Complexity of Estimating Renyi Entropy of Markov Chains

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 Renyi entropy estimation for sources that form Markov chains.

Kamath and Verdú (ISIT'16) showed that good mixing properties are essential for that task. We show that even with very good mixing time, estimation of min-entropy requires $\Omega(K^2)$ sample size, while collision entropy requires $\Omega(K^{3/2})$ samples, where K is the size of the alphabet. Our results hold both in asymptotic and non-asymptotic regimes.

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.

1 Introduction

We follow up after [Han et al., 2018] to investigate efficiency of estimators for other popular notions of entropy - namely min-entropy and collision 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 [Lanctôt et al., 2000; Song et al., 2010; Takaguchi et al., 2011; Wang and Huberman, 2012; Krumme et al., 2013]. In all of those applications one could easily replace Shannon entropy with any other Renyi entropy.

Renyi entropy [Rényi, 1960] arises in many applications as a generalization of Shannon Entropy [Shannon, 2001]. It is also of interests on its own right, with a number of applications including unsupervised learning (like clustering) [Xu, 1998; Jenssen et al., 2003], multiple source adaptation [Mansour et al., 2009], image processing [Ma et al., 2000; Neemuchwala et al., 2006; Sahoo and Arora, 2004], password guessability [Arikan, 1996; Pfister and Sullivan, 2004; Hanawal and Sundaresan, 2011], network anomaly detection [Li et al., 2009], quantifying neural activity [Paninski, 2003] or to analyze information flows in financial data [Jizba et al., 2012].

In particular Renyi entropy of order 2, known also as collision entropy, is used in quality tests for random number generators [Knuth, 1998; van Oorschot and Wiener, 1999], to estimate the number of random bits that can be extracted from a physical source [Impagliazzo and Zuckerman, 1989; Bennett et al., 1995], characterizes security of certain key derivation functions [Barak et al., 2011; Dodis and Yu, 2013], helps testing graph expansion [Goldreich and Ron, 2011] and closeness of distributions to uniformity [Batu et al., 2013; Paninski, 2008] and bounds the number of reads needed to reconstruct a DNA sequence [Motahari et al., 2013].

There are two models of randomness source which we consider when estimating entropy: model with iid samples, and one where samples form a Markov chain. Over the years asymptotic regime for iid

Preprint. Under review.

samples got the most attention [Wyner and Ziv, 1989; Antos and Kontoyiannis, 2001; Effros, 1999; Cai et al., 2006; Han et al., 2017]. More recent work considers an exact, non-asymptotic behaviour of the estimators for iid case [Paninski, 2003; Valiant and Valiant, 2011; Wu and Yang, 2014; Han et al., 2014]. Only recent papers considered Renyi entropy for iid samples [Acharya et al., 2015; Obremski and Skorski, 2017].

Estimation of entropy of Markov chains is a much harder task. [Kamath and Verdú, 2016] gave Renyi entropy estimators for reversible Markov chains in a non-asymptotic regime. They also showed that giving any guarantees on the estimator is impossible for chains with bad mixing time properties. In [Han et al., 2018] authors give bounds for Shannon entropy of Markov chains.

In this paper we investigate lower bounds on sample complexity of Renyi entropy estimator in Markov chain model. Our results hold both when estimating asymptotic entropy of Markov chain, and when estimating entropy of any fixed number of steps with any starting distribution. Our bounds hold even for the Markov chains with close to optimal mixing properties.

1.1 Estimation for Iid Samples

It is interesting to recall the lower bounds for Renyi entropy estimators sample complexity for the case of iid samples, bounds were achieved in a series of papers by [Acharya et al., 2015; Obremski and Skorski, 2017].

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

Table 1: Lower bounds for Renyi entropy α and iid samples from an alphabet of size K, as in [Obremski and Skorski, 2017]

1.2 Our Results and Techniques (Renyi Entropy Rates)

Our main results

- we establish lower bounds for the sample complexity under Markov model of dependency, for Renyi entropy, known results only concern IID samples
- we show that those bounds hold both when estimating asymptotic entropy of Markov chain, and when estimating entropy of any fixed length path taken from any starting distribution.

Our techniques

- we develop a lemma which measures *closeness of sample paths of two chains*; it non-trivially extends the classical result on the distance of two IID sequences and is of independent interest (the motiviation is Le Cam's method on Markov chains)
- we use *perturbation theory* to get insights into spectral properties of matrices; this technique greatly simplifies otherwise complicated calculations and is of independent interest.

2 Preliminaries

2.1 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 $p \times p$

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

Table 2: Lower Bounds for Markov Chain Entropy Estimation

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 $\|\cdot\|_p$.

2.2 Entropy Rates

2.2.1 Entropy $\mathbf{H}_\infty$

Min-entropy of a discrete random variable X is defined as $\mathbf{H}_\infty(X) = -\log\max_x \log\Pr[X = x]$. It is known that the min-entropy rate of a markov chain is determined by the average heaviest cycle [Kamath and Verdú, 2016]. 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

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

2.2.2 Entropy $\mathbf{H}_\alpha$

To evaluate the limiting Renyi entropy of order α, one considers the spectral properties of the *Hadamard power* of the transition matrix. Namely for a chain with a transition matrix M by [Rached et al., 2001] we have

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

2.3 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.

2.4 Perturbation Theory

The spectrum of a matrix remains (somewhat) stable under perturbations. There are many results of this form and we refer to [IPSEN, 2003] or [Zhan and Society, 2013] for more details and a survey; for our needs the classical result due to *Bauer-Firke* will be enough.

Lemma 2.1 (Bauer-Firke Eigenvalue Perturbation [Bauer and Fike, 1960])**.** *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-close to some eigenvalue of A, where $\delta = \|E\|_2$.*

Also the perturbations of eigenvectors have been studied. We will need to apply them to the *stochastic matrices*; in our case we will use bounds depending on a *hitting times*, due to Cho and Meyer.

Lemma 2.2 (Perturbation of MC stationary distributions [Cho et al., 2000])**.** *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 $\frac{m_{i,j}}{m_{j,j}} \leqslant 2\kappa$ for all i, j and $m_{i,j}$ is the expected time of hitting j when the chain starts from i.*

Note that for a uniform random walk over state space S hitting times equal $m_{i,j} = |S|$. Indeed the probability that the walk returns to the starting state after more than n steps equals $(1-|S|^{-1})^n$; thus the expectation equals $|S|$.

2.5 Coupling

Coupling refers to building joint distribution with given marginals and is a powerful technique used to study Markov chains [Frank, 2010]. The following lemma slightly extends the standard construction of coupling

Proposition 2.1 (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 $\Pr[X_1' \neq Y_1'] = d_{TV}(X_1, Y_1)$.*

2.6 Chernoff-Type Bounds for Markov Chains

Chernoff-type bounds hold also for Markov chains with exponentially small tails, but the constant depends on spectral properties of the transition matrix [Lezaud, 1998] or (which is related) on return times of the corresponding random walk [Chung et al., 2012]. In our case this translates to the sample complexity dependency also on the spectral gap.

3 Results

3.1 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 3.1 (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 between $n+1$ samples is bounded by*

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

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 3.1 (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 3.2 (Bounds for IID distributions). *Consider the following matrices $M_X = \left[\frac{1}{m-\ell}\mathbf{1}_{m,m-\ell} \quad \mathbf{0}_{m,\ell}\right]$ and $M_Y = \left[\mathbf{0}_{m,\ell} \quad \frac{1}{m-\ell}\mathbf{1}_{m,m-\ell}\right]$. 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 $E = \left[-\frac{1}{m-\ell}\mathbf{1}_{m,\ell} \quad \mathbf{0}_{m,m-2\ell} \quad \frac{1}{m-\ell}\mathbf{1}_{m,\ell}\right]$. Applying Lemma 3.1 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 3.1- 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.

Coupling. Let $X_0, \ldots, X_n$ and $Y_0, \ldots, Y_n$ be samples from Markov chains that have transition matrices M_X and M_Y respectively. For any coupling

$$\begin{aligned} d_{TV}(X_{\leqslant n}, Y_{\leqslant n}) &= \Pr[X_{\leqslant n-1} = Y_{\leqslant n-1}] \cdot d_{TV}(X_n; Y_n | X_{\leqslant n-1} = Y_{\leqslant n-1}) \quad (1) \\ &\quad + \Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}] \cdot d_{TV}(X_n; Y_n | X_{\leqslant n-1} \neq Y_{\leqslant n-1}) \\ &\leqslant d_{TV}(X_n; Y_n | X_{n-1} = Y_{n-1}) + \Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}] \quad (2) \end{aligned}$$

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

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

If X starts from the stationary distribution μ_X^T we have $X_n = \mu_X$ for all n. Therefore

$$d_{TV}(X_n; Y_n | X_{n-1} = Y_{n-1}) \leqslant \mu_X^T \cdot |M_X - M_Y| \cdot \mathbf{1} \tag{3}$$

There is a coupling such that

$$\Pr[X_{\leqslant n-1} \neq Y_{\leqslant n-1}] = d_{TV}(X_{\leqslant n-1}, Y_{\leqslant n-1}) \tag{4}$$

Putting Equation (3) and Equation (4) into Equation (2) we get

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

so that the statement follows by induction. □

Dynamic Programming. Consider the variation distance of $n+1$ samples

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

Writing $\mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i}$ as the difference of $\mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i} \cdot \left(M_{s_{n-1},s_n} + E\right)$ and $\mu^M_{s_0} \prod_{i=1}^n M_{s_{i-1},s_i} \cdot E$ and and applying the triangle inequality we get $d_{TV} \leqslant I_1 + I_2$ where

$$I_1 = \sum_{s_0,\ldots,s_{n-1}} \left| \mu^M_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} - \mu^{M+E}_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} \right| \cdot \|M+E\|_\infty$$

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. Therefore

$$I_1 \leqslant \sum_{s_0,\ldots,s_{n-1}} \left| \mu^M_{s_0} \prod_{i=1}^{n-1} M_{s_{i-1},s_i} - \mu^{M+E}_{s_0} \prod_{i=1}^{n-1} (M+E)_{s_{i-1},s_i} \right| = d_{\mathrm{TV}}^{n-1} \tag{5}$$

If μ_M is stationary for M then by Chappman-Klomogorov

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

Summing up we get

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

which by induction implies the statement. □

3.2 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 uniform random walk
- perturbation of uniform random walk which overweights one element, the transition matrix of this chain is defined below

$$M = \begin{bmatrix} \frac{1}{m}\mathbf{1}_{m-1,m-1} & \frac{1}{m}\mathbf{1}_{m-1,1} \\ \left(\frac{1}{m} - \frac{\epsilon}{m-1}\right)\mathbf{1}_{1,m-1} & \left(\frac{1}{m} + \epsilon\right) \end{bmatrix} \tag{6}$$

Because the perturbation is sparse, the change in the distance of finite samples will be small. On the other hand we will see that it has a significant effect on the spectrum of Hadamard powers.

3.3 Mixing Time is Good

[Kamath and Verdú, 2016] 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 well-known properties of matrix of ones [Horn and Johnson, 2013]. It follows that the spectral gap is constant. After the perturbation we maintain the constant spectral gap, which follows again by perturbation theory (Lemma 2.1). Note that elementary row operations change eigenvalues, so calculating explicitly the spectrum is hard (although doable for this specific case)!

3.4 Entropy Rates

We state our results for *entropy rates* which, for stochastic sources such as Markov chains, are understood as the limiting entropy per symbol (for Markov chains they exist under standard assumptions such as ergodicity).

3.4.1 Rate Evaluation for $\mathbf{H}_\infty$

We need to find the change in the entropy rate and statistical distance when changing from ϵ to $\epsilon = 0$ in Equation (6).

Claim 3.1 (Min-Entopy Rate)**.** *For the chain with transition matrix as in* (6)

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

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

Claim 3.2 (Statistical Distance Closeness)**.** *The variational distance between n samples from M in Equation (6) 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 3.1 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

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

note that the sparsity of E helps! The distance between stationary distributions can be bounded by $O(\epsilon)$ according to Lemma 2.2. □

Corollary 3.1 (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 3.2 (Statistical Distance)**.** *Let $\epsilon = \frac{1}{m}$, by Claim 3.2 the distance between n samples is bounded by $O(n \cdot m^{-2})$.*

By the two above corollaries and the Le Cam's method described in Section 2.3 we get our lower bound for min-entropy.

3.4.2 Rate Evaluation for $\mathbf{H}_2$

In the lemmas below we estimate the difference in entropy and closeness in statistical distance for these two chains. The results are given in Corollary 3.4 and Corollary 3.3 below.

Lemma 3.2 (Spectral Radious Gap)**.** *Consider the matrix in Equation (6). The spectral radius of its second Hadamard power satisfies*

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

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

Corollary 3.4 (Statistical Distance)**.** *By Claim 3.2, 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 2.3 to above corollaries we get our lower bound for collision entropy.

Proof of Lemma 3.2. We have

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

We want to compute the spectral radious of $M^{\diamond 2}$. We can 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 $O(m^{-2})$. 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 2.1 (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. $\square$

3.4.3 Rate Evaluation for $\mathbf{H}_\alpha$

By proceeding in the same way as for $\mathbf{H}_2$ we arrive at $\rho(M^{\diamond\alpha}) = \rho(Z) + O(m^{-\frac{2\alpha-1}{2}})$ where Z has same structure but the power of 2 is replaced by α. This gives us

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

Let $\epsilon = (2/m)^{\frac{\alpha-1}{\alpha}}$ then we get $\rho(M^{\diamond\alpha}) = (2/m)^{\alpha-1}(1+o(1))$ for large m. This gives a constant entropy gap and the statistical distance of $O(n \cdot m^{-2+\frac{1}{\alpha}})$ between the two paths studied in Le Cam's method.

Note: when deriving formulas above we assumed large m in $o(1)$ terms, but in fact we have lower bounds of form $\rho(M^{\diamond\alpha}) \geqslant (2/m)^{\alpha-1}$ which is sufficient.

3.5 Upper Bounds

We can apply Chernoff-type bounds to get frequencies up to a multiplicative error term. This will cost $|S|^2\mathrm{polylog}(|S|, 1/\epsilon)$ samples. We defer the easy proof to the final version.

4 Finite Sample Bounds

Our bounds were derived for the problem of estimating *asymptotic entropy rate*, but they remain valid also for the task of estimating entropy of *finite* number of samples. We will give the argument for min-entropy, the Renyi entropy case will be discussed in the full version.

For the min-entropy this follows because the entropy of n samples for both matrices considered equals n times the entropy rate. Indeed, the min-entropy of n samples generated from the chain with the transition matrix as in Equation (6) is full when $\epsilon = 0$ and for the case $\epsilon > 0$ achieved for n repetitions of the m-th symbol.

5 Conclusions

We have shown lower bounds for Renyi entropy rate estimation under the Markov chain model.

References

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rényi entropy. In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1855–1869, Philadelphia, PA, USA, 2015. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=2722129.2722253.

András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms*, 19(3-4):163–193, October 2001. ISSN 1042-9832. doi: 10.1002/rsa.10019. URL http://dx.doi.org/10.1002/rsa.10019.

Erdal Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Information Theory*, 42(1):99–105, 1996. doi: 10.1109/18.481781. URL http://dx.doi.org/10.1109/18.481781.

Boaz Barak, Yevgeniy Dodis, Hugo Krawczyk, Olivier Pereira, Krzysztof Pietrzak, François-Xavier Standaert, and Yu Yu. Leftover hash lemma, revisited. In *Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings*, pages 1–20, 2011. doi: 10.1007/978-3-642-22792-9_1. URL http://dx.doi.org/10.1007/978-3-642-22792-9_1.

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. doi: 10.1145/2432622.2432626. URL http://doi.acm.org/10.1145/2432622.2432626.

F. L. Bauer and C. T. Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, Dec 1960. ISSN 0945-3245. doi: 10.1007/BF01386217. URL https://doi.org/10.1007/BF01386217.

Charles H. Bennett, Gilles Brassard, Claude Crépeau, and Ueli M. Maurer. Generalized privacy amplification. *IEEE Trans. Information Theory*, 41(6):1915–1923, 1995. doi: 10.1109/18.476316. URL http://dx.doi.org/10.1109/18.476316.

Haixiao Cai, S. R. Kulkarni, and S. Verdu. Universal entropy estimation via block sorting. *IEEE Trans. Inf. Theor.*, 50(7):1551–1561, September 2006. ISSN 0018-9448. doi: 10.1109/TIT.2004.830771. URL http://dx.doi.org/10.1109/TIT.2004.830771.

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.

Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. In *STACS*, volume 14 of *LIPIcs*, pages 124–135. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2012.

Yevgeniy Dodis and Yu Yu. Overcoming weak expectations. In *Theory of Cryptography - 10th Theory of Cryptography Conference, TCC 2013, Tokyo, Japan, March 3-6, 2013. Proceedings*, pages 1–22, 2013. doi: 10.1007/978-3-642-36594-2_1. URL http://dx.doi.org/10.1007/978-3-642-36594-2_1.

Michelle Effros. Universal lossless source coding with the burrows wheeler transform. In *Proceedings of the Conference on Data Compression*, DCC '99, pages 178–, Washington, DC, USA, 1999. IEEE Computer Society. ISBN 0-7695-0096-X. URL http://dl.acm.org/citation.cfm?id=789086.789615.

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

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 - In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, pages 68–75. 2011. doi: 10.1007/978-3-642-22670-0_9. URL http://dx.doi.org/10.1007/978-3-642-22670-0_9.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61, 11 2014. doi: 10.1109/TIT.2015.2478816.

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

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.

Manjesh Kumar Hanawal and Rajesh Sundaresan. Guessing revisited: A large deviations approach. *IEEE Trans. Information Theory*, 57(1):70–78, 2011. doi: 10.1109/TIT.2010.2090221. URL http://dx.doi.org/10.1109/TIT.2010.2090221.

R.A. Horn and C.R. Johnson. *Matrix Analysis*. Matrix Analysis. Cambridge University Press, 2013. ISBN 9780521839402. URL https://books.google.at/books?id=5I5AYeeh0JUC.

Russell Impagliazzo and David Zuckerman. How to recycle random bits. In *30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, North Carolina, USA, 30 October - 1 November 1989*, pages 248–253, 1989. doi: 10.1109/SFCS.1989.63486. URL http://dx.doi.org/10.1109/SFCS.1989.63486.

ILSE C.F. IPSEN. Departure from normality and eigenvalue perturbation bounds, 2003. URL https://pdfs.semanticscholar.org/dd2a/86188a5d336a9f5c35e1acf6b4583eb1d33e.pdf.

R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft. Clustering using renyi's entropy. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 1, pages 523–528 vol.1, 2003. doi: 10.1109/IJCNN.2003.1223401.

Petr Jizba, Hagen Kleinert, and Mohammad Shefaat. Rényi's information transfer between financial time series. *Physica A: Statistical Mechanics and its Applications*, 391(10):2971 – 2989, 2012. ISSN 0378-4371. doi: http://dx.doi.org/10.1016/j.physa.2011.12.064. URL http://www.sciencedirect.com/science/article/pii/S0378437112000131.

S. Kamath and S. Verdú. Estimation of entropy rate and rényi entropy rate for markov chains. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 685–689, July 2016. doi: 10.1109/ISIT.2016.7541386.

Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998. ISBN 0-201-89685-0.

Coco Krumme, Alejandro Llorente, Manuel Cebrian, Alex Pentland, and Esteban Moro. The predictability of consumer visitation patterns. *Scientific reports*, 3:1645, 04 2013. doi: 10.1038/srep01645.

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

Pascal Lezaud. Chernoff-type bound for finite markov chains. *Ann. Appl. Probab.*, 8(3):849–867, 08 1998. doi: 10.1214/aoap/1028903453. URL https://doi.org/10.1214/aoap/1028903453.

Ke Li, Wanlei Zhou, Shui Yu, and Bo Dai. Effective ddos attacks detection using generalized entropy metric. In *Algorithms and Architectures for Parallel Processing, 9th International Conference, ICA3PP 2009, Taipei, Taiwan, June 8-11, 2009. Proceedings*, pages 266–280, 2009. doi: 10.1007/978-3-642-03095-6_27. URL http://dx.doi.org/10.1007/978-3-642-03095-6_27.

Bing Ma, Alfred O. Hero III, John D. Gorman, and Olivier J. J. Michel. Image registration with minimum spanning tree algorithm. In *Proceedings of the 2000 International Conference on Image Processing, ICIP 2000, Vancouver, BC, Canada, September 10-13, 2000*, pages 481–484, 2000. doi: 10.1109/ICIP.2000.901000. URL http://dx.doi.org/10.1109/ICIP.2000.901000.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 367–374, 2009. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1600&proceeding_id=25.

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. doi: 10.1109/TIT.2013.2270273. URL http://dx.doi.org/10.1109/TIT.2013.2270273.

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. doi: 10.1002/ima.20079. URL http://dx.doi.org/10.1002/ima.20079.

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

Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, June 2003. ISSN 0899-7667. doi: 10.1162/089976603321780272. URL http://dx.doi.org/10.1162/089976603321780272.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Information Theory*, 54(10):4750–4755, 2008. doi: 10.1109/TIT.2008.928987. URL http://dx.doi.org/10.1109/TIT.2008.928987.

C. E. Pfister and W. G. Sullivan. Rényi entropy, guesswork moments, and large deviations. *IEEE Trans. Information Theory*, 50(11):2794–2800, 2004. doi: 10.1109/TIT.2004.836665. URL http://dx.doi.org/10.1109/TIT.2004.836665.

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. doi: 10.1109/18.923736. URL https://doi.org/10.1109/18.923736.

A. Rényi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960. URL http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf.

Prasanna K. Sahoo and Gurdial Arora. A thresholding method based on two-dimensional renyi's entropy. *Pattern Recognition*, 37(6):1149–1161, 2004. doi: 10.1016/j.patcog.2003.10.008. URL http://dx.doi.org/10.1016/j.patcog.2003.10.008.

C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001. ISSN 1559-1662. doi: 10.1145/584091.584093. URL http://doi.acm.org/10.1145/584091.584093.

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. doi: 10.1126/science.1177170.

Taro Takaguchi, Mitsuhiro Nakamura, Nobuo Sato, Kazuo Yano, and Naoki Masuda. Predictability of conversation partners. *Computing Research Repository - CORR*, 1, 04 2011. doi: 10.1103/PhysRevX.1.011008.

Gregory Valiant and Paul Valiant. Estimating the unseen: An n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0691-1. doi: 10.1145/1993636.1993727. URL http://doi.acm.org/10.1145/1993636.1993727.

Paul C. van Oorschot and Michael J. Wiener. Parallel collision search with cryptanalytic applications. *J. Cryptology*, 12(1):1–28, 1999. doi: 10.1007/PL00003816. URL http://dx.doi.org/10.1007/PL00003816.

Chunyan Wang and Bernardo Huberman. How random are online social interactions? *Scientific reports*, 2:633, 09 2012. doi: 10.1038/srep00633.

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. doi: 10.1109/TIT.2016.2548468.

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. ISSN 0018-9448. doi: 10.1109/18.45281.

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

X. Zhan and American Mathematical Society. *Matrix Theory*. Graduate Studies in Mathematics. American Mathematical Society, 2013. ISBN 9781470409456. URL https://books.google.at/books?id=VBgLkAEACAAJ.