Estimation of entropy rate and Rényi entropy rate for Markov chains

Sudeep Kamath, Sergio Verdú
Department of Electrical Engineering
Princeton University
e-mail: sukamath, verdu@princeton.edu

***Abstract*—Estimation of the entropy rate of a stochastic process with unknown statistics, from a single sample path is a classical problem in information theory. While universal estimators for general families of processes exist, the estimates have not been accompanied by guarantees for fixed-length sample paths. We provide finite sample bounds on the convergence of a plug-in type estimator for the entropy rate of a Markov chain in terms of its alphabet size and its mixing properties. We also discuss Rényi entropy rate estimation for reversible Markov chains.**

I. Introduction

Given a stochastic process $\{X_t\}_{t=1}^{\infty}$, where each X_t takes values over the finite alphabet $\mathcal{X}$ of size K, the *entropy rate* of the process defined (whenever the limit exists) by

$$H := \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\imath_{X^n}(X^n)] \tag{1}$$

is a fundamental notion of uncertainty per unit time contained within the process (here and elsewhere, $\imath_Y(y) := \log \frac{1}{p_Y(y)}$ denotes the information[1] in y). Estimating the entropy rate of a process with unknown statistics from the observation of a sample path is an important problem with applications in diverse areas such as data compression, bioinformatics [1] and image processing [2]. This problem has enjoyed a rich history in information theory with its origin in the study of the entropy rate of the English language [3]. Any universal data compression algorithm that achieves the entropy rate can be used as a universal entropy rate estimator. Inspired by the analysis of the Lempel-Ziv algorithm [4], Wyner and Ziv [5] proposed an estimator based on recurrence times and proved that it converges to H in probability for all stationary ergodic processes.

Analogously, the order-α *Rényi entropy rate* for $0 < \alpha < \infty, \alpha \neq 1$ of the stochastic process $\{X_t\}_{t=1}^{\infty}$ is defined by

$$H_\alpha := \lim_{n\to\infty} \frac{1}{n(1-\alpha)} \log \mathbb{E}[e^{(1-\alpha)\imath_{X^n}(X^n)}], \tag{2}$$

and the min-entropy rate (order-∞ Rényi entropy rate) is defined by

$$H_\infty := \lim_{n\to\infty} \frac{1}{n} \min_{x^n} \imath_{X^n}(x^n). \tag{3}$$

The Rényi entropy rate plays a fundamental role in uncertainty in search problems [6], biological sequence analysis [7], and data compression under a risk-averse length criterion [8]. Universal estimators for the Rényi entropy rate have been obtained for stationary ergodic processes under a strong mixing condition [9].

These universal estimators for the entropy rate and Rényi entropy rate converge asymptotically for very general processes. However, so far there is no analysis of their finite-sample performance. At the other extreme, if the process is known to be i.i.d., such finite-sample bounds can be provided for large K from some recent results: it has been shown [10] that for any fixed additive error tolerance and confidence interval, $\Theta\left(\frac{K}{\log K}\right)$ i.i.d. samples is both necessary and sufficient for estimation of the entropy of the unknown distribution. The number of samples necessary and sufficient for estimation of the order-α Rényi entropy of an unknown distribution from its i.i.d. samples has been studied in [11]. To bridge the gap between asymptotic results for the general stationary ergodic process and finite-sample bounds for the i.i.d. process, we consider the simplest and most important family of dependent processes, namely Markov chains. While plug-in estimates for the entropy rate of Markov chains have been investigated before, the focus has been only on proving asymptotic convergence, e.g. [12], [13].

In this paper, we study finite-sample bounds for estimation of the entropy rate of Markov chains. We are specifically interested in the case where the alphabet size of the chain is large, as this is the case in many applications of contemporary interest. Even if the alphabet size is small (such as say, the English alphabet), a higher order Markov source is sometimes employed as a more accurate model for real data and such a source may be viewed as a first order Markov chain over a larger alphabet, so our results are relevant to such higher order Markov sources as well. To provide

[1] All logarithms in this paper are natural logarithms.

any guarantees, a bound on the alphabet size alone is not sufficient and some assumption on the mixing properties of the chain must be made. We assume an upper bound on the relaxation time (reciprocal of the absolute spectral gap) for reversible Markov chains, and an upper bound on the reciprocal of their *pseudo spectral gap*, a quantity recently introduced in [14] for general Markov chains (i.e. without assuming reversibility). While spectral bounds are available for many important Markov chain models, in practice, we may observe a sample path of a chain for which no such spectral bounds are known. In such cases, we could resort to estimation of its spectral properties; for instance, [15] studies estimation of the absolute spectral gap and the minimum stationary probability for reversible Markov chains. Our results emphasize the need for suitable mixing bounds to guarantee convergence of any estimator and also help clarify the dependence on such mixing properties for the plug-in type estimator.

For Rényi entropy rate estimation of Markov chains, we show that bounds on the alphabet size and mixing time cannot suffice to produce finite-sample bounds for *any* estimate of the order-α Rényi entropy rate. By assuming in addition, a lower bound on the minimum stationary probability of any state of the chain, we give (possibly suboptimal) finite-sample bounds on the convergence of a plug-in type estimator of the order-α Rényi entropy rate for reversible Markov chains. We also provide a formula for the min-entropy rate of a Markov chain and finite-sample bounds on its estimation for reversible Markov chains.

The rest of the paper is organized as follows. In Section II, we state basic Markov chain terminology. In Section III, we discuss the necessity of mixing assumptions for providing finite-sample bounds on entropy rate estimates. In Section IV, we show finite-sample bounds on convergence of a simple plug-in type estimator for the entropy rate of a Markov chain. In Section V, we study finite-sample bounds for Rényi entropy rate estimation. We conclude with a discussion and open questions in Section VI.

II. Markov chain preliminaries

In this section, we set up basic terminology about Markov chains.

A. General Markov chains

Let $\boldsymbol{P}$ be the transition matrix of a discrete-time irreducible aperiodic Markov chain over a finite alphabet $\mathcal{X}$ which we assume for simplicity to be $\mathcal{X} := \{1, 2, \ldots, K\}$. Let $\{X_t\}_{t=1}^n$ be a sample path of the Markov chain, with $X_1 \sim \boldsymbol{q}$ for some initial distribution $\boldsymbol{q}$, and

$$\mathbb{P}[X_{t+1} = j | X_t = i] = P_{ij},\ 1 \le t \le n-1. \tag{4}$$

Let $\boldsymbol{\pi}$ denote its unique stationary distribution. The minimum stationary probability is defined as

$$\pi_{\min} := \min_{i\in\mathcal{X}} \pi_i > 0, \tag{5}$$

where the inequality assumes that π charges all points of $\mathcal{X}$. For such a Markov chain, the entropy rate and Rényi entropy rates defined by the limits in (1), (2) always exist, do not depend on the initial distribution, and for $0 < \alpha < \infty, \alpha \neq 1$ are given by the explicit formulae (see [16], [17])

$$H(\boldsymbol{P}) = \sum_{i=1}^K \pi_i \sum_{j=1}^K P_{ij}\ \ \log \frac{1}{P_{ij}}, \tag{6}$$

$$H_\alpha(\boldsymbol{P}) = \frac{1}{1-\alpha} \log\left(\rho\left(\boldsymbol{P}^{\circ\alpha}\right)\right), \tag{7}$$

where $\boldsymbol{P}^{\circ\alpha}$ is the α^{th} Hadamard power of $\boldsymbol{P}$, namely a matrix with $(i,j)^{\text{th}}$ entry given by P_{ij}^α, and $\rho(A)$ is the spectral radius of a matrix A. In Theorem 2 of Section V, we provide a formula for the min-entropy rate $H_\infty(\boldsymbol{P})$.

If the eigenvalues of the transition matrix $\boldsymbol{P}$ are $1 = \lambda_1, \lambda_2, \ldots, \lambda_K$, then the *absolute spectral gap* of the Markov chain is defined to be

$$\gamma_*(\boldsymbol{P}) := 1 - \max_{2\le i\le K} |\lambda_i| > 0, \tag{8}$$

where the inequality in (8) follows from the ergodicity of the Markov chain. The *relaxation time* of the Markov chain is defined to be

$$t_{\text{rel}} := \frac{1}{\gamma_*(\boldsymbol{P})}\ . \tag{9}$$

If $d_{\text{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$ denotes the total variation distance between distributions P and Q, then for $0 < \epsilon < \frac{1}{2}$, the *$\epsilon$-mixing time* of the chain is defined by

$$t_{\text{mix}}(\epsilon) := \min\{t \ge 1 : d_{\text{TV}}(\boldsymbol{P}^t(i,\cdot), \boldsymbol{\pi}) \le \epsilon, \forall i \in \mathcal{X}\}. \tag{10}$$

It is easy to argue that for $0 < \tau < \epsilon < \frac{1}{2}$, (see e.g. [18, Sec 4.5])

$$t_{\text{mix}}(\epsilon) \le t_{\text{mix}}(\tau) \le \left\lceil \frac{\log(\tau^{-1})}{\log\left((2\epsilon)^{-1}\right)} \right\rceil t_{\text{mix}}(\epsilon). \tag{11}$$

We choose the standard terminology,

$$t_{\text{mix}} := t_{\text{mix}}\left(1/4\right) \tag{12}$$

for concreteness, although the bounds we present are easy to adapt to other arguments.

The relationship between the mixing time of a Markov chain and the spectral properties of its transition matrix can be found in terms of the pseudo spectral gap introduced in [14], which we define briefly.

First, let $\boldsymbol{P}^*$ denote the transition matrix of the reverse chain, namely it satisfies

$$\pi_i P^*_{ij} = \pi_j P_{j,i} \quad \forall \;\; i,j \in \mathcal{X}. \tag{13}$$

The chain is defined to be *reversible* if $\boldsymbol{P}^* = \boldsymbol{P}$. If the chain is reversible, then the eigenvalues of the transition matrix $\boldsymbol{P}$ are real. In this case, we define its *spectral gap* as

$$\gamma(\boldsymbol{P}) := 1 - \max_{2 \le i \le K} \lambda_i. \tag{14}$$

The *pseudo spectral gap* of a general Markov chain is then defined as

$$\gamma_{\text{ps}}(\boldsymbol{P}) := \max_{r \ge 1} \frac{\gamma((\boldsymbol{P}^*)^r(\boldsymbol{P})^r)}{r} \ , \tag{15}$$

where we note that for each $r \ge 1$, $(\boldsymbol{P}^*)^r(\boldsymbol{P})^r$ is the transition matrix of a reversible Markov chain and hence, that its spectral gap is well-defined.

If we define the *pseudo relaxation time* of a general Markov chain (reversible or not) as

$$t_{\text{ps}} := \frac{1}{\gamma_{\text{ps}}(\boldsymbol{P})} \ , \tag{16}$$

then, the pseudo relaxation time and the mixing time are related to each other as [14, Prop 3.4]

$$\frac{t_{\text{ps}}}{2} \le t_{\text{mix}} \le t_{\text{ps}} \left(1 + \log \frac{4}{\pi_{\min}}\right). \tag{17}$$

Furthermore, from the definitions (14), (15) and the fact that $\boldsymbol{P}^r$ has an eigenvalue with absolute value $(\max_{2\le i \le K} |\lambda_i|)^r$, we get

$$\begin{aligned} \gamma_{\text{ps}}(\boldsymbol{P}) &\le \max_{r\ge 1} \frac{1 - (\max_{2\le i\le K} |\lambda_i|)^{2r}}{r} && (18)\\ &= 1 - (1 - \gamma_*(\boldsymbol{P}))^2 && (19)\\ &= 2\gamma_*(\boldsymbol{P}) - \gamma_*(\boldsymbol{P})^2 \le 2\gamma_*(\boldsymbol{P}), && (20) \end{aligned}$$

where (19) follows from (8) and Bernoulli's inequality $\frac{1-a}{r} \le 1 - a^{1/r}$ for $0 \le a \le 1, r \ge 1$, Hence,

$$t_{\text{rel}} \le 2 t_{\text{ps}} \ . \tag{21}$$

For a general Markov chain, [19, Prop 1.2] gives an upper bound on the mixing time in terms of the relaxation time and the alphabet size without invoking $\pi_{\min}$. Using that bound in conjunction with (21) yields

$$t_{\text{mix}} \le 4 t_{\text{ps}} \left(K(\log t_{\text{ps}} + 2 + \log 8) + \log 4 - 1\right). \tag{22}$$

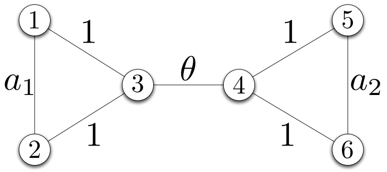

Fig. 1. Random walk on the weighted graph for $a_1, a_2 \in \{1, 2\}$

B. Reversible Markov chains

For reversible chains, the relaxation time and the mixing time are intimately related by [18, Thm. 12.3, 12.4] as

$$(t_{\text{rel}} - 1)\log 2 \le t_{\text{mix}} \le t_{\text{rel}} \log \frac{4}{\pi_{\min}} \ . \tag{23}$$

A recent result for reversible chains [19, Prop 1.1] provides an upper bound on the mixing time in terms of only the relaxation time and alphabet size, without invoking $\pi_{\min}$:

$$t_{\text{mix}} \le 2 t_{\text{rel}} \left(K - 1 + 2\log 4 + \sqrt{2(K-2)\log 4}\right). \tag{24}$$

III. Need for mixing assumptions

Consider the reversible random walk on the six-node weighted graph in Fig. 1, where transition probabilities out of any state are proportional to the weights on its outgoing edges. Suppose that the weights $a_1, a_2 \in \{1, 2\}$ and the weight θ on the bottleneck edge is very small. The entropy rate of the chain is

$$H = \frac{2\log 2 + \sum_{i=1}^2 (1 + a_i) h\left(\frac{1}{1+a_i}\right)}{4 + a_1 + a_2} + O(f(\theta)) \ , \tag{25}$$

where $h(\cdot)$ is the binary entropy function in nats, and $f(\theta) \to 0$ as $\theta \to 0$.

For any initial distribution, the probability that the chain never crosses the bottleneck edge in a sample path of length n is at least $\left(\frac{2}{2+\theta}\right)^n \ge 1 - \frac{n\theta}{2}$. If $n\theta \ll 1$, with high probability, the sample path does not cross over the bottleneck edge. Thus, we cannot infer both the weights a_1, a_2 and it is not possible to estimate the entropy rate within a given level of accuracy with a sufficiently small probability of error.

This example illustrates the necessity of mixing assumptions. Among the most well-known of the mixing properties of a Markov chain are the relaxation time t_{rel} and the mixing time t_{mix}. (23) shows that the relaxation time is always smaller than the mixing time up to a constant factor. Further, [19] shows that the inequality (24) is essentially sharp: for a reversible Markov chain on an alphabet of size K, t_{mix} can be as large as

$\Theta(Kt_{\mathrm{rel}})$. In this paper, we assume upper bounds on the relaxation time for reversible chains and on the pseudo relaxation time for general chains. This leads to a more general setting than imposing the same upper bounds on the mixing time. Let $\mathcal{M}_{\mathrm{rev}}(K, T_{\mathrm{rel}}), \mathcal{M}(K, T_{\mathrm{ps}})$ denote the set of all transition matrices of irreducible aperiodic reversible and irreducible aperiodic general Markov chains respectively on alphabets of size at most K and $t_{\mathrm{rel}} \leq T_{\mathrm{rel}}$ and $t_{\mathrm{ps}} \leq T_{\mathrm{ps}}$ respectively. Let $\mathcal{M}_{\mathrm{rev}}(K, T_{\mathrm{rel}}, \pi_*)$ denote the set of all transition matrices of reversible Markov chains on alphabets of size at most K, $t_{\mathrm{rel}} \leq T_{\mathrm{rel}}$, and minimum stationary probability $\pi_{\min} \geq \pi_*$.

IV. Entropy rate estimation

In this section, we obtain a finite-sample bound on the performance of a plug-in-type estimator for the entropy rate of a Markov chain. If $\{X_t\}_{t=1}^n$ is a sample path of a Markov chain over alphabet $\mathcal{X} = \{1, 2, \ldots, K\}$, then we can define a plug-in estimator for its entropy rate. For $i, j \in \mathcal{X}$,

$$N_{ij} := |\{1 \leq t \leq n-1 : (X_t, X_{t+1}) = (i, j)\}|, \quad (26)$$

$$N_i := |\{1 \leq t \leq n-1 : X_t = i\}|. \quad (27)$$

$$\hat{H}_{\mathrm{plug-in}} = \sum_{i=1}^{K} \frac{N_i}{n-1} \left(\sum_{j=1}^{K} \frac{N_{ij}}{N_i} \, \log \frac{N_i}{N_{ij}} \right). \quad (28)$$

A simple variant of this estimator is used to obtain Theorem 1 whose proof is placed in Appendix A.

Theorem 1. *Let $\{X_t\}_{t=1}^n$ be a sample path of a Markov chain with any transition matrix $\boldsymbol{P} \in \mathcal{M}_{\mathrm{rev}}(K, T_{\mathrm{rel}})$ initiated at any distribution. For any $0 < \epsilon < 1$, there exists an estimate $\hat{H}^{(n)}$ such that with probability at least $1 - \epsilon$, we have*

$$|\hat{H}^{(n)} - H(\boldsymbol{P})| \leq \tfrac{C_1 K^2 T_{\mathrm{rel}}}{n' \epsilon} + \sqrt{\tfrac{C_1 K T_{\mathrm{rel}} \log^2 n'}{n' \epsilon}} \,, \quad (29)$$

where $n' = \max\{n - C_2 K T_{\mathrm{rel}} \log \epsilon^{-1}, 0\}$ for some absolute constants $C_1, C_2 > 0$. If $\boldsymbol{P} \in \mathcal{M}(K, T_{\mathrm{ps}})$ instead, for any $0 < \epsilon < 1$, there exists an estimate $\hat{H}^{(n)}$ such that with probability at least $1 - \epsilon$,

$$|\hat{H}^{(n)} - H(\boldsymbol{P})| \leq \tfrac{C_3 K^2 T_{\mathrm{ps}}}{n'' \epsilon} + \sqrt{\tfrac{C_3 K T_{\mathrm{ps}} (\log T_{\mathrm{ps}})(\log^2 n'')}{n'' \epsilon}} \,, \quad (30)$$

where $n'' = \max\{n - C_4 K T_{\mathrm{ps}} \log T_{\mathrm{ps}} \log \epsilon^{-1}, 0\}$ for some absolute constants $C_3, C_4 > 0$.

Remark 1. In particular, if $\boldsymbol{P} \in \mathcal{M}_{\mathrm{rev}}(K, T_{\mathrm{rel}})$, and if the chain is not too slow mixing, i.e. $T_{\mathrm{rel}} << e^{\sqrt{K}}/K$, then for any fixed desired accuracy and specified upper bound on the error probability, $n = O(K^2 T_{\mathrm{rel}})$ length sample path is sufficient for estimation of the entropy rate.

Remark 2. On the right hand sides of (29) and (30), the first term derives from bounds on the bias of the estimator and the second from those on the variance.

Remark 3. One of the important features of Thm. 1 is that its bounds do not depend on the minimum stationary probability $\pi_{\min}$. In contrast, we shall see in Section V that such dependence on $\pi_{\min}$ is unavoidable for Rényi entropy rate estimation.

V. Rényi entropy rate estimation

We start by providing a simple formula for the min-entropy rate (order-∞ Rényi entropy rate) of a Markov chain. The proof of Theorem 2 is placed in Appendix B.

Given a state space of any Markov chain with transition matrix $\boldsymbol{P}$, a *loop* is a sequence of distinct states of the chain $(i_1, i_2, \ldots, i_l)$ with $l \geq 1$ such that $P_{i_s, i_{s+1}} > 0$ for $s = 1, 2, \ldots, l$ where $i_{l+1} \equiv i_1$. (If $P_{i,i} > 0$, then (i) is a loop.) The set of all loops of length l is denoted by $\mathcal{C}_l(\boldsymbol{P})$.

Theorem 2. *Let $\boldsymbol{P}$ be the transition matrix of an irreducible aperiodic Markov chain on a finite alphabet $\mathcal{X}$. The min-entropy rate of the Markov chain is given by*

$$H_\infty(\boldsymbol{P}) = \min_{1 \leq l \leq K} \min \frac{1}{l} \sum_{s=1}^{l} \imath_{X_2|X_1}(i_{s+1}|i_s), \quad (31)$$

where the inner minimum is taken over all loops $(i_1, \ldots, i_l) \in \mathcal{C}_l(\boldsymbol{P})$, and $\imath_{X_2|X_1}(j|i) := \log \frac{1}{P_{ij}}$. For reversible Markov chains, (31) *simplifies to*

$$H_\infty(\boldsymbol{P}) = \min_{i,j \in \mathcal{X}} \frac{1}{2} \left[\imath_{X_2|X_1}(j|i) + \imath_{X_2|X_1}(i|j) \right]. \quad (32)$$

In parallel with Section III, it can be shown that in addition to a bound on the alphabet size, a bound on the relaxation time is necessary in order to provide guarantees for the estimation of Rényi entropy rates. However, we show that upper bounds on the alphabet size and relaxation time alone do not suffice to provide finite-sample bounds on the accuracy with which the order-α Rényi entropy rate may be estimated for any $\alpha \neq 1$, even for stationary reversible chains. The proof is placed in Appendix C.

Theorem 3. *Fix any $\alpha \in (0, 1) \cup (1, \infty]$. There does not exist any estimate $\hat{H}_\alpha^{(n)}$ based on a sample path $\{X_t\}_{t=1}^n$ of length n of a stationary reversible Markov chain such that for any transition matrix $\boldsymbol{P} \in \mathcal{M}_{\mathrm{rev}}(K, T_{\mathrm{rel}})$, we have $\mathbb{P}[|\hat{H}_\alpha^{(n)} - H_\alpha(\boldsymbol{P})| \geq \delta] \leq \epsilon$, for sufficiently small constants ϵ, δ, if the length n of the sample path is only allowed to depend on $K, T_{\mathrm{rel}}, \epsilon, \delta$.*

However as the next result shows, additional knowledge of a lower bound on the stationary probability

$\pi_{\min}$ opens the possibility of such bounds. The proof of Theorem 4 is in Appendix D.

Theorem 4. *Fix any $\alpha \in (0,1) \cup (1,\infty]$. Let $\{X_t\}_{t=1}^n$ be a sample path of a reversible Markov chain with transition matrix $\boldsymbol{P} \in \mathcal{M}_{\mathrm{rev}}(K, T_{\mathrm{rel}}, \pi_*)$ initiated at any distribution. If $\alpha \in (0,1) \cup (1,\infty)$, then there exists an estimate $\hat{H}_\alpha^{(n)}$ such that for any $0 < \epsilon < 1$, with probability at least $1-\epsilon$,*

$$|\hat{H}_\alpha^{(n)} - H_\alpha(\boldsymbol{P})| \\ \le C_\alpha K^{\alpha \vee 1} \left(\sqrt{\frac{T_{\mathrm{rel}} \log \frac{K}{\epsilon} \log \frac{n}{\pi_* \epsilon}}{\pi_* n}} + \frac{T_{\mathrm{rel}} \log T_{\mathrm{rel}}}{n} \right)^{\alpha \wedge 1}, \quad (33)$$

where $C_\alpha > 0$ is an absolute constant, $a \vee b = \max\{a,b\}, a \wedge b = \min\{a,b\}$.

If $\alpha = \infty$, and $0 < \epsilon < 1$, then there exists an estimate of the min-entropy rate $\hat{H}_\infty^{(n)}$ such that

$$|\hat{H}_\infty^{(n)} - H_\infty(\boldsymbol{P})| \\ \le C_\infty K \left(\sqrt{\frac{T_{\mathrm{rel}} \log \frac{K}{\epsilon} \log \frac{n}{\pi_* \epsilon}}{\pi_* n}} + \frac{T_{\mathrm{rel}} \log T_{\mathrm{rel}}}{n} \right), \quad (34)$$

for some absolute constant $C_\infty > 0$, with probability at least $1-\epsilon$.

VI. Discussion

The estimator studied in Section IV is of the plug-in type. The analysis of this estimator in the proof of Thm. 1 shows a large bias. Efforts to reduce this bias should generally improve performance [20].

A few open questions are as follows. 1) Characterizing (up to constant factors) the minimax risk for estimating the entropy rate for a family of Markov chains (such as all reversible Markov chains with a bound on their alphabet size and relaxation time). 2) Our bounds for entropy rate estimation of non-reversible chains involve bounds on the pseudo relaxation time. We do not know if such bounds could be obtained using bounds on the relaxation time instead. 3) Sharper bounds for Rényi entropy rate estimation could be obtained via a multiplicative error analysis, rather than the additive error analysis we have performed in this paper capitalizing on [15]. For non-reversible Markov chains, an eigenvalue perturbation analysis could be carried out for the estimation of the spectral radius of the Hadamard power of the transition matrix.

Acknowledgements

We would like to thank Ramon van Handel, Aryeh Kontorovich, and Daniel Paulin for helpful discussions. SK would like to acknowledge support from the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. SV would like to acknowledge support from MURI Grant 67470CSMUR.

References

[1] J. Lanctot, M. Li, and E.-H. Yang, "Estimating DNA sequence entropy", in *Proc. Symp. Discrete Algorithms*, San Francisco, CA, 2000.

[2] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs", *IEEE Signal Processing Mag.*, vol. 19, pp. 85–95, September 2002.

[3] C. E. Shannon, "Prediction and Entropy of printed English", *Bell Syst. Tech. J.*, pp. 50–64, 1951.

[4] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression", *IEEE Trans. Inform. Theory*, vol. 23, pp. 337–343, May 1977.

[5] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression", *IEEE Trans. Inform. Theory*, vol. 35, pp. 1250–1258, 1989.

[6] L. Pronzato, H. P. Wynn, and A. A. Zhigljavsky, "Using Rényi entropies to measure uncertainty in search problems", *Lectures in Applied Mathematics*, vol. 33, pp. 253–268, 1997.

[7] S. Vinga, "Information theory applications for biological sequence analysis", *Briefings in bioinformatics*, vol. 15, no. 3, pp. 376–389, 2013.

[8] L. L. Campbell, "A coding theorem and Rényi's entropy", *Information and Control*, vol. 8, pp. 423–429, 1965.

[9] W. Szpankowski, "A generalized suffix tree and its (un)expected asymptotic behaviors", *SIAM Journal on Computing*, vol. 22, no. 6, pp. 1176–1198, 1993.

[10] G. Valiant and P. Valiant, "Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs", in *Proc. of the 43rd annual ACM symposium on Theory of Computing*, 2011.

[11] J. Acharya, A. Orlitsky, H. Tyagi, and A. T. Suresh, "The complexity of estimating Rényi entropy", in *Proc. of the ACM-SIAM Symposium on Discrete Algorithms*, 2015.

[12] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal Entropy Estimation Via Block Sorting", *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1551–1561, July 2004.

[13] G. Ciuperca and V. Girardin, "On the estimation of the entropy rate of finite Markov chains", in *Proc. of the International Symposium on Applied Stochastic Models and Data Analysis*, 2005.

[14] D. Paulin, "Concentration inequalities for Markov chains by Marton couplings and spectral methods", *arXiv:1212.2015v4 [math.PR]*, Jan. 2015.

[15] D. J. Hsu, A. Kontorovich, and C. Szepesvári, "Mixing time estimation in reversible Markov chains from a single sample path", in *Advances in Neural Information Processing Systems*, 2015, pp. 1459–1467.

[16] C. E. Shannon, "A mathematical theory of communication", *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul.-Oct. 1948.

[17] Z. Rached, F. Alajaji, and L.L. Campbell, "Rényi's entropy rate for Discrete Markov sources", in *Proc. of CISS*, Baltimore, MD, March 1999.

[18] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*, American Mathematical Society, Providence, RI, 2009.

[19] D. Jerison, "General mixing time bounds for finite Markov chains via the absolute spectral gap", *arXiv:1310.8021v1 [math.PR]*, Oct. 2013.

[20] L. Paninski, "Estimation of entropy and mutual information", *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[21] A. W. Marcus, D. A. Spielman, and N. Srivastava, "Interlacing families I: Bipartite Ramanujan graphs of all degrees", *Annals of Mathematics*, vol. 182, no. 1, pp. 307–325, 2015.

Appendix A
Proof of Theorem 1

Proof: Let us assume that the Markov chain has alphabet $\mathcal{X} = \{1, 2, \ldots, K\}$. We first analyze the plug-in estimator for the chain initiated at the stationary distribution $\boldsymbol{\pi}$, and will later modify it. Suppose that we observe a sample path $X_1, X_2, \ldots, X_{n+1}$ of length $n+1$.

Define, for $i, j \in \mathcal{X}$,

$$\begin{aligned} N_{ij} &:= \left|\{1 \le t \le n : (X_t, X_{t+1}) = (i,j)\}\right| \ , && (35)\\ N_i &:= \left|\{1 \le t \le n : X_t = i\}\right| \ . && (36) \end{aligned}$$

Consider the following plug-in estimates:

$$\hat{\pi}_i := \frac{N_i}{n} \qquad (37)$$

$$\hat{P}_{ij} := \begin{cases} \frac{N_{ij}}{N_i} & \text{if } N_i > 0 \ , \\ 1_{i=j} & \text{if } N_i = 0 \ , \end{cases} \qquad (38)$$

which we use to obtain a plug-in estimator for the entropy rate as

$$\hat{H} := \sum_{i=1}^{K} \hat{\pi}_i \left(\sum_{j=1}^{K} \hat{P}_{ij} \ \log \frac{1}{\hat{P}_{ij}} \right) \ . \qquad (39)$$

Let the doublet probability distribution on $\mathcal{X} \times \mathcal{X}$ and its estimate be defined respectively by

$$Q_{ij} = \pi_i P_{ij}, \quad \forall i, j \in \mathcal{X}. \qquad (40)$$

$$\hat{Q}_{ij} = \hat{\pi}_i \hat{P}_{ij} = \frac{N_{ij}}{n}, \quad \forall i, j \in \mathcal{X}. \qquad (41)$$

Then, $H = H_2 - H_1$ and $\hat{H} = \hat{H}_2 - \hat{H}_1$ where

$$H_1 = H(\boldsymbol{\pi}) \ , \ \hat{H}_1 = H(\hat{\boldsymbol{\pi}}) \ , \qquad (42)$$

$$H_2 = H(\boldsymbol{Q}) \ , \ \hat{H}_2 = H(\hat{\boldsymbol{Q}}) \ . \qquad (43)$$

Let us bound the bias and variance in the estimate $\hat{H}_2$ of H_2. Note first that

$$\hat{H}_2 - H_2 = \sum_{ij} (Q_{ij} - \hat{Q}_{ij}) \log Q_{ij} - D(\hat{Q} || Q). \quad (44)$$

By taking expectations, and using $\mathbb{E}[\hat{Q}_{ij}] = Q_{ij}$ from stationarity, we obtain

$$\begin{aligned} |\mathbb{E}[\hat{H}_2] - H_2| &= \mathbb{E}\left[D(\hat{Q}||Q) \right] && (45)\\ &\le \mathbb{E}\left[\sum_{i,j \in \mathcal{X}} \frac{(\hat{Q}_{ij} - Q_{ij})^2}{Q_{ij}} \right], && (46) \end{aligned}$$

where (46) follows from $\log u \le u - 1$ for $u > 0$. Now, by Minkowski's inequality,

$$\begin{aligned} &\sqrt{\mathbb{E}\left[(\hat{Q}_{ij} - Q_{ij})^2 \right]} \\ \le \ &\sqrt{\mathbb{E}\left[(\hat{Q}_{ij} - \hat{\pi}_i P_{ij})^2 \right]} + \sqrt{\mathbb{E}\left[(\hat{\pi}_i P_{ij} - Q_{ij})^2 \right]} \ . \end{aligned} \qquad (47)$$

We bound each of the terms on the right side separately. First, we observe that $\left(1_{X_{t-1}=i}(1_{X_t=j} - P_{ij})\right)_{t=2}^{n+1}$ constitutes a martingale difference sequence adapted to the filtration $(\mathcal{F}_t)_{t=2}^{n+1}$ where $\mathcal{F}_t = \sigma(X_1, X_2, \ldots, X_t)$. Since martingale differences are uncorrelated, we get that

$$\begin{aligned} &\mathbb{E}\left[(N_{ij} - N_i P_{ij})^2 \right] \\ =&\mathbb{E}\left[\left(\sum_{t=2}^{n+1} 1_{X_{t-1}=i}(1_{X_t=j} - P_{ij}) \right)^2 \right] && (48)\\ =& \sum_{t=2}^{n+1} \mathbb{E}\left[\left(1_{X_{t-1}=i}(1_{X_t=j} - P_{ij})\right)^2 \right] && (49)\\ =& \sum_{t=2}^{n+1} \mathbb{E}\left[\mathbb{E}\left[\left(1_{X_{t-1}=i}(1_{X_t=j} - P_{ij})\right)^2 | X_{t-1} \right] \right] && (50)\\ =& \sum_{t=2}^{n+1} \mathbb{E}\left[1_{X_{t-1}=i} \mathbb{E}\left[(1_{X_t=j} - P_{ij})^2 | X_{t-1} = i \right] \right] && (51)\\ =& \left(\mathbb{E} N_i\right) P_{ij}(1 - P_{ij}) && (52)\\ =& \ n \pi_i P_{ij}(1 - P_{ij}) \ , && (53) \end{aligned}$$

where (53) follows from stationarity. This gives

$$\mathbb{E}[(\hat{Q}_{ij} - \hat{\pi}_i P_{ij})^2] = \frac{\pi_i P_{ij}(1 - P_{ij})}{n} \ . \qquad (54)$$

For the second term in the right side of (47),

$$\mathbb{E}[(\hat{\pi}_i P_{ij} - Q_{ij})^2] = \frac{P_{ij}^2}{n^2} \operatorname{Var}(N_i) \le \frac{2\pi_i (1 - \pi_i) P_{ij}^2}{n \gamma(\boldsymbol{P})} \qquad (55)$$

where the inequality follows from [14, Thm. 3.5, (3.14)].

From (47), (54) and (55), it follows that

$$\begin{aligned} &\sum_{i,j \in \mathcal{X}} \frac{\mathbb{E}(\hat{Q}_{ij} - Q_{ij})^2}{Q_{ij}} \\ \le& \sum_{i,j \in \mathcal{X}} \frac{\pi_i P_{ij}}{n Q_{ij}} \left(\sqrt{1 - P_{ij}} + \sqrt{\frac{2(1 - \pi_i) P_{ij}}{\gamma(\boldsymbol{P})}} \right)^2 && (56)\\ \le& \frac{1}{n} \sum_{i,j \in \mathcal{X}} \left(1 + \sqrt{\frac{2}{\gamma(\boldsymbol{P})}} \right)^2 && (57)\\ \le& \frac{8K^2}{n \gamma(\boldsymbol{P})} \ , && (58) \end{aligned}$$

where (A) follows from $\gamma(\boldsymbol{P}) \le 2$. From (46) and , we get

$$\left| \mathbb{E}[\hat{H}_2] - H_2 \right| \le \frac{8K^2}{n \gamma(\boldsymbol{P})} \ . \qquad (59)$$

Note that $\hat{H}_2$ is a function of $(X_1, X_2, \ldots, X_{n+1})$ with the bounded differences property: a change in only the t^{th} co-ordinate keeping all the others fixed will change the function by at most

$$c_t := \begin{cases} \frac{4 \log n}{n} & 2 \le t \le n, \\ \frac{2 \log n}{n} & t = 1, n+1. \end{cases} \qquad (60)$$

This is because at most four of the $\hat{Q}_{ij}$ will be affected (two if $t = 1, n+1$) and the function $g(u) := u \log \frac{1}{u}$ satisfies

$$\left| g\left(\frac{a}{n}\right) - g\left(\frac{a-1}{n}\right) \right| \leq \frac{\log n}{n} \quad \forall\ a = 1, 2, \ldots, n. \tag{61}$$

The proof of Corollary 2.11 in [14] then implies that $\hat{H}_2$ is a sub-Gaussian random variable, i.e.

$$\log \mathbb{E} e^{\lambda(\hat{H}_2 - \mathbb{E}[\hat{H}_2])} \leq \frac{\lambda^2 \sigma^2}{2},\ \forall\ \lambda \in \mathbb{R}, \tag{62}$$

with

$$\sigma^2 = (9 t_{\text{mix}}) \cdot \frac{1}{4} \sum_{t=1}^{n+1} c_t^2 \leq \frac{1}{n} \cdot 36 t_{\text{mix}} \log^2 n\ . \tag{63}$$

It follows that

$$\text{Var}(\hat{H}_2) \leq \sigma^2 \leq \frac{1}{n} \cdot 36 t_{\text{mix}} \log^2 n\ . \tag{64}$$

Similar calculations yield

$$|\mathbb{E}\hat{H}_1 - H_1| \leq \frac{2K}{n\gamma(\boldsymbol{P})}\ , \tag{65}$$

$$\text{Var}(\hat{H}_1) \leq \frac{1}{n} \cdot 9 t_{\text{mix}} \log^2 n\ . \tag{66}$$

Again, by Minkowski's inequality,

$$\sqrt{\mathbb{E}[(\hat{H} - H)^2]} \leq \sqrt{\mathbb{E}[(\hat{H}_2 - H_2)^2]} + \sqrt{\mathbb{E}[(\hat{H}_1 - H_1)^2]}, \tag{67}$$

and so from (59), (64), (65), (66), we obtain

$$\mathbb{E}[(\hat{H} - H)^2] \leq \left(\frac{16K^2}{n\gamma(\boldsymbol{P})}\right)^2 + \frac{144 t_{\text{mix}} \log^2 n}{n}\ . \tag{68}$$

By Chebyshev's inequality, we get

$$\mathbb{P}[|\hat{H} - H| \geq \delta] \leq \left(\frac{16K^2}{n\gamma(\boldsymbol{P})\delta}\right)^2 + \frac{144 t_{\text{mix}} \log^2 n}{n\delta^2}\ . \tag{69}$$

Now, suppose that the chain is initiated at a distribution $\boldsymbol{q}$ that is not the stationary distribution. If $\mathbb{P}_{\boldsymbol{q}}$ denotes probabilities under initiation of the chain with distribution $\boldsymbol{q}$, it follows from a coupling argument that

$$\mathbb{P}_{\boldsymbol{q}}[|\hat{H} - H| \geq \delta] \leq \mathbb{P}_{\boldsymbol{\pi}}[|\hat{H} - H| \geq \delta] + d_{\text{TV}}(\boldsymbol{q}, \boldsymbol{\pi}). \tag{70}$$

We therefore "burn" the first $n_0 = t_{\text{mix}}\left(\frac{\epsilon}{2}\right)$ observations and apply the plug-in estimator to the subsequent sequence of length $n' + 1 = n - n_0$. Suppose that $X_{n_0+1} \sim \boldsymbol{q}'$ then $d_{\text{TV}}(\boldsymbol{q}', \boldsymbol{\pi}) \leq \frac{\epsilon}{2}$ from the definition of the mixing time. The result for reversible chains follows from using

$$\frac{1}{\gamma(\boldsymbol{P})} \leq \frac{1}{\gamma_*(\boldsymbol{P})} \leq T_{\text{rel}},$$

the bound on t_{mix} in (24) along with (11), (69), and (70).

The proof for general chains without assuming reversibility is identical, except we use [14, Thm 3.7, (3.17)] to bound $\text{Var}(N_i)$ and (22) to bound the mixing time using K and T_{ps}.

■

Appendix B
Proof of Theorem 2

Proof: Let the right hand side of (31) be denoted by $G_\infty(\boldsymbol{P})$. Let $(j_1, j_2, \ldots, j_l)$ be a loop such that

$$G_\infty(\boldsymbol{P}) = \frac{1}{l} \sum_{s=1}^{l} \imath_{X_2|X_1}(j_{s+1}|j_s), \tag{71}$$

where $1 \leq l \leq K$.

Suppose that the Markov chain is initiated at some arbitrary distribution $\boldsymbol{q}$ and let i be any state such that $q_i \geq \frac{1}{K}$. Since the chain is irreducible, there exists a sequence of distinct vertices $i = i_0, i_1, i_2, \ldots, i_r = j_1$ where $P_{i_s, i_{s+1}} > 0$ for $s = 0, 1, 2, \ldots, r-1$ with $r \leq K$. Let y^n be the sequence of states i_0 through $i_r = j_1$ followed by repetitions of the loop $j_1, j_2, \ldots, j_l$ truncated at a total length of n. This sequence has a positive probability under the chosen initial distribution. Furthermore,

$$\imath_{X^n}(y^n) \leq C + \frac{n-K-l}{l} \sum_{s=1}^{l} \imath_{X_2|X_1}(i_{s+1}|i_s) \tag{72}$$

$$\leq C + nG_\infty(\boldsymbol{P}) \tag{73}$$

where C is a constant that depends on the transition matrix. Dividing by n and taking the lim sup as $n \to \infty$ gives

$$\limsup_{n\to\infty} \frac{1}{n} \min_{x^n} \imath_{X^n}(x^n) \leq G_\infty(\boldsymbol{P}). \tag{74}$$

Now, suppose x^n is any sequence of states of length n with positive probability under an arbitrarily chosen initial distribution. Let $t_1 < t_2$ be indices such that $x_{t_1} = x_{t_2}$ and t_2 is the smallest possible. Then the sequence $(x_{t_1}, \ldots, x_{t_2-1}) = (i_1, i_2, \ldots, i_l)$ is a loop of length $l = t_2 - t_1$. Deleting these l states, we shorten the sequence x^n into a sequence z^{n-l}. By the Markov property,

$$\imath_{X^n}(x^n) = \imath_{X^{n-l}}(z^{n-l}) + \left(\sum_{s=1}^{l} \imath_{X_2|X_1}(i_{s+1}|i_s)\right) \tag{75}$$

$$\geq \imath_{X^{n-l}}(z^{n-l}) + lG_\infty(\boldsymbol{P}). \tag{76}$$

Iteratively repeating this extraction of loops from the sequence z^{n-l}, we will end up with a sequence containing no repeated state, say $w^{l'}$, where $1 \le l' \le K$, resulting in

$$\begin{aligned} \imath_{X^n}(x^n) &\ge \imath_{X^{l'}}(w^{l'}) + (n-l')G_\infty(\boldsymbol{P}) && (77) \\ &\ge (n-K)G_\infty(\boldsymbol{P}). && (78) \end{aligned}$$

Taking the minimum over all sequences x^n of positive probability, dividing by n and computing the lim inf as $n \to \infty$, we obtain

$$\liminf_{n\to\infty} \frac{1}{n} \min_{x^n} \imath_{X^n}(x^n) \ge G_\infty(\boldsymbol{P}). \qquad (79)$$

From (74) and (79), we get that the min-entropy rate exists for any initial distribution, does not depend on it, and is given by $H_\infty(\boldsymbol{P}) = G_\infty(\boldsymbol{P})$.

If $\boldsymbol{P}$ is the transition matrix of a reversible chain and $(i_1, i_2 \dots, i_l)$ is any loop of length $l > 2$, then so is $(i_l, i_{l-1}, \dots, i_1)$ and by invoking the definition of reversibility, we have

$$\begin{aligned} \frac{1}{l}\sum_{s=1}^{l} \imath_{X_2|X_1}(i_{s+1}|i_s) &= \frac{1}{l}\sum_{s=1}^{l} \imath_{X_2|X_1}(i_s|i_{s+1}) && (80) \\ = \frac{1}{l}\sum_{s=1}^{l} \frac{1}{2}&\left[\imath_{X_2|X_1}(i_{s+1}|i_s) + \imath_{X_2|X_1}(i_s|i_{s+1})\right] && (81) \\ \ge \min_{s=1,2,\dots,l} \frac{1}{2}&\left[\imath_{X_2|X_1}(i_{s+1}|i_s) + \imath_{X_2|X_1}(i_s|i_{s+1})\right]. && (82) \end{aligned}$$

This means that it suffices to restrict the minimum in (31) to loops of lengths either one or two, completing the proof. ■

Appendix C
Proof of Theorem 3

Proof: Let $\boldsymbol{Q}$ denote the doublet probability matrix and $\boldsymbol{R}$ denote the normalized affinity matrix for a given reversible Markov chain with transition matrix $\boldsymbol{P}$ and stationary distribution $\boldsymbol{\pi}$ on the alphabet $\mathcal{X}$ given by

$$Q_{ij} = \pi_i P_{ij}, \ \ R_{ij} = \frac{\pi_i P_{ij}}{\sqrt{\pi_i \pi_j}}, \ \ \forall\, i,j \in \mathcal{X}. \qquad (83)$$

On account of reversibility, $\boldsymbol{Q}$ and $\boldsymbol{R}$ are symmetric matrices and so, if $\|A\| := \sup_{\|u\|=1} \|Au\|$ denotes the operator norm, then the spectral radius of $\boldsymbol{P}^{\circ\alpha}$ equals the operator norm of the symmetric matrix $\boldsymbol{R}^{\circ\alpha}$, i.e.

$$\rho\left(\boldsymbol{P}^{\circ\alpha}\right) = \rho\left(\boldsymbol{R}^{\circ\alpha}\right) = \|\boldsymbol{R}^{\circ\alpha}\|, \qquad (84)$$

and therefore,

$$H_\alpha(\boldsymbol{P}) = \frac{1}{1-\alpha} \log \|\boldsymbol{R}^{\circ\alpha}\|. \qquad (85)$$

We discuss below the cases $\alpha > 1$ and $\alpha < 1$ separately.

A. *The case:* $1 < \alpha \le \infty$

We describe a family of reversible Markov chains to show that for $\alpha > 1$, the order-α Rényi entropy of a stationary reversible Markov chain cannot be estimated in a time that depends only on upper bounds on the alphabet size and the relaxation time.

Let $m = K - 1$ and fix $0 < \rho < 1$ and $0 < \tau < \frac{1}{2}$. Let $\bar{\tau} := 1 - \tau$.

Consider a Markov chain $\boldsymbol{P}$ with alphabet $\mathcal{X}$ and $K \times K$ doublet probability matrix $\boldsymbol{Q}$ given by the following:

$$Q_{ij} = \begin{cases} \frac{1-\rho}{m^2}, & 1 \le i,j \le m, \\ \rho(1-2\tau), & i = j = m+1, \\ \frac{\rho\tau}{m}, & \text{else.} \end{cases} \qquad (86)$$

Then, the stationary distribution $\boldsymbol{\pi}(\boldsymbol{P})$ is given by

$$\pi_i = \begin{cases} \frac{1-\rho\bar{\tau}}{m}, & \text{if } 1 \le i \le m\ , \\ \rho\bar{\tau}, & \text{if } i = m+1\ . \end{cases} \qquad (87)$$

Note that the degenerate form of this Markov chain under $\rho = 0$ is simply a process that produces independent samples equiprobably from $\{1, 2, \dots, m\}$. Now, for the given chain with $\rho > 0$, we get

$$R_{ij} = \begin{cases} \frac{1-\rho}{m(1-\rho\bar{\tau})}, & 1 \le i,j \le m, \\ \frac{1-2\tau}{1-\tau}, & i = j = m+1, \\ \frac{\tau\sqrt{\rho}}{\sqrt{m(1-\rho\bar{\tau})(1-\tau)}}, & \text{else.} \end{cases} \qquad (88)$$

The $(m+1) \times (m+1)$ matrix

$$\begin{bmatrix} a & a & \dots & a & b \\ a & a & \dots & a & b \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a & a & \dots & a & b \\ b & b & \dots & b & c \end{bmatrix} \qquad (89)$$

has $m-1$ zero eigenvalues, and the other two eigenvalues are $\frac{1}{2}\left[(am+c) \pm \sqrt{(am-c)^2 + 4mb^2}\right]$. Since $\boldsymbol{R}$ has the form in (89), and the top eigenvalue of $\boldsymbol{R}$ is 1, it follows that its other non-trivial eigenvalue is

$$\text{Trace}(\boldsymbol{R}) - 1 = \frac{1-\rho}{1-\rho\bar{\tau}} - \frac{\tau}{1-\tau}. \qquad (90)$$

The absolute spectral gap of the chain is then, given by:

$$\begin{aligned} \gamma_*(\boldsymbol{P}) &= 1 - \left| \frac{1-\rho}{1-\rho\bar{\tau}} - \frac{\tau}{1-\tau} \right| && (91) \\ &= \frac{\tau}{1-\tau} + \frac{\rho\tau}{1-\rho\bar{\tau}} && (92) \\ &= \frac{\tau}{1-\tau} + O(\rho). && (93) \end{aligned}$$

Likewise, we obtain

$$\begin{aligned} H_\alpha(\boldsymbol{P}) &= \frac{1}{1-\alpha} \log \|\boldsymbol{R}^{\circ\alpha}\| && (94) \\ &= \frac{\alpha}{\alpha-1} \log \frac{1-\tau}{1-2\tau} + O\left(\frac{1}{k^{\alpha-1}}\right). && (95) \end{aligned}$$

Similarly,

$$H_\infty(\boldsymbol{P}) = \log \frac{1-\tau}{1-2\tau} \ . \tag{96}$$

If we choose $\tau \in [0.1, 0.4]$, then (92) gives $\gamma_*(\boldsymbol{P}) \geq= 0.1$, so $t_{\mathrm{rel}} \leq T_{\mathrm{rel}} := 10$.

The probability that the chain initialized at the stationary distribution never visits state $m+1$ in a run of length n is

$$(1-\rho\bar{\tau})\left(1-\frac{\rho\tau}{1-\rho\bar{\tau}}\right)^{n-1} \geq 1-n\rho \ . \tag{97}$$

For any ρ satisfying $0 < \rho n < \frac{1}{100}$, it follows that the chain never visits state $m+1$ in a run of length n with probability at least 99%. Conditioned on this event occuring, the statistics of the observations are an i.i.d. process with samples drawn uniformly from $\{1, 2, \ldots, m\}$ and hence, the statistics do not depend on ρ or τ. Thus, no reliable estimate of τ may be made and the α-Rényi entropy rate for $\alpha > 1$ in (95), (96) which depends critically on τ cannot be estimated to a sufficiently fine precision.

B. The case $0 < \alpha < 1$

Here, again we describe a family of reversible Markov chains to show that for $\alpha < 1$, the order-α Rényi entropy of a reversible Markov chain cannot be estimated in a sample path of length that depends only on upper bounds on the alphabet size and relaxation time. Before we discuss such a family, we briefly recall the notions of conductance of a reversible Markov chain and Ramanujan graphs.

The *conductance* (also called bottleneck ratio) of a reversible Markov chain with transition probability matrix $\boldsymbol{P}$, stationary distribution $\boldsymbol{\pi}$ and doublet probability matrix $\boldsymbol{Q}$ is given by [18, Sec. 7.2]

$$\Phi_*(\boldsymbol{P}) := \min_{S: 0 < \pi(S) \leq \frac{1}{2}} \frac{Q(S, S^c)}{\pi(S)}, \tag{98}$$

where

$$\pi(S) := \sum_{i \in S} \pi_i, \tag{99}$$

$$Q(S, S^c) := \sum_{i \in S, j \in S^c} Q_{ij}. \tag{100}$$

Cheeger's inequality [18, Thm. 13.14] relates the conductance to the spectral gap as

$$\frac{\Phi_*(\boldsymbol{P})^2}{2} \leq \gamma(\boldsymbol{P}) \leq 2\Phi_*(\boldsymbol{P}). \tag{101}$$

A d-regular undirected graph G is called a Ramanujan graph if the associated random walk on its vertices is a reversible Markov chain with transition matrix $\boldsymbol{P}_G$ satisfying

$$\gamma(\boldsymbol{P}_G) \geq 1 - \frac{2\sqrt{d-1}}{d} \ . \tag{102}$$

Bipartite Ramanujan graphs exist for any fixed degree $d \geq 3$ and sufficiently large vertex sets [21].

We now construct a family of reversible Markov chains as follows. Again we have two parameters ρ, τ satisfying $0 < \tau < \frac{\rho}{1-\rho}$.

Suppose K is even. Let G denote a d-regular Ramanujan graph on $\left\{1, 2, \ldots, \frac{K}{2}\right\}$ with edge set $E(G)$. We define the doublet probability matrix $\boldsymbol{Q}$ of a Markov chain on alphabet $\{1, 2, \ldots, K\}$ by

$$Q_{ij} = \begin{cases} \dfrac{2\bar{\rho}\bar{\tau}}{dK} 1_{(i,j) \in E(G)}, & 1 \leq i, j \leq \dfrac{K}{2}, \\ \dfrac{4(\rho - \tau\bar{\rho})}{K^2}, & \dfrac{K}{2} + 1 \leq i, j \leq K, \\ \dfrac{4\bar{\rho}\tau}{K^2}, & \text{else.} \end{cases} \tag{103}$$

Its stationary distribution $\boldsymbol{\pi}$ is given by

$$\pi_i = \begin{cases} \dfrac{2\bar{\rho}}{K}, & 1 \leq i \leq \dfrac{K}{2}, \\ \dfrac{2\rho}{K}, & \dfrac{K}{2} + 1 \leq i \leq K \ . \end{cases} \tag{104}$$

Since $0 < \tau < \frac{\rho}{\bar{\rho}}$, the degenerate form of this Markov chain when $\rho = \tau = 0$ is simply the random walk on the expander graph G.

We claim that the spectral gap of this Markov chain is lower bounded as follows:

$$\gamma(\boldsymbol{P}) \geq \frac{1}{2}\left(\min\left\{\frac{\bar{\rho}\tau}{\rho}, \frac{(1-2\rho)\bar{\tau}}{\left(1 + \frac{\rho K}{2\bar{\rho}}\right)}\left(1 - \frac{2\sqrt{d-1}}{d}\right)\right\}\right)^2 . \tag{105}$$

The claim will follow from Cheeger's inequality (101) and the definition of a Ramanujan graph (102) along with the following relation between the conductance of the Markov chain $\boldsymbol{P}$ and that of the random walk $\boldsymbol{P}_G$ on the graph G:

$$\Phi_*(\boldsymbol{P}) \geq \min\left\{\frac{\bar{\rho}\tau}{\rho}, \frac{(1-2\rho)\bar{\tau}}{\left(1 + \frac{\rho K}{2\bar{\rho}}\right)}\Phi_*(\boldsymbol{P}_G)\right\}. \tag{106}$$

To show (106), consider the different cases for choice of $S \subseteq \{1, 2, \ldots, K\}$. Let $[a : b]$ denote the set $\{a, a+1, \ldots, b\}$.

- $S \subseteq \left[\frac{K}{2} + 1 : K\right]$. Here, $\pi(S) \leq \rho \leq \frac{1}{2}$.

$$\frac{Q(S,S^c)}{\pi(S)} \geq \frac{Q(S, S^c \cap [1 : \frac{K}{2}])}{\pi(S)} \tag{107}$$

$$= \frac{\frac{4\bar{\rho}\tau}{K^2}\frac{K}{2}|S|}{\frac{2\rho}{K}|S|} = \frac{\bar{\rho}\tau}{\rho} \ . \tag{108}$$

- $S \subseteq \left[1 : \frac{K}{2}\right]$. Here, $\pi(S) \leq \frac{1}{2}$ means $1 \leq |S| \leq \frac{K}{4\bar{\rho}}$. Then,

$$\frac{Q(S,S^c)}{\pi(S)}$$
$$\geq \frac{Q(S \cap \left[1 : \frac{K}{2}\right], S^c \cap \left[1 : \frac{K}{2}\right])}{\pi(S \cap \left[1 : \frac{K}{2}\right])} \tag{109}$$
$$\geq \frac{(1-2\rho) Q(S \cap \left[1 : \frac{K}{2}\right], S^c \cap \left[1 : \frac{K}{2}\right])}{\min\{\pi(S \cap \left[1 : \frac{K}{2}\right]), \pi(S^c \cap \left[1 : \frac{K}{2}\right])\}} \tag{110}$$
$$\geq (1-2\rho)\bar{\tau}\Phi_*(\boldsymbol{P}_G) \ , \tag{111}$$

where (110) follows from

$$\frac{\pi(S \cap \left[1 : \frac{K}{2}\right])}{\min\{\pi(S \cap \left[1 : \frac{K}{2}\right]), \pi(S^c \cap \left[1 : \frac{K}{2}\right])\}}$$
$$\leq \frac{\frac{K}{4\bar{\rho}}}{\frac{K}{2} - \frac{K}{4\bar{\rho}}} = \frac{1}{1-2\rho} \ , \tag{112}$$

and (111) follows from the definition of $\Phi_*(\boldsymbol{P}_G)$ noting that for $A = S \cap [1 : \frac{K}{2}], B = S^c \cap [1 : \frac{K}{2}]$, we have $Q(A,B) = \bar{\rho}\bar{\tau}Q_G(A,B)$ and $\min\{\pi(A), \pi(B)\} = \bar{\rho}\min\{\pi_G(A), \pi_G(B)\}$, where $\boldsymbol{Q}_G, \boldsymbol{\pi}_G$ denote the doublet probability matrix and stationary distribution corresponding to $\boldsymbol{P}_G$.

- Suppose $S \cap [1 : \frac{K}{2}] \neq \emptyset, S \cap [\frac{K}{2}+1 : K] \neq \emptyset$. If $\pi(S) \leq \frac{1}{2}$, then

$$\frac{Q(S,S^c)}{\pi(S)}$$
$$\geq \frac{Q(S \cap [1 : \frac{K}{2}], S^c \cap [1 : \frac{K}{2}])}{\left(1 + \frac{\rho K}{2\bar{\rho}}\right)\pi(S \cap [1 : \frac{K}{2}])} \tag{113}$$
$$\geq \frac{(1-2\rho)\bar{\tau}}{\left(1 + \frac{\rho K}{2\bar{\rho}}\right)}\Phi_*(\boldsymbol{P}_G), \tag{114}$$

where (113) follows from using $\pi(S) \leq \pi(S \cap [1 : \frac{K}{2}]) + \rho$ and $\pi(S \cap [1 : \frac{K}{2}]) \geq \frac{2\rho}{K}$, and (114) follows from the previous case.

We now show the following sharp bounds on the order-α Rényi entropy rate of the chain $\boldsymbol{P}$ for $0 < \alpha < 1$:

$$\log\frac{K}{2} - \frac{1}{1-\alpha}\log\frac{1}{\left(1 - \frac{\bar{\rho}\tau}{\rho}\right)^{\alpha}} \tag{115}$$
$$\leq \quad H_\alpha(\boldsymbol{P})$$
$$\leq \quad \log\frac{K}{2} - \frac{1}{1-\alpha}\log\frac{1}{\left(1 - \frac{\bar{\rho}\tau}{\rho}\right)^{\alpha} + \frac{\tau^\alpha\bar{\rho}^{\alpha/2}}{\rho^{\alpha/2}}} \ . \tag{116}$$

To show bounds (115) and (116), consider the spectral norm of $\boldsymbol{R}^{\circ\alpha}$ where $\boldsymbol{R}$ is the normalized affinity matrix given by

$$R_{ij} = \begin{cases} \frac{\bar{\tau}}{d}1_{(i,j)\in E(G)} & 1 \leq i,j \leq \frac{K}{2}, \\ \frac{\rho - \tau\bar{\rho}}{m\rho} & \frac{K}{2} + 1 \leq i,j \leq K, \\ \frac{\tau\sqrt{\bar{\rho}}}{m\sqrt{\rho}} & \text{else.} \end{cases} \tag{117}$$

- From the Perron-Frobenius theorem, $\|\boldsymbol{R}^{\circ\alpha}\|$ is upper bounded by the maximum row sum of $\boldsymbol{R}^{\circ\alpha}$, so

$$\|\boldsymbol{R}^{\circ\alpha}\| \leq \left(\frac{K}{2}\right)^{1-\alpha}\frac{\tau^\alpha\bar{\rho}^{\alpha/2}}{\rho^{\alpha/2}}$$
$$+ \max\left\{\left(\frac{K}{2}\right)^{1-\alpha}\left(1 - \frac{\bar{\rho}\tau}{\rho}\right)^{\alpha}, d^{1-\alpha}\bar{\tau}^{\alpha}\right\} \tag{118}$$
$$\leq \left(\frac{K}{2}\right)^{1-\alpha}\left(\frac{\tau^\alpha\bar{\rho}^{\alpha/2}}{\rho^{\alpha/2}} + \left(1 - \frac{\bar{\rho}\tau}{\rho}\right)^{\alpha}\right) \ , \tag{119}$$

where (119) assumes $\alpha < 1$ and that d is held constant as K increases.

- $\|\boldsymbol{R}^{\circ\alpha}\|$ is lower bounded by the operator norm of the submatrix derived from its rows and columns numbered $\left(\frac{K}{2}+1\right)$ through K. Therefore,

$$\|\boldsymbol{R}^{\circ\alpha}\| \geq \left(\frac{K}{2}\right)^{1-\alpha}\left(1 - \frac{\bar{\rho}\tau}{\rho}\right)^{\alpha} \ . \tag{120}$$

Hence, we find that for very small ρ, the ratio $\frac{\bar{\rho}\tau}{\rho}$ essentially determines the α-Rényi entropy rate for $\alpha < 1$.

The probability that the chain never visits $\left[\frac{K}{2}+1 : K\right]$ in a run of length n when initiated at the stationary distribution is

$$(1-\rho)(1-\tau)^{n-1} \geq 1 - \rho - (n-1)\tau \geq 1 - n\rho \ , \tag{121}$$

where the last inequality in (121) is obtained from the choice $0 < \rho < \frac{1}{100n}$ and $\frac{\bar{\rho}\tau}{\rho} \in [0.3, 0.4]$. Then, any such chain in this family does not visit any state $\left[\frac{K}{2}+1 : K\right]$ with probability at least 99%. Conditioned on this high probability event, the process is a random walk on the Ramanujan graph G, whose statistics do not depend on ρ or τ.

Furthermore, any such chain in this family has a spectral gap at least

$$\frac{1}{8}\left(\min\left\{0.3, \frac{0.5\times 0.5}{1.5}\left(1-\frac{2\sqrt{d-1}}{d}\right)\right\}\right)^2 \geq 10^{-5}, \tag{122}$$

where the inequality in (122) assumes $d=3$. To get a lower bound on the absolute spectral gap instead, we consider a lazy version of this reversible chain with $\boldsymbol{P}_{\text{lazy}} = \frac{1}{2}(\boldsymbol{I}+\boldsymbol{P})$, which makes all the eigenvalues of $\boldsymbol{P}_{\text{lazy}}$ non-negative. Its absolute spectral gap is then equal to its spectral gap which is half the spectral gap of $\boldsymbol{P}$. Similar estimates of the order α-Rényi entropy rate may be obtained for the lazy chain. This shows that the order-α Rényi entropy rate for $0<\alpha<1$ cannot be estimated to a given level of accuracy with a sufficiently small probability of error using a number of samples that depends only on bounds on its alphabet size and relaxation time. ■

Appendix D
Proof of Theorem 4

Proof: For $0<\alpha<\infty, \alpha\neq 1$,

$$H_\alpha(\boldsymbol{P}) = \frac{1}{1-\alpha}\log\rho(\boldsymbol{P}^{\circ\alpha}). \tag{123}$$

The spectral radius is not a norm and does not satisfy the triangle inequality. This makes it tricky to analyze estimation of the order-α Rényi entropy rate. However, for reversible Markov chains,

$$\rho\left(\boldsymbol{P}^{\circ\alpha}\right) = \rho\left(\boldsymbol{R}^{\circ\alpha}\right) = \|\boldsymbol{R}^{\circ\alpha}\|, \tag{124}$$

where $\boldsymbol{R}$ is the normalized affinity matrix for a given reversible Markov chain with transition matrix $\boldsymbol{P}$ and stationary distribution $\boldsymbol{\pi}$ on the alphabet $\mathcal{X}$ obtained by a similarity transformation of $\boldsymbol{P}$ according to

$$R_{ij} = \frac{\pi_i P_{ij}}{\sqrt{\pi_i\pi_j}}, \quad \forall\, i,j\in\mathcal{X}. \tag{125}$$

Thus, for $\alpha\in(0,1)\cup(1,\infty)$,

$$H_\alpha(\boldsymbol{P}) = \frac{1}{1-\alpha}\log\|\boldsymbol{R}^{\circ\alpha}\|. \tag{126}$$

Define now

$$N_{ij} := \left|\{1\leq t\leq n-1 : (X_t, X_{t+1}) = (i,j)\}\right|. \tag{127}$$

Let $\hat{\boldsymbol{Q}}$ denote the doublet probability matrix for the 'empirical reversible chain' given by

$$\hat{Q}_{ij} = \frac{N_{ij}+N_{ji}}{2(n-1)}, \quad \forall\, i,j\in\mathcal{X}. \tag{128}$$

The stationary probability distribution for this chain is $\boldsymbol{\pi}$ given by

$$\hat{\pi}_i = \sum_{j=1}^{K}\hat{Q}_{ij}, \quad \forall\, i\in\mathcal{X}. \tag{129}$$

Define the plug-in estimate for the normalized affinity matrix $\hat{\boldsymbol{R}}$ as

$$\hat{R}_{ij} = \frac{\hat{Q}_{ij}}{\sqrt{\hat{\pi}_i\hat{\pi}_j}}, \ \forall i,j\in\mathcal{X}. \tag{130}$$

We can then produce the estimate

$$\hat{H}_\alpha^{(n)} = \frac{1}{1-\alpha}\log\|\hat{\boldsymbol{R}}^{\circ\alpha}\|. \tag{131}$$

Now, if we let

$$\phi(\alpha,\delta) = \begin{cases}\alpha\delta, & \alpha>1,\\ \delta^\alpha & 0<\alpha<1,\end{cases} \tag{132}$$

then, for any $x,y\in[0,1]$, we have

$$|x^\alpha - y^\alpha| \leq \phi(\alpha, |x-y|). \tag{133}$$

If $\|\cdot\|_{\max}$ is the max norm (maximum absolute entry of the matrix), then

$$\begin{aligned}\|\hat{\boldsymbol{R}}^{\circ\alpha} - \boldsymbol{R}^{\circ\alpha}\| &\leq K\|\hat{\boldsymbol{R}}^{\circ\alpha} - \boldsymbol{R}^{\circ\alpha}\|_{\max} && (134)\\ &\leq K\phi(\alpha, \|\hat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max}) && (135)\\ &\leq K\phi(\alpha, \|\hat{\boldsymbol{R}} - \boldsymbol{R}\|). && (136)\end{aligned}$$

Each row sum of $\boldsymbol{P}^{\circ\alpha}$ is at least 1 if $0<\alpha<1$ and at least $\frac{1}{K^{\alpha-1}}$ if $1<\alpha<\infty$. Since $\boldsymbol{R}^{\circ\alpha}$ is symmetric and similar to $\boldsymbol{P}^{\circ\alpha}$, it follows that

$$\|\boldsymbol{R}^{\circ\alpha}\| = \rho(\boldsymbol{R}^{\circ\alpha}) = \rho(\boldsymbol{P}^{\circ\alpha}) \geq f(\alpha,K), \tag{137}$$

where

$$f(\alpha,K) = \begin{cases}1, & 0<\alpha<1,\\ \frac{1}{K^{\alpha-1}}, & 1<\alpha<\infty.\end{cases} \tag{138}$$

Similarly,

$$\|\hat{\boldsymbol{R}}^{\circ\alpha}\| = \rho(\hat{\boldsymbol{R}}^{\circ\alpha}) = \rho(\hat{\boldsymbol{P}}^{\circ\alpha}) \geq f(\alpha,K). \tag{139}$$

Thus,

$$\begin{aligned}&|\hat{H}_\alpha^{(n)} - H_\alpha(\boldsymbol{P})|\\ &= \left|\frac{1}{1-\alpha}\log\frac{\|\hat{\boldsymbol{R}}^{\circ\alpha}\|}{\|\boldsymbol{R}^{\circ\alpha}\|}\right| && (140)\\ &\leq \left|\frac{1}{1-\alpha}\right|\log\left(1+\frac{\|\hat{\boldsymbol{R}}^{\circ\alpha} - \boldsymbol{R}^{\circ\alpha}\|}{\min\{\|\boldsymbol{R}^{\circ\alpha}\|, \|\hat{\boldsymbol{R}}^{\circ\alpha}\|\}}\right) && (141)\\ &\leq \left|\frac{1}{1-\alpha}\right|\cdot\frac{\|\hat{\boldsymbol{R}}^{\circ\alpha} - \boldsymbol{R}^{\circ\alpha}\|}{\min\{\|\boldsymbol{R}^{\circ\alpha}\|, \|\hat{\boldsymbol{R}}^{\circ\alpha}\|\}} && (142)\\ &\leq \frac{\|\hat{\boldsymbol{R}}^{\circ\alpha} - \boldsymbol{R}^{\circ\alpha}\|}{f(\alpha,K)|1-\alpha|} && (143)\\ &\leq \frac{K\phi(\alpha, \|\hat{\boldsymbol{R}} - \boldsymbol{R}\|)}{f(\alpha,K)|1-\alpha|}. && (144)\end{aligned}$$

Our estimate of the normalized affinity matrix $\hat{\boldsymbol{R}}$ is very close to the estimate described by $(\mathrm{Sym}(\hat{\boldsymbol{L}}) = \frac{\hat{\boldsymbol{L}}+\hat{\boldsymbol{L}}^{\mathrm{T}}}{2})$ proposed in [15, Sec 3.2]. The difference between the two estimates is of $O\left(\frac{1}{n}\right)$ and our choice ensures that $\hat{R}$ is the normalized affinity matrix for a reversible chain. By adapting [15, Lemma 2] to our estimate, there is an absolute constant $C > 0$ such that with probability at least $1-\epsilon$,

$$\|\hat{\boldsymbol{R}} - \boldsymbol{R}\| \leq C \left(\sqrt{\frac{T_{\mathrm{rel}} \log \frac{K}{\epsilon} \log \frac{n}{\pi_* \epsilon}}{\pi_* n}} + \frac{T_{\mathrm{rel}} \log T_{\mathrm{rel}}}{n} \right). \tag{145}$$

Using (145) in (144) proves the theorem for $\alpha \in (0,1) \cup (1,\infty)$

As far as estimation of the min-entropy rate, observe that since the chain is reversible,

$$R_{ij} = \frac{\pi_i P_{ij}}{\sqrt{\pi_i \pi_j}} = \frac{\pi_j P_{ji}}{\sqrt{\pi_i \pi_j}} \tag{146}$$

$$= \sqrt{\frac{\pi_i P_{ij}}{\sqrt{\pi_i \pi_j}} \cdot \frac{\pi_j P_{ji}}{\sqrt{\pi_i \pi_j}}} = \sqrt{P_{ij} P_{ji}}. \tag{147}$$

Then, by Theorem 2,

$$H_\infty(\boldsymbol{P}) = \min_{i,j \in \mathcal{X}} \frac{1}{2} \left[\imath_{X_2|X_1}(j|i) + \imath_{X_2|X_1}(i|j) \right] \tag{148}$$

$$= \log \frac{1}{\|\boldsymbol{R}\|_{\max}}. \tag{149}$$

We can therefore, produce the estimate

$$\hat{H}_\infty^{(n)} = \log \frac{1}{\|\hat{\boldsymbol{R}}\|_{\max}}. \tag{150}$$

Since

$$\min\{\hat{\boldsymbol{R}}\|_{\max}, \|\boldsymbol{R}\|_{\max}\} \geq \frac{\min\{\|\hat{\boldsymbol{R}}\|, \|\boldsymbol{R}\|\}}{K} = \frac{1}{K}, \tag{151}$$

we get

$$|\hat{H}_\infty^{(n)} - H_\infty(\boldsymbol{P})|$$

$$= \left| \log \frac{\|\hat{\boldsymbol{R}}\|_{\max}}{\|\boldsymbol{R}\|_{\max}} \right| \tag{152}$$

$$\leq \log \left(1 + \frac{\|\hat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max}}{\min\{\|\boldsymbol{R}\|_{\max}, \|\hat{\boldsymbol{R}}\|_{\max}\}} \right) \tag{153}$$

$$\leq K \|\hat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max} \tag{154}$$

$$\leq K \|\hat{\boldsymbol{R}} - \boldsymbol{R}\|. \tag{155}$$

Using (145) in (155) proves the theorem for the min-entropy rate.

■