_0 = z_0\right)$ is the conditional forward moment. We marginalize the conditional distribution to $E^{(L-1)}_{x^{L-1}, z_0} = \mathbb{P}\left(X^{L-1} = x^{L-1} | Z_0 = z_0\right) \in \mathbb{R}^{2^{L-1} \times k}$. If E has full column rank k, then the transition matrix (let $O \in \mathbb{R}^{2 \times k}$ be the emission):*

$$Q = \left(O \odot E^{(L-1)}\right)^\dagger E \triangleq B^\dagger E.$$

Specifically, the Khatri-Rao product $O \odot E^{(L-1)} \in \mathbb{R}^{2^L \times k}$ is exactly:

$$B_{x_0^{L-1}, z_0} = \left(O \odot E^{(L-1)}\right)_{x_0^{L-1}, z_0} = \mathbb{P}(X_0^{L-1} = x_0^{n-1} | Z_0 = z_0) \tag{33}$$

and $X^\dagger$ denotes the pseudo-inverse of a matrix X.

Firstly, we show that $\kappa(B) \ge k^{-O(1)}$ whenever the emission is such that Lemma 11 is satisfied and that the perturbation from Q to C_k is not large. This is because when $Q = \mathrm{C}_k$, the corresponding $B^\star$ is exactly a column permutation of $E^\star$. In other words, for all emission matrices O one has that $\kappa(E^\star) = \kappa(B^\star)$. Following the exact same lines as (31) (replacing Z_0 with Z_1), we get that:

$$\|B - B^\star\|_{\mathrm{F}} = \widetilde{O}(k^{1-2c_2})$$

and hence by Mirsky's theorem we know that the condition number for $\kappa(B) = \mathrm{poly}(k)$ with probability at least $1 - 2k^{-d_6}$.

Taking the union of such events, we now have $\kappa(B), \kappa(E)$ all upper bounded by $\mathrm{poly}(k)$, and we wish to show that for $\widehat{Q}_n = \widehat{B}_n^\dagger \widehat{E}_n$, the error

$$Q - \widehat{Q}_n = B^\dagger (E - \widehat{E}_n) + (B^\dagger - \widehat{B}_n^\dagger)\widehat{E}_n$$

has small norm. Note that

$$\begin{aligned}\|Q-\widehat{Q}_n\|_{\mathrm{op}} &\le \|B^\dagger(E-\widehat{E}_n)\|_{\mathrm{op}} + \|(\widehat{B}_n^\dagger - B^\dagger)\widehat{E}_n\|_{\mathrm{op}} \\ &\le \|B^\dagger\|_{\mathrm{F}}\|E-\widehat{E}_n\|_{\mathrm{op}} + \|\widehat{E}_n\|_{\mathrm{F}}\|\widehat{B}_n^\dagger - B^\dagger\|_{\mathrm{op}} \\ &\le 2^L k\left(\|E-\widehat{E}_n\|_{\mathrm{op}} + \|\widehat{B}_n^\dagger - B^\dagger\|_{\mathrm{op}}\right).\end{aligned}$$

By (28), one only need to upper bound $\|\widehat{B}_n^\dagger - B^\dagger\|_{\mathrm{op}} \in O(n^{-c})$ for some constant $c > 0$. Note that

$$\|B-\widehat{B}_n\|_{\mathrm{F}} \le \|E^{(L-1)} - \widehat{E}_n^{(L-1)}\|_{\mathrm{F}} \le 2\|E-\widehat{E}_n\|_{\mathrm{F}}$$

by (33) and therefore $\|\widehat{B}_n - B\|_{\mathrm{F}} \in O(n^{-c'})$ for some $c' > 0$ from (28). Our conclusion follows from the following lemma:

Lemma 13 (Theorem 4.3 in [Ste69])**.** *Let A be an $m \times n$ matrix of rank n, and let the error matrix be E. Let $\kappa = \|A\|_{\mathrm{op}}\left\|A^\dagger\right\|_{\mathrm{op}}$ and $H = (A+E)^\dagger - A^\dagger$. If $\left\|A^\dagger\right\|_{\mathrm{op}}\|E\|_{\mathrm{op}} < 1$, then*

$$\frac{\|H\|_{\mathrm{op}}}{\|A^\dagger\|_{\mathrm{op}}} < (1+\gamma)\beta,$$

where $\gamma = \left(1 - \frac{\kappa\|E\|_{\mathrm{op}}}{\|A\|_{\mathrm{op}}}\right)^{-1}$ and $\beta = \frac{\gamma\kappa\|E\|_{\mathrm{op}}}{\|A\|_{\mathrm{op}}}$.

Plugging in $H = \widehat{B}_n^\dagger - B^\dagger$ into the above lemma and using the fact that $\kappa(B)$ is upper bounded by a polynomial of k, whereas $\|\widehat{B}_n - B\|_{\mathrm{op}}$ is upper bounded by $n^{-c'}$ for some $c' > 0$, we are done.

C.4 Proof of (20)

As a final step, we present the following lemma on a high-entropy construction discussed in the main text that guarantees estimation (19) indeed leads to redundancy lower bounds:

Lemma 14 (Distribution on hidden states)**.** *For any constant $c > 0$, there exists a distribution μ supported on the set $\{Q \in \mathbb{R}^{k\times k} : Q$ is double stochastic, $\|Q - \mathrm{C}_k\|_{\max} \le k^{-c}, Q_{i,i} = 0\}$ such that $h(\mu) \gtrsim -k^2 \log k$.*

Proof. Consider any pair (i,j) such that $0 < i < j-2 < k$. Consider the associated grids $\{(i,j),(j-1,i+1),(i,i+1),(j-1,j)\}$. One can associate an independent random variable $X_{i,j}$ such that $Q_{i,j} = X_{i,j} = Q_{j-1,i+1}$ and $Q_{i,i+1} = 1 - \sum_j X_{i,j} - \sum_j X_{j,i+1}$, $Q_{j-1,j} = 1 - \sum_i X_{i,j} - \sum_i X_{j-1,i}$. This construction will always ensure that the resulting matrix is double stochastic (since each X modifies a 2×2 grid). Furthermore, restricting $0 \le X_{i,j} < k^{-c-2}$ ensures that the max offset from the default C_k (corresponding to all $X = 0$) is at most k^{-c} as desired.

Finally, the entropy is guaranteed as we recover from Q exactly $\Theta(k^2)$ independent random variables that each has a range of $k^{-O(1)}$, when $k > 5$. The cases for small k can be verified easily as when $k = 3$ there exists a trivial construction with constant entropy. □

D Computationally predicting HMMs

D.1 Algorithmic upper bound: small k, ℓ

We show that, for any given matrix M, T one can compute the number of satisfying trajectories $z^n \in [k]^n$ efficiently such that the counts matches exactly M, T.

Algorithm: Count the number of $z^K \in [k]^K$ with given a transition and emission counts.

Input: Matrices $T \in \mathbb{Z}^{k\times\ell}, M \in \mathbb{Z}^{k\times k}$ with non-negative entries, where $\sum T_{ij} = 1 + \sum M_{rs} = K$. Emissions $x^K \in [\ell]^K$.

1. If $K = 1$ and $T_{zx_1} = 1$ for some z, output 1 directly.
2. Check that $\sum_{j\in[k]} M_{ij} = \sum_{j\in[\ell]} T_{ij} = N_i$ for all except for exactly one $i_0 \in [k]$; otherwise, output 0. In this case, $z_K = i_0$ since it is the only item that shows up differently comparing rows of M, T.
3. For $i \in [k]$ let $M^{(i)} \in \mathbb{R}^{k\times k}, T^{(i)} \in \mathbb{R}^{k\times\ell}$ be such that
$$M^{(i)}_{ii_0} = M_{ii_0} - 1$$
$$T^{(i)}_{i_0x_K} = T_{i_0x_K} - 1$$
and all other entries matching M, T otherwise. This i represents candidate z_{K-1}'s.
4. Run algorithm on $(M^{(i)}, T^{(i)}; x^{K-1})$ for all $i \in [k]$, and sum the results over i according to (34).

Output: The number of possible trajectories.

Figure 1: Algorithm for computing $\mathcal{A}(M, T; x^K)$, the number of satisfying hidden state sequences.

Lemma 15 (See Figure 1 and Lemma 1)**.** *The proposed algorithm $\mathcal{A}$ which runs according to the recursion:*
$$\mathcal{A}(M, T; x^K) = \sum_{i\in[k]} \mathcal{A}(M^{(i)}, T^{(i)}; x^{K-1}) \tag{34}$$
computes exactly the count of $z^K \in [k]^K$ with the given transition/emission counts in time $K^{O(k\ell+k^2)}$.

Proof. The proof is via induction on K, assuming that our computation returns the correct result when $K = 1$ (in which case it is straightforward to check the count as either 0 or 1). From matching the number of appearances, $z_K = i_0$ for any trajectory with (M, T) in Figure 1. One thus sums all trajectories with $(z_{K-1}, z_K) = (i, i_0)$, which is a trajectory counting problem on z^{K-1} corresponding to $(M^{(i)}, T^{(i)})$. Assuming that the count is consistent for $K - 1$, the count on K should be consistent as well.

In terms of the runtime: one can simply create an empty array of size $K^{k\ell+k^2}$ first and fill in an item (count) corresponding to some (M, T) at each time some trajectory count is computed. The cost of filling a new item assuming $O(1)$ access to the grid memory is at most $O(k)$, and thus the compute filling the entire grid is at most $K^{O(k\ell+k^2)}$ assuming $K \in \Omega(k + \ell)$. This concludes the runtime. □

Finally, given $Q_{X^n,Z^n}(x^n, z^n) = F(M, T)$ one has that:
$$Q_{X^n}(x^n) = \sum_{z^n} Q_{X^n,Z^n}(x^n, z^n) = \sum_{M,T} F(M, T) \cdot \mathcal{A}(M, T; x^n)$$
can be computed in $n^{O(k^2+k\ell)}$-time.

D.2 Algorithmic upper bound: Markov approximation

When k, ℓ are moderately large, the above algorithm via marginalization becomes intractable, and we need efficient choices of $Q_{X^{n+1}}$ to achieve a small redundancy in Proposition 1. The idea is to drop the structure in Z^{n+1} (hence no marginalization) and apply a Markov approximation directly to X^{n+1}. Specifically, [HJW23, Lemma 23] shows the existence of $Q_{X^{n+1}}$ that

$$\max_{x^{n+1}} \log \frac{P_{X^d}(x^d) \prod_{t=d+1}^{n+1} P_{X_t|X_{t-d}^{t-1}}(x_t|x_{t-d}^{t-1})}{Q_{X^{n+1}}(x^{n+1})} \lesssim \ell^{d+1} \log \frac{n}{\ell^{d+1}} + d \log \ell$$

for $n \geq \ell^{d+1}$, and $Q_{X^{n+1}}$ can be evaluated in time $\mathrm{poly}(n, \ell^d)$. Taking the expectation over $x^{n+1} \sim P_{X^{n+1}}$ leads to the redundancy upper bound of $Q_{X^{n+1}}$:

$$\mathsf{Red}(Q_{X^{n+1}}; \mathcal{P}_n^{\mathsf{HMM}}(k, \ell)) \lesssim \sum_{t=d+1}^{n+1} I(X_t; X^{t-d-1}|X_{t-d}^{t-1}) + \ell^{d+1} \log \frac{n}{\ell^{d+1}} + d \log \ell,$$

where the first term is further upper bounded by

$$(n+1) I(X_{n+1}; X^{n-d}|X_{n+1-d}^n) \overset{\text{(a)}}{\leq} \frac{n+1}{d+1} \sum_{t=0}^{n} I(X_{n+1}; X^{n-t}|X_{n-t+1}^n) \overset{\text{(b)}}{\leq} \frac{n+1}{d+1} \log k.$$

Here (a) is because of the decreasing property of $t \mapsto I(X_{n+1}; X^{n-t}|X_{n-t+1}^n)$, and (b) follows from Proposition 2. Consequently, by Propositions 1 and 2, this choice of $Q_{X^{n+1}}$ leads to the prediction risk

$$\mathsf{Risk}(\widetilde{Q}_{X_{n+1}|X^n}, \mathcal{P}_n^{\mathsf{HMM}}(k, \ell)) \lesssim \frac{\log k}{d} + \frac{\ell^{d+1}}{n} \log \frac{n}{\ell^{d+1}} + \frac{d \log \ell + \log k}{n}.$$

Choosing $d = \log n / (2 \log \ell)$ leads to the risk upper bound in Theorem 2. The overall computational time is $\mathrm{poly}(n, \ell^d) = \mathrm{poly}(n)$, as desired.

D.3 Computational lower bounds

In the last part of our computational discussions we sketch two lower bounds, in contrast with our $O(\frac{\log k \log \ell}{\log n})$ upper bound. Our lower bounds will be based on cryptographic assumptions involving the Learning Parity with Noise (LPN) problem ([BKW03]) and refutation of a class of Constraints Satisfying Problem (CSP) ([FPV15]). In particular, the following assumptions are observed.

Conjecture 1 (Learning Parity With Noise, see e.g. [WS21])**.** *Let the secret key $\mathbf{s} \sim_{\text{unif}} \mathbb{F}_2^k$ and noise $\eta = 0.05$. Any polynomial-time algorithm on the Learning Parity with Noise (LPN) problem with $\Omega(1)$-time query access to a noisy observation $y = \langle \mathbf{s}, \mathbf{x} \rangle \oplus \mathrm{Bern}(\eta)$ for $\mathbf{x} \sim_{\text{unif}} \mathbb{F}_2^k$, requires $2^{\Omega(k/\log k)}$ computational complexity to decide between pure noise ($y \sim \mathrm{Bern}(1/2)$) and noisy parity ($y = \langle \mathbf{s}, \mathbf{x} \rangle \oplus \mathrm{Bern}(\eta)$) correctly with probability $2/3$.*

Conjecture 2 (Refuting CSP's, see e.g. [Kot+17])**.** *Let $k > r$ be constants, Q be any distribution over k-clauses with N variables of complexity r and $0 < \eta < 1$. Any polynomial-time algorithm that, given access to a distribution D that equals either the uniform distribution over k-clauses U_k or a (noisy) planted distribution $Q_{\boldsymbol{\sigma}}^{\eta} = (1-\eta) Q_{\boldsymbol{\sigma}} + \eta U_k$ for some $\boldsymbol{\sigma} \in \{0,1\}^n$ and planted distribution Q_σ, decides correctly whether $D = Q_\sigma^\eta$ or $D = U_k$ with probability $2/3$ needs $\widetilde{\Omega}\left(N^{r/2}\right)$ clauses. (Here $\widetilde{\Omega}$ ignores $\log$ factors.)*

Given the above assumptions, our lower bounds are as follows:

Theorem 6 (Computational lower bounds for HMM prediction)**.** *The following holds:*

1. *For any $\varepsilon > 0$, if $k \geq \log^{1+\varepsilon} n$ and $\ell \geq 2$, then there exists a distribution on HMMs where no efficient algorithm can achieve $o(\frac{\log k}{\log n \log\log n})$ error, assuming Conjecture 1.*
2. *For every $\alpha > 0$ there exists $k_\alpha \geq 2$, such that if $k \geq k_\alpha$ and $\ell \geq n^\alpha$, there exists a distribution on HMMs where no efficient algorithm can achieve $o(1)$ error assuming Conjecture 2.*

Remark 4. *This result, combined with Theorem 2, leaves the following cases of interest open in terms of computational algorithms:*

1. *For $k = n^{\Omega(1)}$ and $\ell = 2$, can there be efficient algorithm achieving $o(1)$ risk?*
2. *For $k = O(1)$ and $\ell = \mathrm{polylog}(n)$, can there be computational lower bounds of $1/\mathrm{polylog}(n)$ for prediction risk?*

The embedding of cryptographically hard models into computational lower bounds in HMM has been long observed in various prior literature (e.g. [MR05; Sha+18]). Here we adopt these constructions into our setting.

Proof of Theorem 6. We divide the two cases:

1. For the $\ell = 2$ case. Let $s = \lfloor \log_2 \frac{k}{r} \rfloor - 2$ for some r and let the k hidden states be labeled:

$$Z = \{(i, b_0, b_1, b_2, \dots b_s); b \in \{0,1\}, i = 1, 2, \dots, r+s\}$$

and hence $|Z| = (r+s)2^{r+1} \leq r2^{r+2} \leq k$. We will choose $r = C \log n \log\log n$ for a large constant C, so that $s \in \Omega(\log k)$ thanks to the assumption $k \geq \log^{1+\varepsilon} n$.

Let $(s_{i,j})_{i\in[r],j\in[s]}$ be independent $\mathrm{Bern}(1/2)$ secret keys, so that there are s secret keys in total, each of length r. The transitions and emissions are defined as follows:

- Emission: state (i, b_0^s) emits b_0 if $i \leq r$, and $b_{i-r} \oplus \mathrm{Bern}(\eta)$ if $i \in \{r+1, \cdots, r+s\}$;
- Transition: state (i, b_0^s) goes to $(i+1, c_0^s)$ (as usual, $r+s$ goes to 1), where:
  (a) If $i \in \{r, r+1, \cdots, r+s-1\}$, let $c_0^s = b_0^s$;
  (b) If $i \in \{r+s, 1, \cdots, r-1\}$, sample $c_0 \sim \mathrm{Bern}(1/2)$, and let $c_j = b_j \oplus s_{i+1,j} c_0$ for all $j = 1, \cdots, s$ (as usual, $s_{i+1,j} = s_{1,j}$ when $i = r+s$).

In other words, the transition runs in cycles of length $r+s$. During the first r rounds in each cycle, the learner observes $\mathbf{b}_0 = (b_{0,1}, \cdots, b_{0,r}) \sim \mathsf{Unif}(\mathbb{F}_2^r)$. For the $(r+j)$-th round with $j \in [s]$, under the current transition the learner observes

$$\langle \mathbf{b}_0, \mathbf{s}_j \rangle \oplus \mathrm{Bern}(\eta), \quad \text{where } \mathbf{s}_j = (s_{1,j}, \cdots, s_{r,j}).$$

In other words, each cycle consists of one query to each of s independent LPN instances. Since there are $\leq n$ cycles in total, each LPN instance has sample size at most n.

Next we understand the prediction problem of the current HMM. Clearly, under the stationary distribution we have $i \sim \mathsf{Unif}([r+s])$, so that $\mathbb{P}(i \in \{r+1, \cdots, r+s\}) = s/(r+s)$ for the state at time $n+1$. Again, using Lemma 5 and (24), we assume without loss of generality that the starting state i at $t = 0$ is known to the learner, so that the learner knows the relative

location of time $n+1$ in a given cycle. Suppose time $n+1$ is the $(r+j)$-th round of some cycle, then predicting X_{n+1} is the same as predicting the distribution $\langle \mathbf{b}_0, \mathbf{s}_j \rangle \oplus \mathrm{Bern}(\eta)$ with observed $\mathbf{b}_0 = (b_{0,1}, \cdots, b_{0,r}) \sim \mathsf{Unif}(\mathbb{F}_2^r)$ and a hidden secret $\mathbf{s}_j$. As $n \le 2^{cr/\log r}$ with $c \to 0$ as $C \to \infty$, Conjecture 1 implies that the KL prediction error is $\Omega(1)$ by choosing C large enough. Consequently, the overall KL prediction risk is lower bounded by

$$\Omega(1) \cdot \frac{s}{r+s} = \Omega\left(\frac{\log k}{\log n \log\log n}\right).$$

Now choosing the growth of ω in the definition of r arbitrarily slow gives the claim.

2. For the $\ell = n^\alpha$ case, this follows directly from plugging in the parameter correspondence to Theorem 2 in [Sha+18] while leaking the first hidden state in the fashion of Lemma 5[3]. In short, when r is a large enough constant in Conjecture 2, there exists a distribution on constant-size clauses such that detection is impossible on the CSP problem which can be embedded with $O(1)$ hidden states. Therefore, with constant probability, one cannot distinguish the next bit from random.

□

E Lower bound proof for renewal processes

In this section we prove the lower bound part of Theorem 4, namely, $\mathsf{Risk}_{\mathsf{rnwl}}(n) \gtrsim \sqrt{n^{-1}}$. Similar to the strategies proving lower bounds in Appendix C, we consider a Bayesian setting where the model parameter (in this case the interarrival distribution μ) is random and drawn from some prior. Then the Bayes prediction KL risk is given by the conditional mutual information $I(\mu; X_{n+1}|X^n)$, which we aim to show is at least $\Omega(\sqrt{n^{-1}})$.

Let us first recall the equivalent HMM representation for renewal processes from Section 3.4 with state space $\mathbb{N}$. The stationary distribution on the hidden states is the same as the distribution of the initial wait time T_0, given by:

$$\pi_\mu(i) = \frac{1}{m(\mu)} \sum_{j>i} \mu(j) \tag{35}$$

where $m(\mu) = \sum_{i\ge 1} i\mu(i)$ is the mean of μ. Notably, X^n has the same law as its time reversal. This reversibility will be exploited in our proof of the lower bound.

We will consider a prior under which μ is always finitely supported. Let the last appearance of "1" in X^n be X_K for $K \le n$, and let $T \triangleq \mathbb{1}_{n+1-K \in \mathrm{supp}(\mu)}$. Note that $T = 0$ implies $X_{n+1} = 0$ almost surely. Denote by $p(X^n, \mu) \triangleq \mathbb{P}(X_{n+1} = 1 | X^n, \mu)$ the optimal predictor who knows the model parameter μ. By (5), this is given by the hazard ratio of μ, namely

$$p(X^n, \mu) = \frac{\mu(n+1-K)}{\sum_{d \ge n+1-K} \mu(d)}. \tag{36}$$

Without knowing μ, the predictor is the average of $p(X^n, \mu)$ over the posterior law of μ given the data X^n. Let $p(X^n) \triangleq \mathbb{P}(X_{n+1} = 1|X^n) = \mathbb{E}_{\mu|X^n}[p(X^n, \mu)]$.

[3]The results in [Sha+18] does not require this lemma as they assumed a slightly different loss; see Appendix A.1. Here we bypass this issue by leaking $U = Z_1$ to both P^{HMM} and Q and adjust the result via Lemma 5.

Let E_1 be some event measurable with respect to $\sigma(X^n)$ to be specified. Let E_2 be the event of $T = 0$, which is measurable with respect to $\sigma(X^n, \mu)$. For any estimator Q, its average risk (with expectations taken over both data X^n and μ according to the prior) satisfies

$$\begin{aligned}
\mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\mu}\|Q_{X_{n+1}|X^n})] &\geq \mathbb{P}(E_1)\cdot\mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\mu}\|Q_{X_{n+1}|X^n})|E_1] \\
&\geq \mathbb{P}(E_1)\cdot\mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n,\mu}\|P_{X_{n+1}|X^n})|E_1] \\
&= \mathbb{P}(E_1)\cdot\mathbb{E}[\mathrm{KL}(\mathrm{Bern}(p(X^n,\mu))\|\mathrm{Bern}(p(X^n)))|E_1] \\
&\geq \mathbb{P}(E_1\cap E_2)\cdot\mathbb{E}[\mathrm{KL}\left(\mathrm{Bern}(0)\|\mathrm{Bern}(p(X^n))\right)|E_1\cap E_2] \\
&\geq \mathbb{P}(E_1\cap E_2)\cdot\mathbb{E}[p(X^n)|E_1\cap E_2] \qquad (37)
\end{aligned}$$

where we used the fact that $T = 0$ implies $p(X^n, \mu) = 0$. Furthermore:

$$p(X^n) = \mathbb{E}_{\mu|X^n}[p(X^n,\mu)] = \mathbb{P}(E_2^c|X^n)\mathbb{E}_{\mu|X^n,E_2^c}[p(X^n,\mu)]. \qquad (38)$$

We will choose a prior under which μ is always uniform over $\Theta(\sqrt{n})$ integers. Therefore, by (36), for all $(\mu, X^n) \in E_2^c$, $p(X^n, \mu) \gtrsim \sqrt{n^{-1}}$. Furthermore, suppose we can show that there exists an event E_1 and a constant $c > 0$ such that: (a) $\mathbb{P}(E_1) > c$ and (b) for all $X^n \in E_1$, one has $\mathbb{P}(T = 0|X^n) = \mathbb{P}(E_2|X^n) \in (c, 1-c)$. Then the last line of (37) is lower bounded by the desired $\Omega(n^{-1/2})$ rate. In the following, we show the construction and proof.

We consider a prior that was previously used for proving the redundancy lower bound in [CS96]. There the goal is to prove that $I(\mu; X^n) \gtrsim \sqrt{n}$ as opposed to $I(\mu; X_{n+1}|X^n) \gtrsim \frac{1}{\sqrt{n}}$ here. Let $a_n = C\sqrt{n}$ be an even number for some large constant C. Let the interarrival distribution μ to be the uniform distribution on an a_n-subset of $[2a_n]$, with its support chosen uniformly over all such sets where exactly half of its elements lies in $[0, a_n]$. In this way, for all $(X^n, \mu) \in E_2^c$ one has that $p(X^n, \mu) \geq \mu(n + 1 - K) \geq \frac{1}{a_n} \asymp \sqrt{n^{-1}}$.

Define E_1 to be the intersection of the following events:

(1) The last appearance of $X_K = 1$ satisfy $K > n - a_n$.

(2) There are at most distinct $\sqrt{n}$ interarrival times in X^n (known as the renewal types [CS96]). Denote this set of interarrivial times by A.

(3) The gap $n + 1 - K$ has never appeared in the past $\sqrt{n}$ interarrivals backwards from X_K.

Clearly E_1 is $\sigma(X^n)$-measurable. We show that $\mathbb{P}(E_1) = \Omega(1)$ for a large enough constant C. By the time reversal property of the renewal process, the distribution of $n + 1 - K$ is given by π_μ in (35). As $m(\mu) \leq 2a_n$ and

$$\sum_{i=1}^{a_n}\sum_{j>i}\mu(j) \geq \sum_{j=a_n+1}^{2a_n} a_n\mu(j) = a_n\mu([a_n+1, 2a_n]) = \frac{a_n}{2},$$

event (1) happens with probability at least 1/4. As for (2), the expectation of the interarrival time is $\geq a_n/2 = C\sqrt{n}/2$. Consequently, for $C > 3$, Hoeffding's inequality shows that event (2) happens with probability $1 - o_n(1)$. For (3), each interarrival time equals to $n + 1 - K$ with probability at most $1/a_n$. By a union bound, event (3) happens with probability at least $1 - \sqrt{n}/a_n = 1 - 1/C$. A union bound then gives that $\mathbb{P}(E_1) \geq 1/4 - 1/C - o_n(1)$, which is $\Omega(1)$ for large enough (n, C).

Next we show that $\mathbb{P}(T = 0|X^n) \in (c, 1-c)$ whenever E_1 holds. Let us first prove a simple lemma:

Lemma 16. *For any μ_1, μ_2 supported in the prior and any $k \in [a_0]$, it holds that*

$$\frac{1}{8} \le \frac{\mathbb{P}(K = n+1-k|\mu_1)}{\mathbb{P}(K = n+1-k|\mu_2)} \le 8.$$

Proof. By the time-reversal property of the renewal process, the conditional distribution of $n+1-K$ conditioned on μ is given by π_μ in (35). Consequently,

$$\mathbb{P}(K = n+1-k|\mu) = \frac{1}{m(\mu)} \sum_{j>k} \mu(j).$$

Since the support of μ has $a_n/2$ elements in $[a_n]$ and $a_n/2$ elements in $[a_n+1, 2a_n]$, we have $m(\mu) \in [a_n/2, 2a_n]$. In addition, as $k \in [a_0]$, we have $\sum_{j>k} \mu(j) \in [1/2, 1]$. This gives $\mathbb{P}(K = n+1-k|\mu) \in [1/(4a_n), 2/a_n]$ for all μ in the support of the prior, and the lemma follows. □

Now we show that for all $X^n \in E_1$ with $K = K(X^n) = n+1-k$, the ratio

$$\frac{\mathbb{P}(T=1|X^n)}{\mathbb{P}(T=0|X^n)} = \frac{\mathbb{P}(T=1|X^n, K=n+1-k)}{\mathbb{P}(T=0|X^n, K=n+1-k)}$$

is bounded above and below by positive constants. Since E_1 holds, $k \le a_n$. First of all,

$$\begin{aligned}
\frac{\mathbb{P}(T=1|K=n+1-k)}{\mathbb{P}(T=0|K=n+1-k)} &= \frac{\mathbb{P}(k \in \mathrm{supp}(\mu)|K=n+1-k)}{\mathbb{P}(k \notin \mathrm{supp}(\mu)|K=n+1-k)} \\
&= \frac{\mathbb{P}(k \in \mathrm{supp}(\mu))}{\mathbb{P}(k \notin \mathrm{supp}(\mu))} \frac{\mathbb{P}(K=n+1-k|k \in \mathrm{supp}(\mu))}{\mathbb{P}(K=n+1-k|k \notin \mathrm{supp}(\mu))} \\
&= \frac{\mathbb{P}(K=n+1-k|k \in \mathrm{supp}(\mu))}{\mathbb{P}(K=n+1-k|k \notin \mathrm{supp}(\mu))} = \Theta(1), \qquad (39)
\end{aligned}$$

where the last step is due to Lemma 16. Second, for the set $A = A(X^n)$ consisting of distinct interarrival times in X^n, the event E_1 implies that $k \notin A$ and $|A| \le \sqrt{n}$. Therefore,

$$\begin{aligned}
&\frac{\mathbb{P}(A \subseteq \mathrm{supp}(\mu)|K=n+1-k, k \in \mathrm{supp}(\mu))}{\mathbb{P}(A \subseteq \mathrm{supp}(\mu)|K=n+1-k, k \notin \mathrm{supp}(\mu))} \\
&= \frac{\mathbb{P}(K=n+1-k|k \notin \mathrm{supp}(\mu))}{\mathbb{P}(K=n+1-k|k \in \mathrm{supp}(\mu))} \frac{\mathbb{P}(K=n+1-k|k \in \mathrm{supp}(\mu), A \subseteq \mathrm{supp}(\mu))}{\mathbb{P}(K=n+1-k|k \notin \mathrm{supp}(\mu), A \subseteq \mathrm{supp}(\mu))} \\
&\quad \cdot \frac{\mathbb{P}(A \subseteq \mathrm{supp}(\mu), k \in \mathrm{supp}(\mu))}{\mathbb{P}(A \subseteq \mathrm{supp}(\mu), k \notin \mathrm{supp}(\mu))} \frac{\mathbb{P}(k \notin \mathrm{supp}(\mu))}{\mathbb{P}(k \in \mathrm{supp}(\mu))} \\
&\overset{(a)}{=} \Theta(1) \cdot \frac{\binom{a_n - |A \cap [a_n]| - 1}{a_n/2}}{\binom{a_n - |A \cap [a_n]| - 1}{a_n/2 - 1}} = \Theta(1) \cdot \frac{a_n/2 - |A \cap [a_n]|}{a_n/2} \overset{(b)}{=} \Theta(1), \qquad (40)
\end{aligned}$$

where (a) is due to Lemma 16, and (b) uses $|A \cap [a_n]| \le |A| \le \sqrt{n} \le a_n/3$ as long as $C \ge 3$. Finally, by writing down the joint pmf of X^n after time reversal, it is clear that

$$\frac{\mathbb{P}(X^n|K=n+1-k, T=1, A \subseteq \mathrm{supp}(\mu))}{\mathbb{P}(X^n|K=n+1-k, T=0, A \subseteq \mathrm{supp}(\mu))} = 1. \qquad (41)$$

Combining the above results leads to

$$\begin{aligned}
\frac{\mathbb{P}(T=1|X^n)}{\mathbb{P}(T=0|X^n)} &= \frac{\mathbb{P}(T=1|X^n, K=n+1-k)}{\mathbb{P}(T=0|X^n, K=n+1-k)} \\
&= \frac{\mathbb{P}(T=1|K=n+1-k)}{\mathbb{P}(T=0|K=n+1-k)} \cdot \frac{\mathbb{P}(X^n|T=1, K=n+1-k)}{\mathbb{P}(X^n|T=0, K=n+1-k)} \\
&\overset{(39)}{=} \Theta(1) \cdot \frac{\mathbb{P}(X^n|T=1, K=n+1-k)}{\mathbb{P}(X^n|T=0, K=n+1-k)} \\
&\overset{(41)}{=} \Theta(1) \cdot \frac{\mathbb{P}(A \subseteq \mathrm{supp}(\mu)|T=1, K=n+1-k)}{\mathbb{P}(A \subseteq \mathrm{supp}(\mu)|T=0, K=n+1-k)} \overset{(40)}{=} \Theta(1).
\end{aligned}$$

Therefore, both conditions $\mathbb{P}(E_1) \geq c$ and $\mathbb{P}(T=0|X^n) \in (c, 1-c)$ for $X^n \in E_1$ are established, and the $\Omega(n^{-1/2})$ lower bound follows from (37) and (38).