Complexity of Estimating Rényi Entropy of Markov Chains

Maciej Obremski*, Maciej Skorski†
* *National University of Singapore*
obremski.math@gmail.com
† *University of Luxembourg*
maciej.skorski@gmail.com

***Abstract*—Estimating entropy of random processes is one of the fundamental problems of machine learning and property testing. It has numerous applications to anything from DNA testing and predictability of human behaviour to modeling neural activity and cryptography. We investigate the problem of Rényi entropy estimation for sources that form Markov chains.**

Kamath and Verd (ISIT'16) showed that good mixing properties are essential for that task. We prove that even with very good mixing time, estimation of entropy of order $\alpha > 1$ requires $\Omega(K^{2-1/\alpha})$ samples, where K is the size of the alphabet; particularly min-entropy requires $\Omega(K^2)$ sample size and collision entropy requires $\Omega(K^{3/2})$ samples. Our results hold both in asymptotic and non-asymptotic regimes (under mild restrictions). The analysis is completed by the upper complexity bound of $O(K^2)$ for the standard plug-in estimator. This leads to an interesting open question how to improve upon a plugin estimator, which looks much more challenging than for IID sources (which tensorize nicely).

We achieve the results by applying Le Cam's method to two Markov chains which differ by an appropriately chosen sparse perturbation; the discrepancy between these chains is estimated with help of perturbation theory. Our techniques might be of independent interest.

***Index Terms*—Sample Complexity, Markov Chains, Entropy**

I. Introduction

We follow up after [16] to investigate efficiency of estimators for other popular notions of entropy - namely min-entropy, collision entropy and in general Rényi entropy.

Entropy estimation is one of the fundamental problems in the field of distribution testing. In addition to being mathematically interesting it has multiple applications to anything from DNA introns identification to predictability of human behaviour [30], [31], [49], [50], [53]. In all of those applications one could use Rényi entropy in place of Shannon entropy.

Rényi entropy [45] arises in many applications as a generalization of Shannon Entropy [48]. It is also of interests on its own right, with a number of applications including unsupervised learning (like clustering) [26], [57], multiple source adaptation [37], image processing [36], [39], [46], password guessability [3], [19], [43], network anomaly detection [35], quantifying neural activity [41] or to analyze information flows in financial data [27].

In particular Rényi entropy of order 2, known also as collision entropy, is used in quality tests for random number generators [29], [52], to estimate the number of random bits that can be extracted from a physical source [8], [23], characterizes security of certain key derivation functions [4], [12], helps testing graph expansion [15] and closeness of distributions to uniformity [6], [42] and bounds the number of reads needed to reconstruct a DNA sequence [38]. In turn the min-entropy is of fundamental importance to cryptography [47].

There are two models of randomness source which we consider when estimating entropy: model with iid samples, and one which samples from a Markov chain. Over the years asymptotic regime for iid samples got the most attention [2], [9], [13], [18], [20], [56]. More recent works consider exact, non-asymptotic behaviours of the estimators under the iid model [17], [41], [51], [55]. Only recent papers considered Rényi entropy for iid samples [1], [40].

Estimation of entropy of Markov chains is a much harder task. [28] gave Rényi entropy estimators for reversible Markov chains in a non-asymptotic regime. They also showed that there are no guarantees on the estimator for chains with bad mixing time properties. In [16] authors give bounds for Shannon entropy of Markov chains. In [21], [54] authors study a general problem of learning Markov chains from limited samples space.

In this paper we develop lower bounds on the sample complexity of Rényi entropy estimators in Markov chain models. Our results hold both when estimating the asymptotic entropy, and when estimating the entropy per symbol of a finite sample. The bounds hold even for the Markov chains with close to optimal mixing properties (i.e. are not due to badly mixing behaviors).

A. *Estimation for Iid Samples*

It is interesting to recall the lower bounds for Rényi entropy estimators sample complexity for the case of iid samples, bounds were achieved in a series of papers by [1], [40].

Entropy	Accuracy	Sample Complexity
$1 < \alpha < 2$	$\delta \leqslant 1$	$\Omega(1) \cdot \min\left(\delta^{-\frac{1}{2}}\|S\|^{\frac{1}{2}}, \delta^{-\alpha}\|S\|^{1-\frac{1}{\alpha}}\right)$
	$\delta > 1$	$\Omega(1) \cdot \min\left(\left(2^{-\delta}\|S\|\right)^{\frac{1}{2}}, 2^{-\left(1-\frac{1}{\alpha}\right)\delta}\|S\|^{1-\frac{1}{\alpha}}\right)$
$2 \leqslant \alpha$	$\delta \leqslant 1$	$\Omega(1) \cdot \delta^{-\frac{1}{\alpha}}\|S\|^{1-\frac{1}{\alpha}}$
	$\delta > 1$	$\Omega(1) \cdot \left(2^{-\left(1-\frac{1}{\alpha}\right)\delta}\|S\|\right)^{1-\frac{1}{\alpha}}$

TABLE I: Lower bounds for estimating Rényi entropy α of iid samples from a finite alphabet S, as in [40]

B. Our Results and Techniques (Rényi Entropy Rates)

We consider *irreducible and aperiodic* Markov chains. Our main results are

- we establish lower bounds for the sample complexity under Markov model of dependency, for Rényi entropy, known results only concern IID samples; we complete by upper bounds for the natural "plugin" estimator.
- we show that those bounds hold both when estimating asymptotic entropy of Markov chain, and when estimating entropy of fixed-length paths

Our techniques

- we develop a lemma on *closeness of sample paths of two chains*; it non-trivially extends the classical result on the distance two IID sequences and is of independent interest. The motivation is to make Le Cam's method work on Markov chains - for the IID case it is easier as certain discrepancies like KL or Hellinger tensorize, here not.
- We use *perturbation theory* to get insights into spectral properties of matrices; this simplifies otherwise complicated calculations and is of independent interest.

Theorem 1 (Lower Bounds for Asymptotic MC Entropy Estimation)**.** *For any state space S and any estimator of the entropy rate of Markov chains on S, the minimum number of samples to achieve a constant additive error is as in Table II. This holds even for chains with contant spectral gap (which mix quickly).*

Renyi Entropy	Min. num. of samples
$\mathbf{H}_\infty$	$\Omega\left(\lvert S\rvert^2\right)$
$\mathbf{H}_2$	$\Omega\left(\lvert S\rvert^{\frac{3}{2}}\right)$
$\mathbf{H}_\alpha \quad (1<\alpha<\infty)$	$\Omega\left(\lvert S\rvert^{2-\frac{1}{\alpha}}\right)$

TABLE II: Lower Bounds for Entropy Estimation under Markov Chain Model on State Space S and spectral gap 0.01.

Theorem 2 (Lower Bounds for Finite-Sample MC Entropy Estimation)**.** *Bounds from Theorem 1 apply to entropy-per-symbol of fixed-length samples, assuming a) the entropy order $1.001 < \alpha$ b) the starting distribution probability masses are at least $|S|^{-O(1)}$ c) the length of samples $n = \omega(\log |S|)$.*

For illustration, we prove the upper bound $\tilde{O}(|S|^2)$ under certain restrictions (relaxing them is outside of the scope).

Theorem 3 (Plugin Estimator for MC Entropy)**.** *The plug-in estimator achieves an additive error of $\epsilon = 0.001$ with $n = O(|S|^2)\log^{O(1)}(|S|)$ samples, assuming a) the entropy order $\alpha > 1.001$ b) the spectral gap is $\Omega(1)$ c) the stationary probability masses are $\Omega(|S|^{-1})$ d) the transition matrix entries are $\Omega(|S|^{-1})$.*

II. Preliminaries

A. Notation

By $\mathbf{1}_{p,q}$ we denote the matrix of ones of size $p \times q$. By $\mathbf{0}_{p,q}$ we denote the matrix of zeros of size $p \times q$. By I_p we denote the identity matrix of size $p \times p$. By A^T we denote the transpose of A.
The spectral radius of M is denoted by $\rho(M)$. The α-th Hadamard power of M is defined as $(M^{\diamond\alpha})_{i,j} = (M_{i,j})^\alpha$ (the entry-wise power).
Matrix norms induced by vector p-th norms are denoted as usual by $\|\cdot\|_p$.
Troughout the paper by the transition matrix M of a Markov chain X we understand the matrix $M_{s_1,s_2} = \Pr[X_n = s_2|X_{n-1} = s_1]$; by the definition of MC it doesn't depend on n. Note that in our convention M is row-stochastic.

B. Entropy Rates

For a single distribution X the Rényi entropy of order $\alpha > 1$ is defined as $\mathbf{H}_\alpha(X) = -\frac{1}{\alpha-1}\log\sum_x \Pr[X = x]^\alpha$. In the case $\alpha = \infty$ we obtain (in the limit) the min-entropy $\mathbf{H}_\infty(X) = -\log\max_x \Pr[X = x]$. The entropy rate of a stochastic process $X_1, X_2, \ldots$ is the limiting entropy per symbol $\frac{1}{n}\mathbf{H}_\alpha(X_1, \ldots, X_n)$ (where α may be also $\alpha = \infty$). For Markov chains this limit exists under standard conditions (irreducibility, aperiodicity) and can be explicitly evaluated.

1) Entropy $\mathbf{H}_\infty$: It is known that the min-entropy rate of a markov chain is determined by the average heaviest cycle [28]. The average weight of a cycle $\mathcal{C} = s_0 \to s_1 \to s_2 \to \ldots s_n = s_0$ is defined as $\mathbf{w}(\mathcal{C}) = \left(\prod_{i=1}^n M(s_{i-1}, s_i)\right)^{\frac{1}{n}}$ where M is the transition matrix; the entropy rate equals

$$\mathbf{H}_\infty(M) = -\log\max_{\mathcal{C}} \mathbf{w}(\mathcal{C}). \tag{1}$$

2) Entropy $\mathbf{H}_\alpha$: To evaluate the limiting Rényi entropy of order α, one considers spectral properties of the *Hadamard power* of the chain transition matrix M. Namely [44]

$$\mathbf{H}_\alpha(M) = \frac{1}{1-\alpha}\log\rho\left(M^{\diamond\alpha}\right) \tag{2}$$

where $\rho(\cdot)$ denotes the spectral radius of a matrix.

C. Le Cam's method

The popular technique of proving lower bounds on a minimax estimator is to find two sample distributions such that (a) they are statistically close and (b) the true values of estimated parameters or functionals are far away.

Since the values of estimated parameters are far away, we can use the estimator as a distinguisher between two sample distributions. But the samples are close together (say ϵ-close) thus any distinguisher with constant chance of success requires at least $\Omega(1/\epsilon)$ samples, which provides lower bound.

D. Perturbation Theory

The spectrum of a matrix remains (somewhat) stable under perturbations. There are many results of this form and we refer to [24], [25], [58] for more details and a survey; for our needs the classical result due to *Bauer-Firke* will be enough.

Lemma 1 (Bauer-Firke Eigenvalue Perturbation [7])**.** *If A is a real normal matrix, that is $AA^T = A^TA$ then each eigenvalue of the matrix $A + E$ is at most δ-apart from some eigenvalue of A, where $\delta = \|E\|_2$.*

Also the perturbations of eigenvectors have been studied. We recall bounds depending on a *hitting times*

Lemma 2 (Perturbation of MC stationary distributions [10]). *The stationary distribution before and after the perturbation by a matrix E differ in ℓ_1-norm by at most $\kappa \cdot \|E\|_\infty$, for any κ such that $\max_{i,j} \frac{m_{i,j}}{m_{j,j}} \leqslant 2\kappa$ where $m_{i,j}$ is the expected time of hitting j when the chain starts from i.*

E. Coupling

Coupling refers to building joint distribution with given marginals and is very useful in studying Markov chains [14]. The following slightly extends the standard construction

Proposition 1 (Consistent Coupling). *For any four discrete random variables X_1, X_2, Y_1, Y_2 there exist distributions X_1', X_2', Y_1', Y_2' over same probability space, such that $X_1' = X_1, X_2' = X_2, Y_1' = Y_1, Y_2' = Y_2$ and $\Pr[X_1' \neq Y_1'] = d_{TV}(X_1, Y_1)$.*

F. Chernoff-Type Bounds for Markov Chains

Chernoff-type bounds hold for Markov chains with exponentially small tails, but the constant depends on spectral properties of the transition matrix [34] or on (related) mixing times [11]. We use them to estimate the transition matrix and analyze the plugin estimator.

III. Results and Proofs

A. Sample Paths of Perturbed Markov Chains

The lemma below states that sample paths of two chains with close transition matrices remains statistically close, when the number of samples is not too big.

Lemma 3 (Total Variation of Markov Chains with Close Transitions). *Consider two Markov chains with transition matrices M and $M+E$, starting from their stationary distributions μ^M, μ^{M+E}. The total variation δ between $n+1$ samples from both chains is bounded by*

$$\delta \leqslant \|\mu^M - \mu^{M+E}\|_1 + n \cdot (\mu^M)^T \cdot |E| \cdot \mathbf{1}$$

where $|E|$ is the matrix of absolute entries of E and $\mathbf{1}$ is the vector of ones.

Before we proceed to the proof let us make few remarks.

Remark 1 (Sparsity of Perturbation Helps). *Note that $(\mu^M)^T \cdot |E| \cdot \mathbf{1}$ is a combination of row-sums of $|E|$ with weights μ^M. For fixed μ the mapping $E \to \mu^T \cdot |E| \cdot \mathbf{1}$ is a matrix norm which captures sparsity.*

Remark 2 (Bounds for IID distributions). *Consider the following matrices $M_X = \begin{bmatrix} \frac{1}{m-\ell}\mathbf{1}_{m,m-\ell} & \mathbf{0}_{m,\ell} \end{bmatrix}$ and $M_Y = \begin{bmatrix} \mathbf{0}_{m,\ell} & \frac{1}{m-\ell}\mathbf{1}_{m,m-\ell} \end{bmatrix}$. They describe IID distributions μ^X uniform over $1,\ldots,m-\ell$ and μ^Y uniform over $\ell,\ldots,m$ respectively. We can write $M_Y = M_X + E$ where $E = \begin{bmatrix} -\frac{1}{m-\ell}\mathbf{1}_{m,\ell} & \mathbf{0}_{m,m-2\ell} & \frac{1}{m-\ell}\mathbf{1}_{m,\ell} \end{bmatrix}$. Applying Lemma 3 we get that the total variation between n samples from X and n samples from Y is bounded by $n \cdot \frac{\ell}{m-\ell} = n \cdot d_{\mathrm{TV}}(\mu^X; \mu^Y)$, as in the standard bound for the distance of IID variables.*

We give two proofs of Lemma 3- one by a coupling, the other by a dynamic programming technique where the distance for n samples is expressed in terms of the distance of $n-1$ samples, and the connection is explicit due to factorization of finite-sample distributions under the Markov assumption.

by Coupling. Let $X_0,\ldots,X_n$ and $Y_0,\ldots,Y_n$ be samples from two Markov chains with transition matrices M_X and M_Y respectively; let $X_{\leqslant k} = (X_1,\ldots,X_k)$. For any coupling

$$\begin{aligned}
&d_{TV}(X_{\leqslant n}, Y_{\leqslant n}) = \\
&\Pr[X_{\leqslant n-1} = Y_{\leqslant n-1}] \cdot d_{\mathrm{TV}}(X_n; Y_n | X_{\leqslant n-1} = \\
&Y_{\leqslant n-1} + \Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}] \cdot \\
&\quad \cdot d_{\mathrm{TV}}(X_n; Y_n | X_{\leqslant n-1} \neq Y_{\leqslant n-1}) \\
&\leqslant d_{\mathrm{TV}}(X_n; Y_n | X_{n-1} = Y_{n-1}) + \Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}] \quad (3)
\end{aligned}$$

where we used $d_{\mathrm{TV}}(X_n; Y_n | X_{\leqslant n-1} = Y_{\leqslant n-1}) = d_{\mathrm{TV}}(X_n; Y_n | X_{n-1} = Y_{n-1})$ which follows from the Markov property. For two Markov matrices M_X, M_Y and any (common) distribution μ we have

$$\|\mu^T \cdot (M_X - M_Y)\|_1 \leqslant \mu^T \cdot |M_X - M_Y| \cdot \mathbf{1}$$

If X starts from the stationary distribution μ^X we have $X_n \overset{d}{=} \mu^X$ for all n. Therefore

$$d_{TV}(X_n; Y_n | X_{n-1} = Y_{n-1}) \leqslant (\mu^X)^T \cdot |M_X - M_Y| \cdot \mathbf{1} \quad (4)$$

There is a coupling such that

$$\Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}] = d_{\mathrm{TV}}(X_{\leqslant n-1}, Y_{\leqslant n-1}) \qquad (5)$$

Putting Equations (4) and (5) into Equation (3) we get

$$\begin{aligned}
&d_{\mathrm{TV}}(X_{\leqslant n}, Y_{\leqslant n}) \leqslant \\
&\mu^T \cdot |M_X - M_Y| \cdot \mathbf{1} + \Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}].
\end{aligned}$$

so that the statement follows by induction. □

by Dynamic Programming. Consider the total variation distance of $n+1$ samples, and let μ^M, μ^{M+E} be as in Lemma 3.

$$d_{\mathrm{TV}}^n = \sum_{s_0,\ldots,s_n} \left| \mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i} - \mu^{M+E}_{s_0} \prod_{i=1}^n (M+E)_{s_{i-1},s_i} \right|$$

Consider $\mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i}$ as the difference between $\mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i} \cdot \left(M_{s_{n-1},s_n} + E\right)$ and $\mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i} \cdot E$; then by the triangle inequality $d_{TV} \leqslant I_1 + I_2$ where I_1 equals:

$$\sum_{s_0,\ldots,s_{n-1}} \left| \mu^M_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} - \mu^{M+E}_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} \right| \cdot \|M+E\|_\infty$$

with $\|M+E\|_\infty = \max_{s_{n-1}} \sum_{s_n} |(M+E)_{s_{n-1},s_n}|$ and

$$I_2 = \sum_{s_0,\ldots,s_{n-1}} \mu^M_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} \cdot \sum_{s_n} |E_{s_{n-1},s_n}|$$

with $\|E\|_\infty = \max_{s_{n-1}} \sum_{s_n} |E_{s_{n-1},s_n}|$. Observe that $\|M + E\|_\infty = 1$ because $M + E$ is stochastic. Thus I_1 is at most

$$\sum_{s_0,\ldots,s_{n-1}} \left| \mu^M_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} - \mu^{M+E}_{s_0} \prod_{i=1}^{n-1} (M+E)_{s_{i-1},s_i} \right| = \tag{6}$$

$$= d^{n-1}_{\mathrm{TV}}$$

If μ^M is stationary for M then by Chapman-Kolmogorov

$$\begin{aligned} I_2 &= (\mu^M)^T \cdot (M+E)^{n-1} \cdot |E| \cdot \mathbf{1} \\ &= (\mu^M)^T \cdot M^{n-1} \cdot |E| \cdot \mathbf{1} \\ &= (\mu^M)^T \cdot |E| \cdot \mathbf{1} \end{aligned}$$

Summing up we get

$$d^n_{\mathrm{TV}} \leqslant d^{n-1}_{\mathrm{TV}} + (\mu^M)^T \cdot |E| \cdot \mathbf{1}$$

which by induction implies the statement. □

B. Construction of Extreme Matrix

From now on we assume that the state space has $|S| = m$ elements. We apply Le Cam's method to two Markov chains:

- the random walk with uniform transitions $\frac{1}{m}\mathbf{1}_{m,m}$
- perturbation of the uniform random walk which over-weights one element. For a parameter $0 < \epsilon < \frac{1}{2}$ the transition matrix of this chain is defined as

$$M = \begin{bmatrix} \frac{1}{m}\mathbf{1}_{m-1,m-1} & \frac{1}{m}\mathbf{1}_{m-1,1} \\ \left(\frac{1}{m} - \frac{\epsilon}{m-1}\right)\mathbf{1}_{1,m-1} & \left(\frac{1}{m}+\epsilon\right) \end{bmatrix} \tag{7}$$

Our perturbation is sparse (affects only one row and one column), and thus we expect the change in the distance of finite samples to be small. On the other hand it will have a significant effect on the spectrum of Hadamard powers.

C. Mixing Time is Good

[28] showed that bad mixing properties heavily impact the efficiency of an estimator. Here we argue that Markov chains we mentioned above have very good mixing times, thus concluding that estimation of entropy is still hard even when restricted to Markov chains with good mixing properties.

For the unperturbed matrix eigenvalues are 1 (single) and 0 (multiplicity of $m-1$) (this follows from known properties of matrix of ones [22]); it follows that the spectral gap is one. After the perturbation we maintain the *constant spectral gap*; by perturbation theory (Lemma 1) eigenvalues changes by at most $\|E\| = O(m^{-1/2})$, smaller than 0.01 for sufficiently big m. We avoided calculating eigenvalues explicitly.

D. Entropy Rates

In this section we prove Theorem 1, our result on *entropy rates*. Entropy rates for stochastic sources are understood as the limiting entropy per symbol (for Markov chains they exist under standard assumptions such as ergodicity).

1) Rate Evaluation for $\boldsymbol{H}_\infty$: We find the change in the entropy rate and statistical distance when setting $\epsilon > 0$ and $\epsilon = 0$ in Equation (7).

Claim 1 (Min-Entopy Rate). *For the chain with transition matrix as in* (7)

$$\mathbf{H}_\infty(M) = -\log\left(\frac{1}{m}+\epsilon\right)$$

Proof. The heaviest cycle is the self-loop at the m-th state. □

Claim 2 (Statistical Distance Closeness). *The variational distance between n samples from M in Equation (7) and the random walk, assuming both chains start from their stationary distributions, is bounded by $O(\epsilon + n\epsilon/m)$.*

Proof. This follows from Lemma 3 applied to M being the matrix of the random walk and E equal to

$$E = \begin{bmatrix} \mathbf{0}_{m-1,m-1} & \mathbf{0}_{m-1,1} \\ -\frac{\epsilon}{m-1}\mathbf{1}_{1,m-1} & \epsilon \end{bmatrix}$$

Since $\mu^M = \frac{1}{m}\mathbf{1}_{m,1}$ we get

$$\mu^M \cdot |E| \cdot \mathbf{1}_{m,1} = O(\epsilon/m)$$

The distance between stationary distributions can be bounded by $O(\epsilon)$ according to Lemma 2. □

Corollary 1 (Entropy Separation). *If we take $\epsilon = 1/m$, then min-entropy of perturbed chain will be $\log(\frac{m}{2})$ while min-entropy of uniformly random walk remains $\log(m)$, thus the min-entropies of two Markov chains differ by 1.*

Corollary 2 (Statistical Distance). *Let $\epsilon = \frac{1}{m}$, by Claim 2 the distance between n samples is bounded by $O(n \cdot m^{-2})$.*

By the above corollaries and the Le Cam's method described in Section II-C we get our lower bound for min-entropy.

2) Rate Evaluation for $\boldsymbol{H}_2$: Below we estimate the difference in entropy and closeness in statistical distance for these two chains, summarizing in Corollary 4 and Corollary 3.

Lemma 4 (Spectral Radius). *For the matrix in Equation (7). the spectral radius of its second Hadamard power is*

$$\rho(M^{\circ 2}) = \max\left(\frac{1}{m}, \left(\frac{1}{m}+\epsilon\right)^2\right) + O(m^{-\frac{3}{2}})$$

More generally, the eigenvalues are $O(m^{-\frac{3}{2}})$ (with $m-2$ repeats), $\frac{1}{m} + O(m^{-\frac{3}{2}})$ and $\left(\frac{1}{m}+\epsilon\right)^2 + O(m^{-\frac{3}{2}})$.

Corollary 3 (Entropy Separation). *For $\epsilon = \sqrt{2/m}$ one obtains $\rho(M^{\circ 2}) = \frac{2+o(1)}{m}$ for large m. For $\epsilon = 0$ we have $\rho(M^{\circ 2}) = \frac{1}{m}$. Therefore collision entropy rates of these two Markov chains differ by at least 1 bit.*

Corollary 4 (Statistical Distance). *By Claim 2, for $\epsilon = \sqrt{2/m}$ the distance between n samples is bounded by $O(n \cdot m^{-3/2})$.*

Again by applying the Le Cam's method described in Section II-C to above corollaries we get our lower bound for collision entropy claimed in Theorem 1.

Proof of Lemma 4. We have

$$M^{\diamond 2} = \begin{bmatrix} \frac{1}{m^2}\mathbf{1}_{m-1,m-1} & \frac{1}{m^2}\mathbf{1}_{m-1,1} \\ \left(\frac{1}{m}-\frac{\epsilon}{m-1}\right)^2 \mathbf{1}_{1,m-1} & \left(\frac{1}{m}+\epsilon\right)^2 \end{bmatrix}$$

To compute the spectral radius of $M^{\diamond 2}$ we write

$$M^{\diamond 2} = Z + E$$

where Z is the block-diagonal matrix given by

$$Z = \begin{bmatrix} \frac{1}{m^2}\mathbf{1}_{m-1,m-1} & \mathbf{0}_{m-1,1} \\ \mathbf{0}_{1,m-1} & \left(\frac{1}{m}+\epsilon\right)^2 \end{bmatrix}$$

and E has non-zero elements only in the last row and column, of magnitude $O(m^{-2})$ (we assume $\epsilon = O(m^{-1/2})$. In particular we obtain $\|E\|_2 \leqslant O(m^{-\frac{3}{2}})$ (for example by bounding the Frobenius norm which in turn bounds the second norm) and by Lemma 1 (Z is symmetric hence normal!)

$$\rho(M^{\diamond 2}) = \rho(Z) + O(m^{-\frac{3}{2}})$$

so that we can focus on finding the spectrum of Z. But they follow from the block-diagonal structure - the first $m-1 \times m-1$ minor has eigenvalues $\frac{m-1}{m^2}$ (simple) and 0 (repeated $m-2$ times); the m-th eigenvalue is $\left(\frac{1}{m}+\epsilon\right)^2$. In view of the previous bound this finishes the proof. □

3) Rate Evaluation for $\boldsymbol{H}_\alpha$, $1 < \alpha < \infty$: We proceed as for $\mathbf{H}_2$. Now Z has same structure but the power of 2 is replaced by α; also $\|E\|_2 = O((m \cdot m^{-2\alpha})^{1/2})$. Thus

$$\rho(M^{\diamond\alpha}) = \max\left(\frac{1}{m^{\alpha-1}}, \left(\frac{1}{m}+\epsilon\right)^\alpha\right) + O(m^{\frac{1}{2}-\alpha})$$

We choose $\epsilon = (2/m)^{\frac{\alpha-1}{\alpha}}$ then

$$\rho(M^{\diamond\alpha}) \geqslant (2/m)^{\alpha-1}(1 + O(m^{\frac{3}{2}-2\alpha}))$$

Since $\alpha \geqslant 1$ we have $O(m^{\frac{3}{2}-2\alpha}) = o(1)$ for large m. Thus for the two paths studied in Le Cam's method entropy rates are $\log(m/2)+o(1)$ and $\log m$, differring by at approximately 1 while the statistical distance is $O(n \cdot m^{-2+\frac{1}{\alpha}})$.

E. Upper Bounds

We sketch a proof for Theorem 3 when $\alpha < \infty$. The pluggin estimator is using the empirical (maximum-likelihood) estimate of the transition matrix in Equation (1) or Equation (2).

Remark 3. *The estimator is used to validate physical random number generators [5], without a proof or reference.*

Let M and μ be the transition matrix and stationary distribution for X. Consider the *two-step chain* $Y_n = (X_{n-1}, X_n)$ on $S \times S$. The stationary distribution of Y over (s_1, s_2) is such that s_1 follows μ and then probability of s_2 given s_1 equals $M(s_1, s_2)$. It is easy to see that X_n and μ are 0.1-close in d_{TV} then also Y_{n+1} is 0.1-close to its stationary distribution; on the other hand Y_{n+1} is 0.1-close to its stationary distribution then X_n is 0.1-close to μ. Thus the mixing times differ at most by 1. Relating them to spectral gaps [33] we get that the $\Omega(1)$ gap in X implies $\mathrm{polylog}(|S|)$ mixing time for Y; this uses the fact that the probability masses after first step are $|S|^{-O(1)}$.

Let $n = O(\epsilon^{-2}m^2)\log^{O(1)} m$ with sufficiently big constants, $m = |S|$. We estimate frequencies of single symbols s_1 from X, with a relative error of ϵ/m, and frequencies of tuples (s_1, s_2) from Y with relative error $O(\epsilon)$ by Hoeffding-type bounds [11], [34]; this holds simultaneously for all frequencies w.h.p. We then know the transition matrix up to additive error $O(\epsilon/m)$ (element-wise). By our assumptions this gives the relative error $O(\alpha\epsilon)$ for the Hadamard power. The spectral radius is monotone on non-negative matrices [22] so the estimated matrix when plugged in Equation (2) gives the entropy rate up to additive error $O(\alpha\epsilon/(\alpha-1)) = O(\epsilon)$.

F. Finite Sample Lower Bounds

Our bounds were derived for the *asymptotic entropy rate*, but they remain valid also for the task of estimating entropy of *finite* number of samples. Here we prove Theorem 2.

For $\alpha = \infty$ this holds because the entropy of n samples for both matrices considered equals n times the entropy rate. Indeed, the min-entropy of n samples from the chain with the transitions as in Equation (7) is full when $\epsilon = 0$ and otherwise it is achieved for n repetitions of the m-th symbol.

For $\alpha < \infty$ we give a reduction, invoking the derivation of Equation (2) [32]; let $Z = M^{\diamond\alpha}$, then the entropy per sample H_n of the sequence $X_1, ..., X_n$ satisfies $2^{-H_n \cdot n \cdot (\alpha-1)} = \mu^T \cdot Z^n \cdot \mathbf{1}$ where μ is the starting distribution. Thus

$$H_n = -\frac{1}{\alpha-1} \cdot \frac{\log(\mu^T \cdot Z^n \cdot \mathbf{1})}{n}$$

Note that $\mathbf{1}^T \cdot Z^n$ and $\mu^T \cdot Z^n \cdot \mathbf{1}$ differ by a factor at most $m^{-O(1)}$ because of the assumption b). Next, the mapping $X \to \mathbf{1}^T X \mathbf{1}$ for non-negative X is away by a factor $m^{O(1)}$ from the Frobenius norm of X and by another factor $m^{O(1)}$ from the spectral norm, by the known equivalence of matrix norms. Therefore $\mu^T \cdot Z^n \cdot \mathbf{1} \geqslant m^{-O(1)}\|Z^n\|$ which combined with the above formula for H_n gives

$$H_n \leqslant \log(\|Z^n\|)/((1-\alpha)n + O(\log m)/((1-\alpha)n)$$

The second term is $o(1)$ bcause of a) and c). Since $\|Z^n\| \geqslant \rho(Z^n)$ (for any matrix norm) and $\rho(Z^n) = \rho(Z)^n$ (by the Jordan form) we eventually get $H_n \leqslant H + o(1)$, where $H = -\frac{\log\rho}{n(1-\alpha)}$ is the asymptotic entropy rate.

Revise now the application of Le Cam's method with same matrices. The claim on statistical distances is unchanged. As for the entropy, for the perturbed matrix is at most $H + o(1) = \log(m) - 1 + o(1)$ by the above analysis, and for the unperturbed matrix equals $\log m$; this gives the gap of $1 - o(1)$ so the same bounds apply and we conclude Theorem 2.

IV. Conclusions

We have shown upper and lower bounds for Rényi entropy rate estimation under the Markov chain model.

V. ACKNOWLEDGEMENTS

This work was supported in part by the Luxembourg National Research Fund under grant no. C17/IS/11613923. MO was supported by MOE2019-T2-1-145 Foundations of quantum-safe cryptography grant.

REFERENCES

[1] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rényi entropy. In *SODA '15*.

[2] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms*, 19(3-4):163–193, October 2001.

[3] Erdal Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Information Theory*, 42(1):99–105, 1996.

[4] Boaz Barak, Yevgeniy Dodis, Hugo Krawczyk, Olivier Pereira, Krzysztof Pietrzak, François-Xavier Standaert, and Yu Yu. Leftover hash lemma, revisited. In *CRYPTO'11*.

[5] Elaine Barker and John Kelsey. Recommendation for the entropy sources used for random bit generation. *Draft NIST Special Publication*, 2012.

[6] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4:1–4:25, 2013.

[7] F. L. Bauer and C. T. Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, Dec 1960.

[8] Charles H. Bennett, Gilles Brassard, Claude Crépeau, and Ueli M. Maurer. Generalized privacy amplification. *IEEE Trans. Information Theory*, 41(6):1915–1923, 1995.

[9] Haixiao Cai, S. R. Kulkarni, and S. Verdu. Universal entropy estimation via block sorting. *IEEE Trans. Inf. Theor.*, 50(7).

[10] Grace E. Cho, Carl D. Meyer, Carl, and D. Meyer. Comparison of perturbation bounds for the stationary distribution of a markov chain. *Linear Algebra Appl*, 335:137–150, 2000.

[11] Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. In *STACS'12*.

[12] Yevgeniy Dodis and Yu Yu. Overcoming weak expectations. In *Theory of Cryptography Conference*, pages 1–22. Springer, 2013.

[13] Michelle Effros. Universal lossless source coding with the burrows wheeler transform. In *Proceedings of the Conference on Data Compression*, DCC '99.

[14] den Hollander Frank. *http://websites.math.leidenuniv.nl/probability/lecturenotes/CouplingLectures.pdf*, 2010.

[15] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

[16] Yanjun Han, Jiantao Jiao, Chuan-Zheng Lee, Tsachy Weissman, Yihong Wu, and Tiancheng Yu. Entropy rate estimation for markov chains with large state space. 02 2018.

[17] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61, 11 2014.

[18] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over lipschitz balls. 11 2017.

[19] Manjesh Kumar Hanawal and Rajesh Sundaresan. Guessing revisited: A large deviations approach. *IEEE Trans. Information Theory*, 57(1):70–78, 2011.

[20] Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems 2019*.

[21] Yi Hao, Alon Orlitsky, and Venkatadheeraj Pichapati. On learning markov chains. In *NeurIPS*, 2018.

[22] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Matrix Analysis. Cambridge University Press, 2013.

[23] Russell Impagliazzo and David Zuckerman. How to recycle random bits. In *FOCS 1989*.

[24] Ilse CF Ipsen. Relative perturbation results for matrix eigenvalues and singular values. *Acta numerica*, 7, 1998.

[25] ILSE C.F. IPSEN. Departure from normality and eigenvalue perturbation bounds, 2003.

[26] Robert Jenssen, KE Hild, Deniz Erdogmus, Jose C Principe, and Torbjørn Eltoft. Clustering using renyi's entropy. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 1, pages 523–528. IEEE, 2003.

[27] Petr Jizba, Hagen Kleinert, and Mohammad Shefaat. Rnyis information transfer between financial time series. *Physica A: Statistical Mechanics and its Applications*, 391(10):2971 – 2989, 2012.

[28] S. Kamath and S. Verd. Estimation of entropy rate and rnyi entropy rate for markov chains. In *ISIT 2016*.

[29] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. 1998.

[30] Coco Krumme, Alejandro Llorente, Manuel Cebrian, Alex Pentland, and Esteban Moro. The predictability of consumer visitation patterns. *Scientific reports*, 3:1645, 04 2013.

[31] J. Kevin Lanctôt, Ming Li, and En-Hui Yang. Estimating dna sequence entropy. In *SODA*, 2000.

[32] P.D. Lax. *Functional analysis*. Pure and applied mathematics. Wiley, 2002.

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

[34] Pascal Lezaud. Chernoff-type bound for finite markov chains. *Ann. Appl. Probab.*, 8(3):849–867, 08 1998.

[35] Ke Li, Wanlei Zhou, Shui Yu, and Bo Dai. Effective ddos attacks detection using generalized entropy metric. In *ICA3PP 2009*.

[36] Bing Ma, Alfred Hero, John Gorman, and Olivier Michel. Image registration with minimum spanning tree algorithm. In *Proceedings 2000 International Conference on Image Processing*.

[37] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *UAI 2009*.

[38] Abolfazl S. Motahari, Guy Bresler, and David N. C. Tse. Information theory of DNA shotgun sequencing. *IEEE Trans. Information Theory*, 59(10):6273–6289, 2013.

[39] Huzefa Neemuchwala, Alfred O. Hero III, Sakina Zabuawala, and Paul L. Carson. Image registration methods in high-dimensional space. *Int. J. Imaging Systems and Technology*, 16(5):130–145, 2006.

[40] Maciej Obremski and Maciej Skorski. Renyi entropy estimation revisited. *Random Approx*, 2017:588, 2017.

[41] Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, June 2003.

[42] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Information Theory*, 54(10):4750–4755, 2008.

[43] C. E. Pfister and W. G. Sullivan. Rényi entropy, guesswork moments, and large deviations. *IEEE Trans. Information Theory*, 50(11), 2004.

[44] Ziad Rached, Fady Alajaji, and L. Lorne Campbell. Rényi's divergence and entropy rates for finite alphabet markov sources. *IEEE Trans. Information Theory*, 47(4):1553–1561, 2001.

[45] A. Rényi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960.

[46] Prasanna K. Sahoo and Gurdial Arora. A thresholding method based on two-dimensional renyi's entropy. *Pattern Recognition*, 37(6):1149–1161, 2004.

[47] Ronen Shaltiel. An introduction to randomness extractors. In *International colloquium on automata, languages, and programming*, pages 21–41. Springer, 2011.

[48] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[49] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Laszlo Barabasi. Limits of predictability in human mobility. *Science (New York, N.Y.)*, 327:1018–21, 02 2010.

[50] Taro Takaguchi, Mitsuhiro Nakamura, Nobuo Sato, Kazuo Yano, and Naoki Masuda. Predictability of conversation partners. *Computing Research Repository - CORR*, 1, 04 2011.

[51] 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 *STOC'11*.

[52] Paul C. van Oorschot and Michael J. Wiener. Parallel collision search with cryptanalytic applications. *J. Cryptology*, 12(1):1–28, 1999.

[53] Chunyan Wang and Bernardo Huberman. How random are online social interactions? *Scientific reports*, 2:633, 09 2012.

[54] G. Wolfer and A. Kontorovich. Minimax learning of ergodic markov chains. In *International Conference on Algorithmic Learning Theory'19*.

[55] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62, 07 2014.

[56] A. D. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258, Nov 1989.

[57] Dongxin Xu. *Energy, Entropy and Information Potential for Neural Computation*. PhD thesis, Gainesville, FL, USA, 1998. AAI9935317.

[58] X. Zhan. *Matrix Theory*. Graduate Studies in Mathematics. American Mathematical Society, 2013.