Next-symbol Prediction from Compression: Statistical and Computational Foundations

Yanjun Han, Yihong Wu

***Abstract*—Next-symbol prediction is a central task in modern generative models, yet its statistical and computational foundations for dependent data remain poorly understood. This tutorial introduces an information-theoretic framework that connects prediction to universal compression, revealing a fundamental distinction between prediction and parameter estimation. Addressing model classes with finite memory (Markov chains) and infinite memory (Hidden Markov Model and renewal processes), we characterize the minimax prediction risks without imposing mixing conditions, and show how optimal prediction can be achieved using ideas from universal compression. We further quantify the impact of memory and mixing on statistical efficiency, and establish computational hardness results demonstrating that achieving statistically optimal prediction may be computationally infeasible for certain models.**

I. Introduction

Consider the following "ChatGPT-style" problem: Observing a time series $X^n := (X_1, \ldots, X_n)$, one is tasked to predict the next (unseen) symbol X_{n+1}. This problem, known as *next-token prediction*, is the core objective of modern paradigms of generative AI and large language models (LLMs) [1] such as transformers. These technologies have achieved astounding success in the generation and recognition of natural languages using predictive models such as high-order Markov chains [2], a classical idea rooted in information theory dating back to Shannon [3].

The problem of next-symbol prediction boils down to estimating the *conditional* distribution $P_{X_{n+1}|X^n}$, which informs downstream tasks such as sampling for text generation or estimating the most likely realizations for autocomplete. This is a well-defined but non-standard statistical problem, in that the estimand $P_{X_{n+1}|X^n}$ is random and data-dependent, unless the data are independent and identically distributed (iid), in which case the problem reduces to distribution estimation, one of the most well-studied paradigm in statistical inference. While it is common in the statistical learning literature to model data as iid, for applications such as language models the dependency in the data can no longer be ignored. In fact, one must embrace models with (possibly long) memory and learn the temporal dependency from data.

However, for data-generating models with memory, this basic prediction problem is not well-understood even for the simplest instances such as first-order Markov chains. In statistics, most of the literature relies on favorable mixing assumptions, such as a large spectral gap, so that Bernstein-style concentration inequalities continue to hold. These assumptions essentially require the memory in the data is relatively weak, so that both theory and algorithms designed for iid data are still applicable. This presents a conceptual paradox: Although such mixing conditions are necessary for parameter estimation, they are *not* needed for prediction. Indeed, a chain that moves at a glacial speed is 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 without model identifiability.

In this tutorial, building on recent developments by the authors [4]–[6], we connect the problem of next-symbol prediction to the classical topic of *universal compression* [7]–[9]. Using Markov and hidden Markov models as guiding examples, this tutorial will: (a) characterize the information-theoretic limits of

prediction; (b) decouple learning and mixing; (c) quantify the effect of memory on statistical and computational efficiency.

Notation. For $k \in \mathbb{N}$, let $[k] := \{1, \ldots, k\}$. For a discrete random variable X with pmf P, let $H(X) = \mathbb{E}[\log \frac{1}{P(X)}]$ be its Shannon entropy. For distributions P and Q on the same probability space, let $D(P\|Q) = \int \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}$ if $P \ll Q$, and $+\infty$ otherwise, be the Kullback–Leibler (KL) divergence. We also define the conditional KL divergence by

$$D(P_{Y|X}\|Q_{Y|X}|P_X) = \int D(P_{Y|X=x}\|Q_{Y|X=x})\mathrm{d}P_X(x). \quad (1)$$

Under the above notation, the conditional mutual information $I(X;Y|Z)$ can be written as $D(P_{X,Y|Z}\|P_{X|Z} \otimes P_{Y|Z}|P_Z)$. Throughout this paper, we use standard asymptotic notations O, Ω, Θ and $\lesssim, \gtrsim, \asymp$ interchangably.

II. Redundancy and prediction risk

In this section, we review basic concepts of redundancy and prediction risk, and provide a survey of their properties.

A. Compression and prediction

The concepts of redundancy and prediction risk reflect two different perspectives: compression and prediction. For $n \in \mathbb{N}$, let $\mathcal{P} = \{P_{X^{n+1}|\theta} : \theta \in \Theta\}$ be a collection of joint distributions parameterized by θ.

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

$$\mathsf{Red}(Q_{X^n}) := \sup_{\theta\in\Theta} D(P_{X^n|\theta}\|Q_{X^n}). \quad (2)$$

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

$$\mathsf{Red}_n(\mathcal{P}) := \inf_{Q_{X^n}} \mathsf{Red}(Q_{X^n}), \quad (3)$$

where the infimum is over all joint distribution Q_{X^n}. In universal compression, Q_{X^n} is called a probability assignment, and $D(P_{X^n|\theta}\|Q_{X^n})$ is the excess number of bits compared to the optimal compressor of X^n that knows θ [10].

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

$$\mathsf{Risk}(Q_{X_{n+1}|X^n}) := \sup_{\theta\in\Theta} D(P_{X_{n+1}|X^n,\theta}\|Q_{X_{n+1}|X^n}|P_{X^n|\theta}), \quad (4)$$

where the conditional KL divergence is defined in (1). The minimax prediction risk is then defined as

$$\mathsf{Risk}_n(\mathcal{P}) := \inf_{Q_{X_{n+1}|X^n}} \mathsf{Risk}_n(Q_{X_{n+1}|X^n}), \quad (5)$$

While (3) does not directly correspond to a statistical estimation problem, (5) is exactly the familiar setting of "density estimation", where $Q_{X_{n+1}|X^n}$ is understood as an estimator for the distribution of the unseen X_{n+1} based on the available data $X_1, \ldots, X_n$.

Next we introduce some basic properties of redundancy and prediction risk. First, by using $Q_{X^{n+1}} = \prod_{t=0}^{n} Q_{X_{t+1}|X^t}$, the chain rule of KL divergence tells that

$$\mathsf{Red}_{n+1}(\mathcal{P}) \le \sum_{t=0}^{n} \mathsf{Risk}_t(\mathcal{P}). \quad (6)$$

This inequality is known as the *compression-prediction inequality*.

Second, both quantities admit equivalent formulations in the Bayesian setting, where θ is drawn from a prior $\pi(\theta)$. By replacing $\sup_{\theta\in\Theta}$ in (2), (4) by an average with respect to the prior π, the Bayes redundancy and prediction risk are $I_\pi(\theta; X^n)$ and $I_\pi(\theta; X_{n+1}|X^n)$, respectively, with the Bayes compressor/predictor

$$Q_{X^n}^{\mathrm{Bayes}} = \int_\Theta P_{X^n|\theta}\pi(\mathrm{d}\theta), \quad (7)$$

$$Q_{X_{n+1}|X^n}^{\mathrm{Bayes}} = \frac{\int_\Theta P_{X^{n+1}|\theta}\pi(\mathrm{d}\theta)}{\int_\Theta P_{X^n|\theta}\pi(\mathrm{d}\theta)}. \quad (8)$$

Next, taking the supremum over prior π gives the following dual representations of redun-

dancy and prediction risk.

Theorem II.1. *In general,*

$$\mathsf{Red}_n(\mathcal{P}) = \sup_{\pi} I_\pi(\theta; X^n), \tag{9}$$

where the supremum is over all distributions (priors) $\pi(\theta)$ on Θ. In addition, if $|\mathcal{X}| < \infty$,

$$\mathsf{Risk}_n(\mathcal{P}) = \sup_{\pi} I_\pi(\theta; X_{n+1}|X^n). \tag{10}$$

The identity (9) is a general result known as the *capacity-redundancy theorem* [11]. The identity (10) follows from the minimax theorem; the condition $|\mathcal{X}| < \infty$ guarantees the existence of regular conditional probabilities and the weak compactness of $Q_{X_{n+1}|X^n}$. See [5, Lemma 34] for the details of the proof.

B. A simple example

When $\mathcal{P}$ is the class of iid distributions over a finite alphabet $[k]$, the following upper bounds hold for the redundancy and prediction risk.

Theorem II.2. *For fixed $k \geq 2$,*

$$\mathsf{Red}_n(\mathcal{P}) \leq \frac{k-1}{2}\log n + O(1), \tag{11}$$

$$\mathsf{Risk}_n(\mathcal{P}) \leq \frac{k-1}{n+1}. \tag{12}$$

The upper bound (11) is obtained in [7], with a tight leading term for fixed k by Example 4. The upper bound (12) can be improved to $\frac{k-1}{2n}(1+o(1))$ using an involved predictor [12], and this is also tight by classical asymptotics [13, Chapter 8]. Therefore, for iid discrete distributions, the compression-prediction inequality (6) is essentially tight. For completeness we present a self-contained proof of (11) and (12).

Proof. The redundancy upper bound (11) can be established by an add-$\frac{1}{2}$ rule, also known as the Krichevsky–Trifomov probability assignment [14]. For a sequence $x^t \in [k]^t$ and $i \in [k]$, let $n_i(x^t)$ be the number of appearances of i's in x^t. Let $\Gamma(\cdot)$ be the Gamma function, and

$$\begin{aligned} Q_{X^n}(x^n) &= \frac{1}{k}\prod_{t=1}^{n-1}\frac{n_{x_{t+1}}(x^t)+\frac{1}{2}}{t+\frac{k}{2}} \\ &= \frac{\Gamma(\frac{k}{2})}{2\Gamma(n+\frac{k}{2})}\prod_{i=1}^{k}\frac{\Gamma(n_i(x^n)+\frac{1}{2})}{\Gamma(\frac{1}{2})}. \end{aligned}$$

To upper bound the redundancy, note that

$$\sup_{\theta\in\Theta} P_{X^n|\theta}(x^n) = \prod_{i=1}^{k}\Big(\frac{n_i(x^n)}{n}\Big)^{n_i(x^n)}.$$

Using Stirling's approximation $|\log\Gamma(x+\frac{1}{2}) - (x\log x + x)| \leq C$ for $x \geq 0$, we obtain

$$\begin{aligned} &\log\frac{P_{X^n|\theta}(x^n)}{Q_{X^n}(x^n)} \leq O(1) \\ &+ (n+\frac{k-1}{2})\log(n+\frac{k-1}{2}) - n\log n \\ &\leq \frac{k-1}{2}\log n + O(1). \end{aligned}$$

Since this likelihood ratio bound holds for all $\theta \in \Theta$ and $x^n \in [k]^n$, this choice Q_{X^n} attains the claimed redundancy bound.

The prediction risk upper bound (12) can be established using a similar add-1 rule:

$$Q_{X_{n+1}=i|X^n} = \frac{n_i(X^n)+1}{n+k}, \quad i \in [k].$$

Let the true pmf be $P = (p_1, \ldots, p_k)$, then the prediction risk of the add-1 rule is

$$\begin{aligned} \mathbb{E}[D(P\|Q_{X_{n+1}|X^n})] &= \mathbb{E}\Big[\sum_{i=1}^{k} p_i \log\frac{p_i}{\frac{n_i(X^n)+1}{n+k}}\Big] \\ &\leq \log\mathbb{E}\Big[\sum_{i=1}^{k}\frac{(n+k)p_i^2}{n_i(X^n)+1}\Big], \end{aligned}$$

by applying Jensen's inequality twice. Since for $X \sim \mathsf{B}(n,p)$, $\mathbb{E}[\frac{1}{X+1}] = \frac{1-(1-p)^{n+1}}{(n+1)p} \leq \frac{1}{(n+1)p}$, the above quantity is further upper bounded by

$$\log\sum_{i=1}^{k}\frac{(n+k)p_i}{n+1} = \log\frac{n+k}{n+1} \leq \frac{k-1}{n+1}.$$

This is the upper bound (12). □

C. Redundancy upper bounds

In this section we review some existing procedures for upper bounding the redundancy. We begin with the iid case where $P_{X^n|\theta} = P_\theta^{\otimes n}$. We will use a shorthand $\mathcal{P} = \mathcal{P}_0^{\otimes n}$.

Definition II.3 (KL covering number)**.** *For a family $\mathcal{P}_0$ of distributions, $(P_1, \ldots, P_N)$ is called an* ε-cover *under the KL divergence iff*

$$\sup_{P\in\mathcal{P}_0}\min_{i\in[N]} D(P\|P_i) \leq \varepsilon^2.$$

The smallest integer N such that an ε-cover of size N exists is called the KL covering number of $\mathcal{P}$, or $N_{\mathrm{KL}}(\mathcal{P}_0, \varepsilon)$ in short.

Using the KL covering number, an entropic upper bound of $\mathsf{Red}_n(\mathcal{P}_0^{\otimes n})$ for iid families is known [15].

Theorem II.4.

$$\mathsf{Red}_n(\mathcal{P}_0^{\otimes n}) \le \inf_{\varepsilon>0} \Big(\log N_{\mathrm{KL}}(\mathcal{P}_0, \varepsilon) + n\varepsilon^2 \Big).$$

The proof of Theorem II.4 is simple: for an ε-cover $P_1, \dots, P_N$, the average distribution $Q_{X^n} = \frac{1}{N}\sum_{i=1}^N P_i^{\otimes n}$ attains this bound.

Example 1. *When $\mathcal{P}_0$ is a distribution family with $\log N_{\mathrm{KL}}(\mathcal{P}_0, \varepsilon) \sim d \log \frac{1}{\varepsilon}$, Theorem II.4 gives $\mathsf{Red}_n(\mathcal{P}_0^{\otimes n}) \le \frac{d}{2}\log n + O(1)$. This is the standard scaling of redundancy, where d is the "degree of freedom" of the family.*

For general (non-iid) families, a more combinatorial argument is usually used to upper bound the redundancy. For instance, the analysis in Theorem II.2 upper bounds a pointwise likelihood ratio rather than the KL divergence. This leads to the definition of *pointwise/worst-case* redundancy:

$$R^\star(\mathcal{P}) = \inf_{Q_{X^n}} \sup_{\theta\in\Theta} \sup_{x^n \in \mathcal{X}^n} \log \frac{P_{X^n|\theta}(x^n)}{Q_{X^n}(x^n)}. \quad (13)$$

Clearly, $\mathsf{Red}_n(\mathcal{P}) \le R^\star(\mathcal{P})$. It turns out that $R^\star(\mathcal{P})$ admits an explicit expression [16].

Theorem II.5. *For discrete $\mathcal{X}$,*

$$R^\star(\mathcal{P}) = \log Z := \log \Big(\sum_{x^n \in \mathcal{X}^n} \sup_{\theta\in\Theta} P_{X^n|\theta}(x^n) \Big),$$

where Z is called the Shtarkov sum. *The minimizing distribution $Q_{X^n}(x^n)$ is the normalized maximum likelihood (NML) distribution $Q_{X^n}(x^n) = Z^{-1} \sup_{\theta\in\Theta} P_{X^n|\theta}(x^n)$.*

By Theorem II.5, for $R^\star(\mathcal{P})$ we only need to compute the Shtarkov sum, usually via a combinatorial counting argument. The quantities $\mathsf{Red}_n(\mathcal{P})$ and $R^\star(\mathcal{P})$ usually differ in lower-order terms; see [17] for the simplex case and a recent work [18] for the Gaussian case. Even if we do not choose to evaluate the exact Shtarkov sum, the idea of shifting to $R^\star(\mathcal{P})$ is often helpful for general $\mathcal{P}$ with dependence.

Example 2. *Let $\mathcal{P}$ be the family of time-homogeneous Markov chains on the state space $[k]$, i.e., $P_{X^n}(x^n) = p_1(x_1)\prod_{t=1}^{n-1} M(x_t|x_{t-1})$, with initial distribution p_1 and transition matrix M. In this case, by applying the Krichevsky–Trifomov probability assignment to each symbol succeeding $j \in [k]$, we obtain kernels $Q_{X_{t+1}|X^t}$ such that*

$$\sup_M \sup_{x^n} \log \prod_{t\in[n-1]:x_t=j} \frac{M(x_{t+1}|j)}{Q_{X_{t+1}|X^t}(x_{t+1}|x^t)} \le \frac{k-1}{2}\log n + O(1).$$

Assigning $Q_{X_1}(x_1) = \frac{1}{k}$ and summing over $j \in [k]$, we obtain

$$\mathsf{Red}_n(\mathcal{P}) \le R^\star(\mathcal{P}) \le \frac{k(k-1)}{2}\log n + O(1).$$

Again, $k(k-1)$ is the number of free parameters in the transition matrix M.

Similar ideas can be applied to other time series like higher-order Markov chains or hidden Markov models. We also note that both the mixing distribution $Q_{X^n} = \frac{1}{N}\sum_{i=1}^N P_i^{\otimes n}$ and the NML distribution can be hard to compute; a notable example of computationally efficient probability assignments Q_{X^n} is the context tree weighting (CTW) method. We refer to the monograph [19] on these topics.

D. Redundancy lower bounds

In this section we survey some lower bound techniques for the redundancy, and once again start with iid models $\mathcal{P} = \mathcal{P}_0^{\otimes n}$.

Definition II.6 (Hellinger packing number)**.** *For a sequence of distributions $P_1, \dots, P_M \in \mathcal{P}_0$, it is called an ε-*packing *under the Hellinger distance iff*

$$\min_{i\ne j\in[M]} H(P_i, P_j) \ge \varepsilon.$$

The smallest such integer M is the Hellinger packing number, denoted by $M_{\mathrm{H}}(\mathcal{P}_0, \varepsilon)$.

The following result, due to [15], provides an entropic lower bound of the redundancy.

Theorem II.7.

$$\mathsf{Red}_n(\mathcal{P}_0^{\otimes n}) \geq \sup_{\varepsilon>0} \min\left\{ \log M_{\mathrm{H}}(\mathcal{P}_0,\varepsilon), \frac{n\varepsilon^2}{2}\right\} - \log 2.$$

Example 3. *For iid models $\mathcal{P} = \mathcal{P}_0^{\otimes n}$ with d degrees of freedom, one typically has $\log M_{\mathrm{H}}(\mathcal{P}_0,\varepsilon) \sim d\log\frac{1}{\varepsilon}$. Then Theorem II.7 shows that $\mathsf{Red}_n(\mathcal{P}_0^{\otimes n}) \geq \frac{d}{2}\log(\frac{n}{\log n}) - O(1)$.*

Next we move on to general families and describe a useful lower bound program due to Rissanen [8]. The key of this program is that "good estimation implies high redundancy".

Theorem II.8. *Suppose $\Theta \subseteq \mathbb{R}^d$ has a non-empty interior, and there exists an estimator $\widehat{\theta}(X^n)$ such that*

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\widehat{\theta}(X^n) - \theta\|_2^2] \leq \varepsilon_n^2.$$

Then

$$\mathsf{Red}_n(\mathcal{P}) \geq \log \mathrm{Vol}_d(\Theta) - \frac{d}{2}\log\Big(\frac{2\pi e\varepsilon_n^2}{d}\Big).$$

Proof. Let $\theta \sim \mathrm{Unif}(\Theta)$, and $h(\cdot)$ denote the differential entropy on $\mathbb{R}^d$. Then

$$I(\theta; X^n) = h(\theta) - h(\theta|X^n),$$

where $h(\theta) = \log \mathrm{Vol}_d(\Theta)$. By the Gaussian maximum entropy principle,

$$\begin{aligned} h(\theta|X^n) &= h(\theta - \widehat{\theta}(X^n)|X^n) \\ &\leq h(\theta - \widehat{\theta}(X^n)) \\ &\leq \frac{d}{2}\log\Big(2\pi e\cdot\frac{\mathbb{E}_\theta[\|\widehat{\theta}(X^n)-\theta\|_2^2]}{d}\Big) \\ &\leq \frac{d}{2}\log\Big(\frac{2\pi e\varepsilon_n^2}{d}\Big). \end{aligned}$$

The rest follows from Theorem II.1. □

Example 4. *For the iid discrete distribution model in Theorem II.2, choosing $\widehat{\theta}(X^n)$ being the empirical distribution gives $\varepsilon_n^2 = \frac{1}{n}$. In addition, the volume of a $(k-1)$-dimensional probability simplex is $\frac{\sqrt{k}}{(k-1)!}$, a constant. Therefore, Theorem II.8 gives a redundancy lower bound of $\frac{k-1}{2}\log n - O(1)$.*

Rissanen's program is a powerful technique to deal with models with dependence, as long as one can find a sufficiently large submodel $\Theta_0 \subseteq \Theta$ and an estimator that works well under this submodel. An example is the Markov chain in Example 2, where a redundancy lower bound of $\Omega(k^2\log\frac{n}{k^2})$ is shown for all $k \in [2, c\sqrt{n}]$ using Rissanen's program [20].

III. Optimal prediction risk in Markov models

While there are several techniques for upper and lower bounding the redundancy of families with dependence, much less is known about the prediction risk. In this section, we review some recent progresses on the optimal prediction risk of Markov chains.

Consider a time-homogeneous Markov chain $P_{X^n}(x^n) = \pi(x_1)\prod_{t=1}^{n-1} M(x_{t+1}|x_t)$ over a finite state space $[k]$. We assume that the Markov chain is stationary, i.e., $\pi M = \pi$. The optimal prediction risk (5) is then

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

This prediction problem is a non-standard statistical problem: it is distinct from the parameter estimation problem such as estimating the transition matrix M, in that the quantity to be estimated (conditional distribution of the next state) depends on the sample path itself. As a result, this formulation allows more flexibility with far less assumptions compared to the estimation framework. For example, if certain state has very small probability under the stationary distribution, consistent estimation of the transition matrix with respect to usual loss function, e.g. squared risk, may not be possible, whereas the prediction problem is unencumbered by such rare states. This brings the central question of this section: *is prediction possible even when estimation is impossible?*

A. Characterization of optimal prediction risk

For prediction in Markov chains, surprising phenomena already arise even in the binary case $k = 2$.

Theorem III.1 (Binary chain $k = 2$; [21])**.**

$$\mathsf{Risk}_{2,n} \asymp \frac{\log\log n}{n}.$$

The main feature of Theorem III.1 is that the optimal prediction risk is strictly larger than the parametric rate $\Theta(\frac{1}{n})$. Compared with the redundancy bound $\mathsf{Red}_{2,n} = O(\log n)$ in Example 2, the compression-prediction inequality (6) is no longer tight. In other words, optimal prediction does not imply optimal compression even in binary Markov chains.

Although we will not prove Theorem III.1, below we describe the estimator that attains the optimal rate. Based on the trajectory $X^n \in \{0,1\}^n$, let $N_i = \sum_{t=1}^{n-1} \mathbf{1}_{\{X_t=i\}}$, and $N_{ij} = \sum_{t=1}^{n-1} \mathbf{1}_{\{X_t=i,X_{t+1}=j\}}$. The estimator is

$$\widehat{M}(0|X^n) = \begin{cases} \frac{N_{X_n,0}+1}{N_{X_n}+2} & \text{if } X^n \notin \mathcal{L}, \\ \frac{N_{X_n,0}+\alpha}{N_{X_n}+2\alpha} & \text{if } X^n \in \mathcal{L}, \end{cases}$$

and $\widehat{M}(1|X^n) = 1 - \widehat{M}(0|X^n)$, where $\alpha \asymp (\log n)^{-1}$, and $\mathcal{L}$ is the collection of "lazy" trajectories taking the form $0\cdots01\cdots1$ or $1\cdots10\cdots0$. In other words, the same add-1 estimator is applied to most trajectories, while for lazy trajectories a different add-α estimator is used. To gain some intuition on the choice of α, consider a slow-mixing chain with $M(0|0) = M(1|1) = 1 - \frac{1}{n}$ and $M(0|1) = M(1|0) = \frac{1}{n}$. In this case, each of the trajectory in $\mathcal{L}$ occurs with probability $\Omega(\frac{1}{n})$, and the risk of an add-α estimator on these trajectories is at least

$$\Omega(\frac{1}{n}) \cdot \sum_{t=1}^{n-1} D\Big((\frac{1}{n}, 1-\frac{1}{n}) \| (\frac{\alpha}{t+2\alpha}, \frac{t+\alpha}{t+2\alpha})\Big)$$
$$= \Omega\Big(\frac{\log(\frac{1}{\alpha}) + \alpha\log n}{n}\Big)$$

after some algebra. Therefore, the choice $\alpha \asymp (\log n)^{-1}$ attains the optimal risk $O(\frac{\log\log n}{n})$, even restricting only to the trajectories in $\mathcal{L}$.

At a high level, the above computation shows that the prediction risk is strictly larger than the parametric rate due to slow-mixing Markov chains. On the other hand, consistent prediction remains possible even for these slow-mixing chains. These intuitions continue to hold for general Markov chains with state space $k \geq 3$, though with a different optimal rate.

Theorem III.2 (General chain $k \geq 3$; [4], [5])**.** *For $3 \leq k \leq c\sqrt{n}$ with some universal $c > 0$,*

$$\mathsf{Risk}_{k,n} \asymp \frac{k^2}{n} \log \frac{n}{k^2}.$$

Again, for constant $k \geq 3$, the optimal prediction risk exceeds the parametric rate $\Theta(\frac{k^2}{n})$ only by a logarithmic factor, without requiring mixing conditions of the chain. However, when moving from $k = 2$ to $k = 3$, this logarithmic factor increases from $\log\log n$ to $\log n$. The proof of Theorem III.2 is presented in the next few sections. In particular, we show that although the compression-prediction inequality (6) is not tight, the optimal prediction risk in Theorem III.2 can nevertheless be derived from redundancy in both the upper and lower bounds.

B. Proof of upper bound

The upper bound of Theorem III.2 relies on the following inequality between the redundancy and prediction risk. In iid models, this type of reduction relating cumulative risks and individual risks is known as *online-to-batch conversion* dating back to [22].

Lemma III.3. *Suppose each $P_{X^{n+1}|\theta} \in \mathcal{P}$ is stationary and m^{th}-order Markov. For any joint distribution $Q_{X^{n+1}} = \prod_{t=1}^{n+1} Q_{X_t|X^{t-1}}$, the Cesàro-mean-type predictor*

$$\widetilde{Q}_{X_{n+1}|X^n}(x_{n+1}|x^n) \tag{14}$$
$$= \frac{1}{n+1-m} \sum_{t=m+1}^{n+1} Q_{X_t|X^{t-1}}(x_{n+1}|x_{n+2-t}^n)$$

atttains the prediction risk

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

Proof. Based on $Q_{X^{n+1}}$, denote the t-th estimator by $\widehat{P}_t(\cdot|x^{t-1}) = Q_{X_t|X^{t-1}=x^{t-1}}$, then

$$\widetilde{Q}_{X_{n+1}|X^n=x^n} = \mathbb{E}_t[\widehat{P}_t(\cdot|x_{n+2-t}^n)],$$

where we use the short-hand $\mathbb{E}_t$ to denote the average of $t = m+1, \ldots, n+1$. In other words, we apply $\widehat{P}_t$ to the most recent $t-1$ symbols prior to X_{n+1} for predicting its distribution,

then average over t. To analyze its prediction risk, for any $P_{X^{n+1}} \in \mathcal{P}$,

$$\begin{aligned} & D(P_{X_{n+1}|X^n} \| \widetilde{Q}_{X_{n+1}|X^n} | P_{X^n}) \\ & \overset{\text{(a)}}{\le} \mathbb{E}_t \mathbb{E}\Big[D(P_{X_{n+1}|X^n} \| \widehat{P}_t(\cdot|X_{n+2-t}^n)) \Big] \\ & \overset{\text{(b)}}{=} \mathbb{E}_t \mathbb{E}\Big[D(P_{X_{n+1}|X_{n+1-m}^n} \| \widehat{P}_t(\cdot|X_{n+2-t}^n)) \Big] \\ & \overset{\text{(c)}}{=} \mathbb{E}_t \mathbb{E}\Big[D(P_{X_t|X_{t-m}^{t-1}} \| \widehat{P}_t(\cdot|X^{t-1})) \Big] \\ & \overset{\text{(d)}}{\le} \frac{\mathsf{Red}(Q_{X^{n+1}})}{n-m+1}. \end{aligned}$$

Here (a) uses the convexity of the KL divergence, (b) follows from m^{th}-order Markovity, (c) is due to stationarity, and (d) uses the chain rule of KL divergence. □

Specializing Lemma III.3 to the Markov case $m = 1$ gives the following corollary:

Corollary III.4. *For Markov chains,*

$$\mathsf{Risk}_{k,n} \le \frac{\mathsf{Red}_{k,n+1}}{n}.$$

Using the redundancy upper bound (shown via unfolding the k-dependence in Example 2)

$$\mathsf{Red}_{k,n+1} = O\Big(\frac{k^2}{n} \log \frac{n}{k^2}\Big),$$

the prediction risk upper bound in Theorem III.2 follows from Corollary III.4. In addition, Lemma III.3 implies a computationally efficient predictor: let $\widehat{M}^{+1}(x_{n+1}|x^n)$ be the add-1 estimator under the trajectory x^n, our predictor is an average of add-1 estimators:

$$\widehat{M}^{\star}(x_{n+1}|x^n) := \frac{1}{n} \sum_{t=1}^{n} \widehat{M}^{+1}(x_{n+1}|x_{n-t+1}^n).$$

Finally, in comparison with the usual statistical analysis of Markov chains, we remark that the key advantage of our approach via Lemma III.3 is that it does not rely on mixing properties. At the heart, this is because the worst-case redundancy, an upper bound on Red obtained by replacing $\mathbb{E}$ with $\max$, becomes a purely combinatorial quantity. In contrast, classical statistical approaches work directly with $\mathbb{E}$, where mixing properties are often required to make the underlying distribution tractable.

Remark III.5. *For m^{th}-order Markov chains, the same program shows that the optimal prediction risk is $O(\frac{k^m}{n} \log \frac{n}{k^m})$ if $2 \le k \le c_m n^{1/m}$. This is also shown to be tight [5].*

C. Proof of lower bound

Although the compression-prediction inequality (6) seems to suggest a prediction risk lower bound, since $t \mapsto \mathsf{Risk}_t$ is non-increasing, it does not imply a lower bound of an individual term Risk_n. Interestingly, our lower bound of Theorem III.2 relies on a different reduction from prediction risk to redundancy.

Lemma III.6. *Let $\mathcal{P}_k^{\mathsf{sym}}$ be the family of all stationary and* symmetric *Markov chains over the state space $[k]$, then*

$$\mathsf{Risk}_{k,n} \ge \frac{1}{4en}\Big(\mathsf{Red}_n(\mathcal{P}_{k-1}^{\mathsf{sym}}) - \log(k-1)\Big).$$

Proof. The crux of the proof is to embed any symmetric Markov chain on $k-1$ states into a k-state Markov chain, by adding a "lazy" state. Specifically, let T be the transition matrix of a general symmetric Markov chain on $[k-1]$, we construct the overall transition matrix as

$$M = \begin{bmatrix} 1-\frac{1}{n} & \frac{1}{n(k-1)} & \cdots & \frac{1}{n(k-1)} \\ 1/n & & & \\ \vdots & & \big(1-\frac{1}{n}\big)\, T & \\ 1/n & & & \end{bmatrix}. \quad (15)$$

In other words, the "lazy" state 1 has a heavy self-loop and the probability of leaving it scales as $\frac{1}{n}$. For the other states, the transition matrix T is unknown and the parameter of interest. An example with $k = 3$ is displayed in Figure 1, where state 1 is the "lazy" state, and the transition probability p between states 2 and 3 is the parameter of interest.

Next let T follow a generic prior over $\mathcal{P}_{k-1}^{\mathsf{sym}}$. For $t \in [n-1]$, let $\mathcal{X}_t$ be the set of trajectories x^n with $x_s \equiv 1$ for all $s \le t$ and $x_s \ne 1$ for all $s \ge t+1$. On a trajectory $x^n \in \mathcal{X}_t$, by (8) the Bayes estimator for $M(j|x_n)$ with $j \ne 1$ is

$$\begin{aligned} \widehat{M}(j|x^n) &= \frac{\mathbb{P}(X^{n+1} = (x^n, j))}{\mathbb{P}(X^n = x^n)} \\ &= \Big(1-\frac{1}{n}\Big) \widehat{T}_{n-t}(j|x_{t+1}^n). \end{aligned}$$

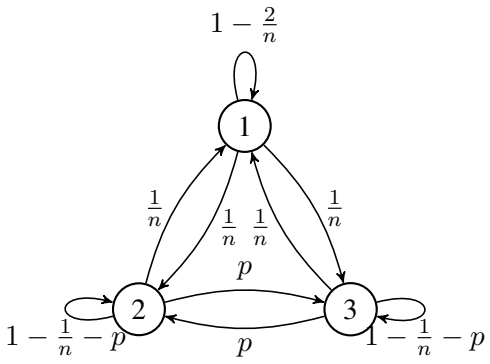

Fig. 1: Lower bound construction for three-state Markov chains, with state 1 being the lazy state.

Here $\widehat{T}_t(\cdot|y^t)$ is the Bayes estimator of $T(\cdot|y_t)$ given the length-t trajectory y^t from the $(k-1)$-state chain (call it Y-chain), and the second identity follows from the construction of M in (15) and algebra. Since $M(1|x_n) \equiv \frac{1}{n}$ is deterministic for $x_n \neq 1$, we obtain

$$\widehat{M}(\cdot|x^n) = \frac{1}{n}\delta_1 + \Big(1-\frac{1}{n}\Big)\widehat{T}_{n-t}(\cdot|x_{t+1}^n)$$

for all $x^n \in \mathcal{X}_t$. This parallels with the construction of M that

$$M(\cdot|x) = \frac{1}{n}\delta_1 + \Big(1-\frac{1}{n}\Big)T(\cdot|x),\ x \neq 1.$$

In other words, estimation of M on a trajectory $x^n \in \mathcal{X}_t$ is equivalent to the estimation of T on a shorter trajectory x_{t+1}^n. In particular, if Y^{n-t} is generated from the Y-chain, this shows that

$$\begin{aligned}
&\mathbb{E}_T\Big[\mathbb{E}[D(M(\cdot|X^n)\|\widehat{M}(\cdot|X^n))|X^n \in \mathcal{X}_t]\Big] \\
&= \Big(1-\frac{1}{n}\Big)\mathbb{E}_T\Big[\mathbb{E}[D(T(\cdot|Y_{n-t})\|\widehat{T}(\cdot|Y^{n-t}))]\Big] \\
&= \Big(1-\frac{1}{n}\Big)I(T;Y_{n-t+1}|Y^{n-t}). \qquad (16)
\end{aligned}$$

The key observation now is that each trajectory class $\mathcal{X}_t$ occurs with a large probability:

$$\mathbb{P}(\mathcal{X}_t) = \frac{1}{2}\Big(1-\frac{1}{n}\Big)^{(t-1)+(n-1-t)} \cdot \frac{1}{n} \geq \frac{1}{2en}.$$

Combining with (16),

$$\begin{aligned}
&\mathbb{E}_T\Big[\mathbb{E}[D(M(\cdot|X^n)\|\widehat{M}(\cdot|X^n))]\Big] \\
&\geq \frac{1}{2en}\Big(1-\frac{1}{n}\Big)\sum_{t=1}^{n-1} I(T;Y_{n-t+1}|Y^{n-t}) \\
&\geq \frac{1}{4en}\Big(I(T;Y^n) - I(T;Y_1)\Big).
\end{aligned}$$

The proof is completed by the dual representation of redundancy in Theorem II.1, and that $I(T;Y_1) \leq H(Y_1) \leq \log(k-1)$. □

At a high level, in our embedding (15), the effective number of observations on other states is roughly *uniformly distributed* on $[n]$. Therefore, $\mathsf{Risk}_{k,n}$ is lower bounded by the average prediction risks on a smaller chain with sample size $\sim \mathrm{Unif}([n])$, which by the compression-prediction inequality (6) is lower bounded by the redundancy of the smaller chain. Finally, to complete the proof of lower bound in Theorem III.2, we use Rissanen's program (Theorem II.8) in a similar way to [20] to obtain a redundancy lower bound $\mathsf{Red}_n(\mathcal{P}_{k-1}^{\mathsf{sym}}) = \Omega(\frac{k^2}{n}\log\frac{n}{k^2})$.

Remark III.7. *This embedding critically relies on $k \geq 3$, for the chain T is trivial when $k = 2$. This explains the technical distinction between the cases $k = 2$ and $k \geq 3$.*

D. Role of mixing condition

Although the previous program of prediction via compression yields optimal prediction risks for general Markov chains, we note that an improved rate is possible with favorable mixing conditions while not captured by this program.

Specifically, we focus on the class of irreducible and reversible Markov chains. It is well-known that for such chains, the transition matrix M has k real eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq -1$. The *absolute spectral gap* of M, defined as

$$\gamma^\star(M) := 1 - \max\{|\lambda_i| : 2 \leq i \leq k\}, \qquad (17)$$

quantifies the memory of the Markov chain. For example, the mixing time is determined by $1/\gamma^\star$ (relaxation time) up to logarithmic factors. To model the class of Markov chains with favorable mixing property, let $\mathsf{Risk}_{k,n}(\gamma_0)$ be the optimal prediction risk in (5) restricted to all stationary, irreducible, and reversible Markov chains on $[k]$ that have absolute spectral gap at least γ_0. Our question is to understand how the optimal prediction risk depends on γ_0.

For binary Markov chains, we have a complete characterization of $\mathsf{Risk}_{2,n}(\gamma_0)$ [5].

Theorem III.8. *For any* $\gamma_0 \in (0,1)$,

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

where $\log_+(x) := \max\{1, \log(x)\}$.

Theorem III.8 shows that the parametric rate $\Theta(\frac{1}{n})$ is attainable in binary Markov chains if and only if there is a constant spectral gap $\gamma_0 = \Omega(1)$; below this threshold, the prediction risk is strictly larger than $\Theta(\frac{1}{n})$, and saturates at $\Theta(\frac{\log\log n}{n})$ in Theorem III.1 even if $\gamma_0 = 0$. For the general case $k \geq 3$, only a sufficient condition for achieving the parametric rate $O(\frac{k^2}{n})$ is known.

Theorem III.9. *If* $k \gtrsim \log n \log\log n$ *and* $\gamma_0 \gtrsim \frac{\log^2 n}{k}$, *then* $\mathsf{Risk}_{k,n}(\gamma_0) = O(\frac{k^2}{n})$.

For large k, Theorem III.9 shows that the parametric rate $O(\frac{k^2}{n})$ remains attainable even as the spectral gap shrinks. Remarkably, the predictor achieving this risk is extremely simple: the add-1 estimator $\widehat{M}^{+1}(\cdot|x^n)$. The proof of Theorem III.9 is primarily statistical, and uses a recent KL concentration result [23] for the add-1 estimator in place of [5, Lemma 17].

Finally, we note that, in contrast to prediction risk, redundancy is insensitive to the spectral gap. Indeed, the redundancy of the class of Markov chains with an $\Omega(1)$ spectral gap remains $\Theta(\frac{k^2}{n}\log\frac{n}{k^2})$, shown by the lower bound in [20]. Therefore, the program of prediction via compression ceases to be effective in the presence of mixing conditions.

IV. Optimal prediction risk in models with infinite memory

In this section, we move from simple Markov chains to general stochastic processes with possibly infinite memory. For such processes, our online-to-batch conversion in Lemma III.3 may no longer work as it requires a bounded length of memory. A natural solution is to approximate the time series by an m^{th}-order Markov chain; however, we usually need to take $m \to \infty$ to achieve a small approximation error, while the optimal prediction risk for m^{th}-order Markov chains is $\Omega(\frac{k^m}{n}\log\frac{n}{k^m})$ (cf. Remark III.5). As a result, such approaches (like the one in [24]) only achieve an extremely slow rate, like $O(\frac{1}{\log n})$, for prediction. As we will show in this section, such approaches are highly suboptimal.

A. A general online-to-batch conversion

In this section, we generalize the online-to-batch conversion in Lemma III.3 from Markov processes to general stationary processes. The next result shows that, the expected relationship $\mathsf{Risk}_n \leq \frac{\mathsf{Red}_n}{n}$ holds up to a "memory" term.

Lemma IV.1. *Suppose each* $P_{X^{n+1}|\theta} \in \mathcal{P}$ *is stationary. For any joint distribution* $Q_{X^{n+1}} = \prod_{t=0}^n Q_{X_{t+1}|X^t}$, *the predictor*

$$\widetilde{Q}_{X_{n+1}|X^n}(x_{n+1}|x^n) = \frac{1}{n+1}\sum_{t=0}^n Q_{X_{t+1}|X^t}(x_{n+1}|x_{n-t+1}^n)$$

atttains the prediction risk

$$\mathsf{Risk}(\widetilde{Q}_{X_{n+1}|X^n}) \leq \frac{\mathsf{Red}(Q_{X^{n+1}})}{n+1} + \mathsf{Mem}_{n+1}(\mathcal{P}),$$

where the memory term is defined as

$$\mathsf{Mem}_{n+1}(\mathcal{P}) := \sup_{P_{X^{n+1}|\theta}\in\mathcal{P}} \frac{1}{n+1}\sum_{t=0}^n I(X_{n+1}; X^{n-t}|X_{n-t+1}^n). \tag{18}$$

Proof. The proof is similar to Lemma III.3. Let $\widehat{P}_t(\cdot|x^t) := Q_{X_{t+1}|X^t=x^t}$ and $\mathbb{E}_t$ denote the average over $t = 0, 1, \ldots, n$, we have

$$\begin{aligned}
&D(P_{X_{n+1}|X^n}\|\widetilde{Q}_{X_{n+1}|X^n}|P_{X^n}) \\
&\overset{\text{(a)}}{\leq} \mathbb{E}_t\mathbb{E}\Big[D(P_{X_{n+1}|X^n}\|\widehat{P}_t(\cdot|X_{n-t+1}^n))\Big] \\
&\overset{\text{(b)}}{=} \mathbb{E}_t\mathbb{E}\Big[D(P_{X_{n+1}|X_{n-t+1}^n}\|\widehat{P}_t(\cdot|X_{n-t+1}^n))\Big] \\
&\quad + \mathbb{E}_t\mathbb{E}\Big[D(P_{X_{n+1}|X^n}\|P_{X_{n+1}|X_{n-t+1}^n})\Big] \\
&\overset{\text{(c)}}{=} \mathbb{E}_t\mathbb{E}\Big[D(P_{X_{t+1}|X^t}\|\widehat{P}_t(\cdot|X^t))\Big] \\
&\quad + \mathbb{E}_t I(X_{n+1}; X^{n-t}|X_{n-t+1}^n) \\
&\overset{\text{(d)}}{\leq} \frac{\mathsf{Red}(Q_{X^{n+1}})}{n+1} + \mathsf{Mem}_{n+1}(\mathcal{P}).
\end{aligned}$$

Here (a) uses the convexity of KL divergence, (b) follows from simple algebra, (c) is due to stationarity and the definition of the conditional mutual information, and (d) uses the chain rule of KL divergence and the definition (18). □

We verify that the memory term is indeed negligible in Markov models. Indeed, for m^{th}-order Markov chains, we have $I(X_{n+1}; X^{n-t}|X_{n-t+1}^n) = 0$ for $t \geq m$. Therefore, $\mathsf{Mem}_{n+1} \leq \frac{m \log k}{n+1}$, which is negligible compared to the target quantity $O(\frac{k^m}{n} \log \frac{n}{k^m})$. In the next section, we will see that this memory term remains small for many natural time series with even infinite memory.

B. Applications: HMMs and renewal processes

A natural time series with infinite memory is the hidden Markov model (HMM), a useful tool for modeling practical data such as natural language and speech signals. An HMM is obtained by passing a Markov process through a memoryless noisy channel. Specifically, fix $k, \ell \in \mathbb{N}$. Let $\{Z_t\}_{t\geq 1}$ be a stationary Markov chain on the state space $[k]$ with transition matrix $M \in \mathbb{R}^{k\times k}$. Let $T \in \mathbb{R}^{k\times \ell}$ denote a probability transition kernel from $[k]$ to $[\ell]$. Let $\{X_t\}_{t\geq 1}$ be an $[\ell]$-valued process such that $P_{X^n|Z^n} = \prod_{t=1}^n T(x_t|z_t)$. We refer to $\{X_t\}_{t\geq 1}$ as a hidden Markov process with *transition probabilities* M and *emission probabilities* T, while $\{Z_t\}_{t\geq 1}$ are called the *hidden states*. It is easy to verify that HMMs have infinite memory, i.e., $I(X_{t+1}; X_1|X_2^t) \neq 0$ for all $t \geq 1$.

Let $\mathsf{Risk}_{k,\ell,n}$ be the optimal prediction risk in the above family of HMMs. Using Lemma IV.1, we obtain the following bound.

Theorem IV.2 ([6]). *For $n \geq k(k+\ell)$,*

$$\mathsf{Risk}_{k,\ell,n} = O\Big(\frac{k^2}{n} \log \frac{n}{k^2} + \frac{k\ell}{n} \log \frac{n}{k\ell}\Big).$$

If $\ell \geq k$ and $n \geq k\ell$, or $n \geq k^C$ and $k, \ell \geq 2$, this upper bound is also tight.

Proof of upper bound. For the redundancy, since the HMM has $k(k+\ell-2)$ free parameters, a redundancy $O(k^2 \log \frac{n}{k^2} + k\ell \log \frac{n}{k\ell})$ is shown in [19]. By Lemma IV.1, it remains to show

$$\mathsf{Mem}_{n+1} \leq \frac{I(Z_1; X^{n+1})}{n+1} \leq \frac{\log k}{n+1}. \qquad (19)$$

The second inequality of (19) trivially follows from $I(Z_1; X^{n+1}) \leq H(Z_1) \leq \log k$. For the first inequality, note that

$$\begin{aligned} & I(X_{n+1}; X^{n-t}|X_{n-t+1}^n) \\ & \overset{\text{(a)}}{\leq} I(X_{n+1}; Z_{n-t+1}|X_{n-t+1}^n) \\ & \overset{\text{(b)}}{=} I(X_t; Z_1|X^{t-1}), \end{aligned}$$

where (a) applies the data-processing inequality to the Markov chain $X^{n-t} \to Z_{n-t+1} \to X_{n-t+1}^{n+1}$, and (b) is due to stationarity. The chain rule yields the first inequality of (19). □

The above proof shows that, despite the infinite memory of HMM, the memory term in (18) remains small and is controlled by a discrete hidden state. As a result, the optimal prediction risk $O(\frac{\log n}{n})$, obtained via compression, is much better than the upper bound $O(\frac{1}{\log n})$ obtained in [24] via Markov approximations.

Another example of time series with infinite memory is the 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 $X_t = 1$ if and only if $t \in \{T_0, T_0 + T_1, T_0 + T_1 + T_2, \ldots\}$. We refer to T_0 and $\{T_i : t \geq 1\}$ as the initial wait time and the interarrival times.

Let $\mathsf{Risk}_{\mathsf{rwnl},n}$ be the optimal prediction risk in all stationary renewal processes. Lemma IV.1 again leads to a tight characterization.

Theorem IV.3 ([6]). $\mathsf{Risk}_{\mathsf{rwnl},n} = \Theta(\frac{1}{\sqrt{n}})$.

Proof of upper bound. For renewal processes, [25] establishes a well-known redundancy bound $\mathsf{Red}_{\mathsf{rwnl},n} = \Theta(\sqrt{n})$. By Lemma IV.1, it remains to show that $\mathsf{Mem}_{n+1} = O(n^{-1/2})$.

To this end, note that a renewal process with interarrival distribution μ can be represented by an HMM with a countably infinite state space:

1) The hidden states $\{Z_t\} \subseteq \mathbb{N}$ represent 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) Deterministic emissions: $X_t = \mathbf{1}_{\{Z_t=1\}}$.

Under this representation, we may apply (19) to get $\mathsf{Mem}_{n+1} \leq \frac{I(Z_1;X^{n+1})}{n+1}$. Although Z_1 takes infinitely many values, we may take $\widetilde{Z}_1 \triangleq$

$\min\{Z_1, n+2\}$ so that $Z_1 \to \widetilde{Z}_1 \to X^{n+1}$ is a Markov chain. This is 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:

$$\begin{aligned} I(Z_1; X^{n+1}) &\le I(\widetilde{Z}_1; X^{n+1}) \\ &\le H(\widetilde{Z}_1) \le \log(n+2), \end{aligned}$$

so that $\mathsf{Mem}_{n+1} = O(\frac{\log n}{n})$. □

C. Computational barriers

Despite the statistical efficiency of the online-to-batch conversion in Lemma IV.1, such a conversion may encounter computational barriers when achieving the redundancy bound. For instance, unlike the Markov case where the optimal redundancy is attained by the computationally efficient add-$\frac{1}{2}$ rule, the redundancy upper bound for HMMs is established via a combinatorial argument based on the Shtarkov sum in Theorem II.5. Specifically, this argument requires marginalization over all hidden state sequences $Z^n \in [k]^n$. This exponential complexity can be reduced using dynamic programming:

Lemma IV.4 ([6]). *The algorithm for Theorem IV.2 can be implemented in $n^{O(k^2+k\ell)}$ time.*

For small k and ℓ such as $k = \ell = 2$, Lemma IV.4 yields a polynomial-time algorithm. However, computational barriers still exist for moderately large k and ℓ. The next result shows that this barrier is inherent under certain cryptographic hardness assumptions, thereby establishing a statistical-computational gap for the general prediction problem.

Theorem IV.5 ([6]). *For prediction in HMMs, the following hardness results hold under certain cryptographic hardness assumptions:*

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

Theorem IV.5 shows that as long as either k or ℓ is large, no polynomial-time predictor can achieve a reasonable prediction risk. These hardness results rely on cryptographic assumptions, specifically the hardness of learning parity with noise (LPN) and of refuting constrained satisfaction problems (CSPs); see [6] and [24] for details.

V. Conclusion and Outlook

This tutorial develops a statistical and computational foundation for next-symbol prediction, a non-standard statistical task of estimating a random and data dependent distribution. Our information-theoretic framework clarifies the fundamental distinction between prediction and parameter estimation, and shows that prediction can remain possible even when estimation is impossible. By connecting prediction to the classical problem of universal compression through redundancy, we characterize statistically optimal prediction risk for Markov chains and processes with infinite memory without imposing mixing assumptions, and quantify the effects of memory and mixing on statistical efficiency. Beyond statistical limits, we identify inherent computational barriers, showing that statistically optimal prediction may be unattainable by computationally efficient algorithms. Together, these results provide a unified framework for understanding both the possibilities and limitations of prediction in dependent data.

Looking ahead, an important theoretical challenge is to explain the empirical successes of large language models in next-symbol prediction, and identify their potential weaknesses. To this end, one can study a sequence of questions of increasing difficulty. The first concerns the representational power of transformers, namely their ability to implement a broad class of prediction algorithms. When specialized to Markov chains and related models, this question has received considerable recent attention [26]–[28]. The second concerns the prediction risk achieved by the minimizer of the training loss. For example, recent work [29] establishes an upper bound on the in-context KL risk under an implicit mixing condition, but it remains unclear whether similar guarantees hold in the absence of mixing. The third question concerns the dynamics of transformer training, including

how gradient-based optimization learns Markov models [28], [30]. A closely related question is why training overparameterized transformers on simple Markov models does not appear to induce spurious long-range memory that degrades prediction performance. Together, these questions point toward a unified theory that bridges information-theoretic limits, optimization dynamics, and architectural inductive biases in modern large language models.

REFERENCES

[1] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] W. D. Heaven, "Large language models can do jaw-dropping things. but nobody knows exactly why." *MIT Technology Review*, March 2024.

[3] C. E. Shannon, "Prediction and entropy of printed English," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[4] Y. Han, S. Jana, and Y. Wu, "Optimal prediction of Markov chains with and without spectral gap," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 233–11 246, 2021.

[5] ——, "Optimal prediction of Markov chains with and without spectral gap," *IEEE Transactions on Information Theory*, 2023.

[6] Y. Han, T. Jiang, and Y. Wu, "Prediction from compression for models with infinite memory, with applications to hidden markov and renewal processes," in *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, 2024, pp. 2270–2307.

[7] L. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 783–795, 1973.

[8] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information theory*, vol. 30, no. 4, pp. 629–636, 1984.

[9] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory, 2nd Ed.* New York, NY, USA: Wiley-Interscience, 2006.

[11] J. Kemperman, "On the Shannon capacity of an arbitrary channel," in *Indagationes Mathematicae (Proceedings)*, vol. 77, no. 2. North-Holland, 1974, pp. 101–115.

[12] D. Braess and T. Sauer, "Bernstein polynomials and learning theory," *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187–206, 2004.

[13] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[14] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[15] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.

[16] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 3–17, 1987.

[17] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[18] J. Mourtada, "Universal coding, intrinsic volumes, and metric complexity," *Journal of the European Mathematical Society*, 2025.

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

[20] K. Tatwawadi, J. Jiao, and T. Weissman, "Minimax redundancy for Markov chains with large state space," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 216–220.

[21] M. Falahatgar, A. Orlitsky, V. Pichapati, and A. Suresh, "Learning Markov distributions: Does estimation trump compression?" in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, July 2016, pp. 2689–2693.

[22] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, pp. 1564–1599, 1999.

[23] J. Mourtada, "Estimation of discrete distributions in relative entropy, and the deviations of the missing mass," *arXiv preprint arXiv:2504.21787*, 2025.

[24] V. Sharan, S. Kakade, P. Liang, and G. Valiant, "Prediction with a short memory," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018, pp. 1074–1087.

[25] I. Csiszár and P. C. Shields, "Redundancy rates for renewal and other processes," *IEEE Transactions on Information theory*, vol. 42, no. 6, pp. 2065–2072, 1996.

[26] N. Rajaraman, M. Bondaschi, K. Ramchandran, M. Gastpar, and A. V. Makkuva, "Transformers on markov data: Constant depth suffices," *Advances in Neural Information Processing Systems*, vol. 37, pp. 137 521–137 556, 2024.

[27] R. Zhou, C. Tian, and S. Diggavi, "Transformers learn variable-order markov chains in-context," *arXiv preprint arXiv:2410.05493*, 2024.

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

[29] O. K. Yüksel and N. Flammarion, "On the sample complexity of next-token prediction," in *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[30] B. L. Edelman, E. Edelman, S. Goel, E. Malach, and N. Tsilivis, "The evolution of statistical induction heads: In-context learning markov chains," *arXiv preprint arXiv:2402.11004*, 2024.