Entropy Rate Estimation for Markov Chains with Continuous Distributions

Puning Zhao
Zhejiang Lab
pnzhao@zhejianglab.com

Lifeng Lai
Department of Electrical and Computer Engineering
University of California, Davis
lflai@ucdavis.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206685

***Abstract*—Entropy rate estimation has a broad range of applications such as bioinformatics, feature clustering etc. Although there are many existing work on the estimation of entropy rate for Markov chains with discrete distributions, the understanding for the entropy rate estimation of Markov chains with continuous distributions is limited. In this paper, efficient methods for estimating the entropy rate for Markov chains with continuous distributions are proposed. Moreover, we derive bounds on the convergence rate of the proposed entropy rate estimators.**

I. Introduction

Entropy rate is an important quantity in information theory and statistics. It can be understood as the fundamental limit of predicting the next step in a stochastic process. Correspondingly, the estimation of entropy rate for a stochastic process, especially a stationary Markov chain, has been used in a wide variety applications. For example, it can be used in bioinformatics [1–5] for analyzing DNA sequences, feature clustering and image registration [6], blind source separation [7, 8], economics [9], and many other signal processing related fields [10, 11].

The estimation of the entropy rate for Markov chains with discrete distributions has been discussed in some recent interesting works [12–15]. A simple and intuitive method is plug-in method, in which one estimates the stationary distribution and the transition probability matrix first and then calculates the entropy rate. [13] proved that this estimator converges almost surely and is asymptotically normal. [14] provided a finite sample bound on the estimation error. These results show that the simple plug-in method is efficient if the alphabet size is finite and fixed. If the alphabet size grows with the sequence length, then the simple plug-in method is no longer optimal. In [15], a new method was proposed, which can handle the case with large state space. This method estimates the conditional entropy given each previous state based on some efficient entropy estimators for identical and independently distributed (i.i.d) samples, such as those proposed in [16, 17]. It exhibits clear advantage over the simple plug-in estimator, since it estimates the conditional entropy directly, instead of estimating the full transition matrix. Such an improvement is significant if the alphabet size is large. Moreover, it is shown in [15] that this new method achieves the minimax optimal sample complexity.

Despite that the entropy rate estimation has been widely discussed for discrete distributions, the previous methods and the theoretical analysis can not be straightforwardly generalized to continuous distributions. For discrete distributions, for all samples at a fixed state s, their next states are i.i.d conditional on s. However, for continuous distributions, we can not use this property, because we can not expect that there are a large number of samples located at the same state. As a result, it is impossible to find some states that are conditionally i.i.d, which makes the analysis much harder.

In this paper, our goal is to estimate the entropy rate for continuous distributions. In particular, we propose two competitive methods to estimate the entropy rate of Markov Chains with continuous distributions.

For the first method, our design is based on the fact that for stationary and homogeneous Markov chain, the entropy rate equals to the conditional entropy of a state given its previous state. Therefore, it is natural to design a method based on the combination of two Kozachenko-Leonenko (KL) entropy estimators [18], in which one of them estimates the joint entropy, while the other estimates the marginal entropy. The final estimate of the entropy rate can then be calculated by subtracting the marginal entropy estimate from the joint entropy estimate. We name this method as 2KL entropy rate estimator. The 2KL method is simple to use with little parameter tuning, and has good empirical performance. However, it is very difficult to rigorously characterize the convergence rate of this 2KL method. The analysis techniques for the KL entropy estimator [19–24] cannot be applied for the analysis of the proposed 2KL entropy rate estimator. The main reason is that the existing techniques for analyzing the KL entropy estimator relies heavily on the fact that the available samples are independent, while the samples obtained from the Markov chain case are not independent anymore.

To overcome the lack of rigorous convergence rate characterization of the 2KL entropy rate estimator, we propose another estimator that is amenable to analysis and has similar or better performance than the 2KL estimator. The main idea of the new estimator is to divide the support set into bins. For each bin, we find all samples falling in

This work was supported by the National Science Foundation under grant CCF-21-12504.

this bin and then find their next states. After that, we apply the KL entropy estimator, on these states whose previous steps are all in the same bin. We call this as Bin-KL entropy rate estimator. The Bin-KL entropy rate estimator shares some similarity with that in [15], since both [15] and our method estimates the conditional entropy directly instead of estimating the full probability mass function (pmf) or probability density function (pdf). However, unlike discrete distributions, for continuous distributions, the states whose previous steps are within the same bin are not i.i.d or conditional i.i.d given the previous state, since their distributions are still slightly different due to different locations of the previous states, even if those previous states are in the same bin. As a result, the analysis of the new proposed method becomes harder. For this estimator, we are able to provide a new analysis to show that this method is consistent, and derive its convergence rate. Our analysis uses some techniques from previous works on the KL entropy estimator [21, 23, 24]. Consider that the samples are no longer i.i.d, we modify the previous analysis and carefully bounded the effect of the mutual dependence between each steps. To the best of our knowledge, this is the first attempt to propose a method to estimate the entropy rate of Markov chains with continuous distributions, and bound its convergence rate.

II. Preliminaries

Consider a first order ergodic Markov chain $\mathbf{X}_1, \mathbf{X}_2, \ldots$, in which $\mathbf{X}_i \in S$, $S \subset \mathbb{R}^d$ is a compact set. Each $\mathbf{X}_i$ is a continuous random variable, is conditionally independent with $\mathbf{X}_1, \ldots, \mathbf{X}_{i-2}$ given $\mathbf{X}_{i-1}$. The entropy rate is defined as

$$\bar{h} = \lim_{n\to\infty} \frac{1}{n} h(\mathbf{X}_1, \ldots, \mathbf{X}_n), \tag{1}$$

in which $h(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the joint differential entropy of $\mathbf{X}_1, \ldots, \mathbf{X}_n$:

$$\begin{aligned} & h(\mathbf{X}_1, \ldots, \mathbf{X}_n) \\ = & -\int f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \ln f(\mathbf{x}_1, \ldots, \mathbf{x}_n) d\mathbf{x}_1 \ldots d\mathbf{x}_n. \end{aligned}$$

The joint pdf $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is usually unknown in practice. We need to estimate the entropy rate from samples. Suppose we are given a realization of the Markov chain, which is a sequence with length N, i.e. $\mathbf{X}_1, \ldots, \mathbf{X}_N$, in which $\mathbf{X}_1$ is an arbitrary initial state. Our goal is to estimate the entropy rate $\bar{h}$ based on these N samples.

In this paper, we assume that the Markov chain is time homogeneous, which means that there exists a transition kernel p, such that $p(\mathbf{x}, \cdot) = f_{i+1}(\cdot|\mathbf{X}_i = \mathbf{x})$ for all $i = 1, 2, \ldots$ and $\mathbf{x} \in S$, in which $f_{i+1}(\cdot|\mathbf{X}_i = \mathbf{x})$ is the conditional pdf of $\mathbf{X}_{i+1}$ given $\mathbf{X}_i = \mathbf{x}$. Denote π as the pdf of the stationary distribution, which satisfies $\int \pi(\mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} = \pi(\mathbf{y})$.

Then we have

$$\begin{aligned} \bar{h} &= \lim_{n\to\infty} \frac{1}{n} h(\mathbf{X}_1, \ldots, \mathbf{X}_n) \\ &= \lim_{n\to\infty} h(\mathbf{X}_n|\mathbf{X}_{n-1}) = \int \pi(\mathbf{x}) h(p(\mathbf{x}, \cdot)) d\mathbf{x}, \end{aligned} \tag{2}$$

in which $h(p(\mathbf{x}, \cdot)) = -\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$. is the conditional entropy of a state given the previous state.

III. Proposed Methods

In this section, we propose two methods, called 2KL method and Bin-KL method respectively, to estimate the entropy rate based on the expression (2).

A. 2KL Method

For a time homogeneous and uniformly ergodic Markov chain, we have

$$\bar{h} = \lim_{n\to\infty} \frac{1}{n} h(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \lim_{n\to\infty} h(\mathbf{X}_n|\mathbf{X}_{n-1}). \tag{3}$$

Therefore, we can estimate the entropy rate by estimating the conditional entropy of next state given the previous state. Define $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{X}_{i+1})$ for $i = 1, \ldots, N-1$, then $\bar{h} = \lim_{n\to\infty} [h(\mathbf{Z}_n) - h(\mathbf{X}_n)]$. $h(\mathbf{Z}_n)$ and $h(\mathbf{X}_n)$ slightly change over n. However, they will converge as n increases. Hence, a possible idea is to use two entropy estimators to estimate $h(\mathbf{Z})$ using $\mathbf{Z}_i$, $i = 1, 2, \ldots, N-1$ and to estimate $h(\mathbf{X})$ using $\mathbf{X}_i$, $i = 1, 2, \ldots, N$ separately, and then calculate the estimated conditional entropy.

The most common method for estimating the entropy for continuous random variable is the KL estimator, which calculates the differential entropy based on k nearest neighbor distances. If the nearest neighbor distances are large, then the random variable has a high differential entropy, and vice versa. It was first proposed in [18], and was then analyzed in [20, 21, 23–26]. In our case, the expressions of the KL estimates for $h(\mathbf{X})$ and $h(\mathbf{Z})$ are

$$\hat{h}(\mathbf{X}) = -\psi(k) + \psi(N) + \ln c_d + \frac{d}{N} \sum_{i=1}^{N} \ln \epsilon_{\mathbf{X}_i}, \tag{4}$$

$$\hat{h}(\mathbf{Z}) = -\psi(k) + \psi(N-1) + \ln c_{2d} + \frac{d}{N} \sum_{i=1}^{N} \ln \epsilon_{\mathbf{Z}_i}, \tag{5}$$

in which ψ is the digamma function, $\psi(t) = \Gamma'(t)/\Gamma(t)$, Γ is the Gamma function, $\Gamma(t) = \int u^{t-1} e^{-u} du$. c_d and c_{2d} are the volumes of the d-dimension and $2d$-dimension unit balls respectively. If we use ℓ_2 metric, then $c_d = \pi^{d/2}/\Gamma(d/2+1)$. c_{2d} can be defined similarly. $\epsilon_{\mathbf{X}_i}$ is the distance of $\mathbf{X}_i$ to its k nearest neighbors among $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, and $\epsilon_{\mathbf{Z}_i}$ is the distance of $\mathbf{Z}_i$ to its k nearest neighbors among $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\}$. We call this method as 2KL method, as it is a combination of two KL estimators.

This method is simple to use since the only parameter we need to tune is k. According to the analysis of KL estimators in [23, 24, 27], the optimal k is fixed with respect to sample size N. Furthermore, we will show in Section V that the 2KL method has good empirical performance. However, it is challenging to rigorously characterize the convergence property of this method. In

particular, although there are many existing works that analyze the performance of KL entropy estimators for the case with i.i.d data under various assumptions [19–24], these analysis can not be easily generalized to analyze the 2KL entropy rate estimator for the Markov chains because samples are no longer i.i.d. The challenge with the rigorous performance analysis of the 2KL method motivates us to design an alternative entropy rate estimator that is amenable to analysis and has similar or better empirical performance as the 2KL method in the next subsection.

B. Bin-KL Method

To address the lack of rigorous performance analysis of the 2KL method issue, we propose an alternative entropy rate estimator that has rigorous convergence characterization and has similar or better empirical performance. We name this new method as Bin-KL method. The main idea of the Bin-KL method is to estimate $\pi(\mathbf{x})$ and $h(p(\mathbf{x},\cdot))$. However, since the distribution of each $\mathbf{X}_i$ is continuous, we can not estimate them at every $\mathbf{x}$. Therefore, we divide the support S into m bins, denoted as $B_1,\ldots,B_m$, such that each bin is a connected set and is sufficiently small. Assume that the transition kernel p and the corresponding pdf of the stationary distribution π is continuous, then they will be sufficiently close within each bin. Hence, we can let the estimation of π and $\hat{h}(p(\mathbf{x},\cdot))$ to be the same in each bin.

$\pi(\mathbf{x})$ can then be estimated by $\hat{\pi}(\mathbf{x}) = n(B_j)/NV(B_j)$, in which $V(B_j)$ is the volume of B_j, and $n(B_j) = \sum_{i=1}^{N-1}\mathbf{1}(\mathbf{X}_i \in B_j)$ is the number of samples falling in B_j among the first $N-1$ samples. This means that $\pi(\mathbf{x})$ can be simply estimated by the fraction of all samples falling in B_j.

For the estimation of $h(p(\mathbf{x},\cdot))$ for $\mathbf{x} \in B_j$, we denote $I_j = \{i|\mathbf{X}_{i-1} \in B_j\}$, which is a set of indices of samples whose previous step belong to B_j. Obviously, the cardinality of I_j is $|I_j| = n(B_j)$. Denote $(\mathbf{X}_{j1},\mathbf{Y}_{j1}),\ldots,(\mathbf{X}_{j,n(B_j)},\mathbf{Y}_{j,n(B_j)})$ as a random permutation of $(\mathbf{X}_{i-1},\mathbf{X}_i)$ for $i \in I_j$. Note that the distributions of $\mathbf{Y}_{j1},\ldots,\mathbf{Y}_{j,n(B_j)}$ are close to each other, since their previous steps are all in B_j, and the previous steps are close to each other. With this observation, we can use the KL entropy estimator to estimate $h(p(\mathbf{x},\cdot))$ for $\mathbf{x} \in B_j$. The estimated value of $h(p(\mathbf{x},\cdot))$ is the same for all $\mathbf{x} \in B_j$. Therefore, we use $\hat{h}_j$ to denote such estimated result. If $n(B_j) \geq k$, then

$$\begin{aligned}\hat{h}_j &= -\psi(k) + \psi(n(B_j)) + \ln c_d \\ &\quad + \frac{d}{n(B_j)}\sum_{l=1}^{n(B_j)} \ln \epsilon_{jl}, \end{aligned} \tag{6}$$

in which ϵ_{jl} is the distance from $\mathbf{Y}_{jl}$ to its k-th nearest neighbor among $\mathbf{Y}_{j1},\ldots,\mathbf{Y}_{j,l-1}$, $\mathbf{Y}_{j,l+1},\ldots,\mathbf{Y}_{j,n(B_j)}$. If $n(B_j) < k$, then the KL entropy estimator can not be used. In this case, we just set $\hat{h}_j = 0$.

Setting $\hat{h}_j = 0$ when the number of samples within B_j is less than k will inevitably cause some estimation bias. However, we can ensure that the number of bins grows slower than the total sample size N, then the expected number of samples within each bin will also grows with N. As a result, the probability that $n(B_j) < k$ becomes smaller with the increase of N. This ensures that the additional bias caused by setting $\hat{h}_j = 0$ converges to zero.

Combining these two steps, the proposed Bin-KL entropy rate estimator is written as

$$\begin{aligned}\hat{\bar{h}} &= \int \hat{\pi}(\mathbf{x})\hat{h}(p(\mathbf{x},\cdot))d\mathbf{x} \\ &= \sum_{j=1}^{m}\int_{B_j}\hat{\pi}(\mathbf{x})\hat{h}(p(\mathbf{x},\cdot))d\mathbf{x} = \sum_{j=1}^{m}\frac{n(B_j)}{N}\hat{h}_j. \end{aligned} \tag{7}$$

The Bin-KL method has two design parameters, m and k. The optimal choice of m grows with sample size N. We will characterize the optimal growth rate of m with sample size N in the convergence analysis section. On the contrary, the optimal k does not grow with N. Therefore, k can be selected as a fixed value. Although the selection of k can partially affect the bias and variance of this estimator, it does not impact their convergence rates.

We now compare the Bin-KL method and the 2KL method. As will be shown in the numerical simulation, both methods have good empirical performance. The mean square errors of both methods converge to zero. The performances of these two methods depend on the distributions, but they generally have comparable empirical performances. A major difference is that we have rigorous convergence analysis for the Bin-KL method (shown in Section IV), while for the 2KL method, we are not able to provide such analysis due to difficulties discussed in Section III-A.

Another aspect to compare is the time complexity. We discover that the Bin-KL method usually requires less time than the 2KL method. This is because the k nearest neighbor search has a higher time complexity than assigning bins. For the Bin-KL method, the KL estimator is used for every bin, in which the number of samples is much less than the total sample size N. However, for the 2KL method, the KL estimator is used on the whole dataset. Hence, the 2KL method is slower than the Bin-KL method, especially when the sequence length is large.

Finally, we would like to remark that the 2KL method has a broader range of applications. It can be used for both distributions with bounded and unbounded support, while the Bin-KL method can only be used on distributions with bounded support. If the support is unbounded, the number of bins will be infinite. To address this, it is possible to improve the bin method so that it can adaptively divide the support into bins with different sizes, such that the bin size is larger where the pdf of the stationary distribution is low.

IV. Convergence Analysis

In this section, we provide a theoretical analysis of the convergence rate of the Bin-KL entropy rate estimator (7). Our analysis is based on the following assumptions.

Assumption 1. *We make the following assumptions:*

(a) The support set S has finite volume V_S, finite surface area A_S and finite diameter D;

(b) The conditional pdf is lower bounded, i.e. $p(\mathbf{x},\mathbf{y}) \geq f_L$ for some constant $f_L > 0$.

(c) $p(\mathbf{x},\mathbf{y})$ is L-Lipschitz;

(d) There exist two constant R and α, such that for all $\mathbf{x} \in S$ and $r \leq R$, we have

$$V(B(\mathbf{x},r) \cap S) \geq \alpha V(B(\mathbf{x},r)), \tag{8}$$

in which V_S, A_S, D, C_b, f_L, f_U, L, R are all finite positive constants, and $\alpha \in (0,1)$.

We now comment on these assumptions. Assumption (a) is an assumption that is necessary for our bin splitting method. It is possible to design an adaptive bin splitting strategy to the estimate the entropy rate if the distribution of $\mathbf{X}_i$ has an unbounded support. However, in this paper, we focus on the case that S is bounded for simplicity. Assumption (b) restricts the lower bound of the transition probability. It can also be shown from Assumption (a) and (c) that the $p(\mathbf{x},\mathbf{y})$ is also upper bounded. These bounds are important to calculate the convergence rate. Such assumption has already been made in similar works on the estimation of information theoretic functionals [23, 24, 27]. In Assumption (c), we assume that $p(\mathbf{x},\mathbf{y})$ is Lipschitz in both $\mathbf{x}$ and $\mathbf{y}$. The overall convergence of the mean square error may be faster if we assumes smoothness of $p(\mathbf{x},\mathbf{y})$ with a higher order. Assumption (d) is a regularity assumption on the shape of the support, which is satisfied by almost all common support sets. For example, Assumption (d) holds if the support set is convex or is the union of a finite number of convex sets. We would like to remark that in previous works on the estimation of entropy rate for discrete distributions [14, 15], there are some assumptions about the uniform ergodicity of Markov chain as well as the corresponding mixing time, which indicates how fast the distribution converges to the stationary distribution. This assumption is necessary to get the convergence rate of the entropy rate estimator in [14, 15]. In our Assumptions (a)-(d), we do not state such assumption explicitly, because the uniform ergodicity can actually be derived from Assumption (b), which restricts the lower and upper bound of the transition kernel.

Based on these assumptions, we have the following theorem regarding the convergence rate of the Bin-KL entropy rate estimator defined in (7).

Theorem 1. *The mean square error of the Bin-KL entropy rate estimator can be bounded by*

$$\mathbb{E}[(\hat{\bar{h}} - \bar{h})^2] \lesssim \frac{m}{N}\ln^2 N + m^{-\frac{2}{d}} + \left(\frac{m}{N}\right)^{\frac{2}{d}}. \tag{9}$$

To optimize the convergence rate, we let m grow with N as

$$m \sim \begin{cases} N^{\frac{d}{d+2}} & \text{if} \quad d \leq 2 \\ N^{\frac{1}{2}} & \text{if} \quad d > 2. \end{cases} \tag{10}$$

With this choice, the convergence rate of the mean square error of the Bin-KL estimator becomes

$$\mathbb{E}[(\hat{\bar{h}} - \bar{h})^2] \lesssim \begin{cases} N^{-\frac{2}{d+2}}\ln^2 N & \text{if} \quad d \leq 2 \\ N^{-\frac{1}{d}} & \text{if} \quad d > 2. \end{cases} \tag{11}$$

The main idea of the proof of Theorem 1 is to bound the estimation error of $\hat{\pi}$ and $\hat{h}_j$ separately. The main difficulty is that $\hat{h}_j$ are not i.i.d for different j, thus the overall bound of the estimator can not be obtained by simply bounding the bias and variance of each $\hat{h}_j$. To cope with this problem, we designed a new approach that is different from the traditional analysis on KL estimator [23, 24].

Theorem 1 shows the convergence rate of the Bin-KL entropy rate estimator. An intuitive understanding of (9) is that the first term comes from the variance of $\hat{h}_j$ defined in (6), the second term comes from the bias of $\hat{h}_j$, while the third term comes from the bias and variance of $\hat{\pi}_j$.

V. Numerical Examples

In this section, we provide numerical simulations to validate our theoretical analysis. We use the following distributions as the ground truth. For all of the distributions used in this section, $\mathbf{X}_1 \sim \mathbb{U}([0,1]^d)$, in which $\mathbb{U}$ denotes uniform distribution, d is the dimensionality. In this section, we use $d = 1, 2, 3$. For each $i = 2, 3, \ldots$,

$$\mathbf{X}'_{i-1} = \mathbf{X}_{i-1} + \mathbf{W}_{i-1}, \tag{12}$$

$$\mathbf{X}_{ij} = \mathbf{X}'_{i-1,j} - \lfloor \mathbf{X}'_{i-1,j} \rfloor, \tag{13}$$

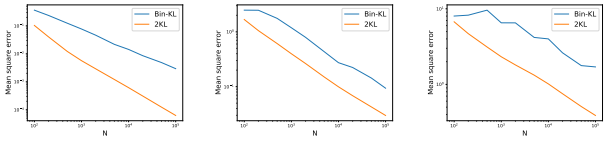
in which $\lfloor \cdot \rfloor$ is the floor function. By operation (13), it is ensured that all $\mathbf{X}_i$'s are within $[0,1]^d$. The distribution of $\mathbf{W}$ is different for different cases. In the first case, $\mathbf{W} \sim \mathbb{U}([-0.1, 0.1]^d)$. In the second case, $\mathbf{W} \sim \mathbb{U}([-0.3, 0.3]^d)$. In the third case, each component $\mathbf{W}_j$ follows a distribution with pdf $f_{Z_j}(z_j) = 1.5 - z_j$, for $z_j \in [0,1]$ and $j = 1, \ldots, d$. In the fourth case, each component $\mathbf{W}_j$ follows a Gaussian distribution with $\mu = 0, \sigma = 0.2$.

From the above construction, $\mathbf{X}_1, \ldots, \mathbf{X}_N$ is a Markov chain, in which each random variable is supported on $[-1,1]^d$. The entropy rates of these four distributions are $\bar{h}_1 = \ln(0.2)d$, $\bar{h}_2 = \ln(0.6)d$, $\bar{h}_3 = \left(\frac{1}{2} - \frac{1}{8}\ln 2 - \frac{9}{8}\ln\frac{3}{2}\right)d$, $\bar{h}_4 = -0.2249d$, in which $\bar{h}_4$ is calculated numerically.

In all of the trials, we fix $k = 3$ for both Bin-KL and 2KL methods. Moreover, for the Bin-KL method, we use

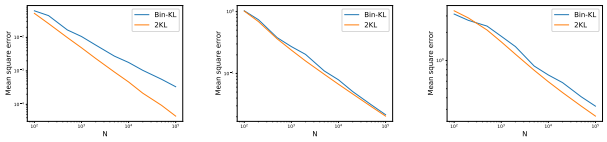
$$m = \begin{cases} \lfloor N^{\frac{1}{3}} \rfloor & \text{if} \quad d = 1 \\ \max\{m' | (m')^{\frac{1}{d}} \in \mathbb{N}, m' \leq \sqrt{N}\} & \text{if} \quad d \geq 2. \end{cases} \tag{14}$$

Such setting is based on (10), which requires $m \sim N^{1/3}$ for $d = 1$ and otherwise $m \sim \sqrt{N}$. The reason that we



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$

Fig. 1. Plots of the mean square error of the Bin-KL method and the 2KL method for the first distribution.



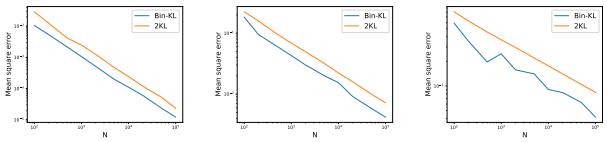
(a) $d = 1$ (b) $d = 2$ (c) $d = 3$

Fig. 2. Plots of the mean square error of the Bin-KL method and the 2KL method for the second distribution.

require $m^{1/d}$ to be an integer is that we would like the number of bins to be the same in each dimension, so that each bin is a regular hexahedron.

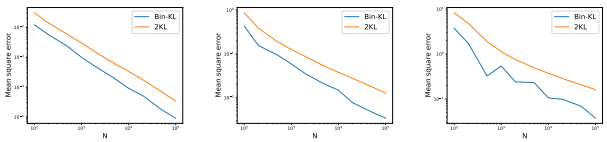
Now we show the plots of the mean square error vs the sample size. Both two coordinates are set to be log scale, so that we can have a clear view of the convergence rates. In each plot, we compare the results of the Bin-KL method and the 2KL method. Each point on the curves in the figures are averaged from $T = 500$ trials. The sample sizes range from 100 to $100,000$. The results are shown in Figures 1, 2, 3, 4 respectively for the four different cases discussed above.

From Figures 1, 2, 3 and 4, it can be observed that the mean square errors of both Bin-KL method and 2KL method converge to zero as the sequence length N increases. Both methods have comparable performance, with each one being slightly better than the other one depending on the underly distribution. For the first and the second distribution, the 2KL method performs better, while for the third and fourth distribution, the Bin-KL method performs better.



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$

Fig. 3. Plots of the mean square error of the Bin-KL method and the 2KL method for the third distribution.



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$

Fig. 4. Plots of the mean square error of the Bin-KL method and the 2KL method for the fourth distribution.

Distribution	d	Bin	2KL	Theoretical rate
1	1	0.71	1.05	0.67
	2	0.51	0.58	0.50
	3	0.33	0.40	0.33
2	1	0.77	1.02	0.67
	2	0.56	0.57	0.50
	3	0.35	0.39	0.33
3	1	0.98	1.01	0.67
	2	0.52	0.50	0.50
	3	0.34	0.33	0.33
4	1	1.04	0.97	0.67
	2	0.67	0.58	0.50
	3	0.59	0.56	0.33

TABLE I
THE EMPIRICAL AND THEORETICAL CONVERGENCE RATES OF BIN-KL AND 2KL METHOD FOR THE ENTROPY RATE ESTIMATION.

Moreover, it can be observed that the curves of 2KL method are smooth, while for the Bin-KL method, the curves are less smooth, especially when the dimension is high. This is because according to (14), m does not change continuously with N. When the sequence length N reaches some threshold, m suddenly changes, and thus the resulting mean square error changes abruptly. As a result, there are usually several turning points in the curve. This effect is especially obvious if the dimensionality is high.

Finally, we list the empirical convergence rates of the Bin-KL method and the 2KL method for different cases, and compare them with the theoretical rates. The empirical rates are calculated by finding the negative slope of the curves by linear regression, while the theoretical rates come from (11). The results are shown in Table I, in which the theoretical rate is denoted as β if the mean square error converges with $\mathcal{O}(N^{-\beta})$ or $\mathcal{O}(N^{-\beta}\text{poly}(\ln N))$, in which poly denotes any polynomial.

From Table I, it can be observed that some empirical convergence rates agree with the theoretical rates, and other empirical results are actually faster than the theoretical one. We explain such difference as following. The assumption we make is a relatively weak condition, and practical distributions may satisfy some stronger conditions. For example, we assume that the transition kernel p is Lipschitz, while actually p may be second order smooth, i.e. p may have bounded Hessian in the support. As a result, the real convergence rate of the mean square error can actually be faster than our theoretical prediction.

VI. CONCLUSION

In this paper, we have proposed two methods to estimate the entropy rate of Markov chains, in which each step follows a continuous distribution. The first method, the 2KL method, combines two KL entropy estimators directly. This method is intuitively correct but hard to analyze. The second method, the Bin-KL method, is based on a hybrid of bin splitting and the KL entropy estimator. For this method, we have conducted a theoretical analysis to show the consistency of this method, and have derived its convergence rate. Numerical simulations have shown that both methods perform well for the distributions satisfying our assumptions, and the convergence rate of the Bin-KL method agrees with our theoretical prediction.

REFERENCES

[1] A. O. Schmitt and H. Herzel, "Estimating the entropy of DNA sequences," *Journal of Theoretical Biology*, vol. 188, no. 3, pp. 369–377, 1997.

[2] J. K. Lanctot, M. Li, and E.-h. Yang, "Estimating DNA sequence entropy," in *Proc. ACM-SIAM Symposium on Discrete Algorithms*, vol. 9, no. 11, San Francisco, Feb 2000, pp. 409–418.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[4] T. Koski, *Hidden Markov models for bioinformatics*. Springer Science & Business Media, 2001, vol. 2.

[5] S. Vinga, "Information theory applications for biological sequence analysis," *Briefings in Bioinformatics*, vol. 15, no. 3, pp. 376–389, 2014.

[6] A. O. Hero, B. Ma, O. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, 2002.

[7] G.-S. Fu, R. Phlypo, M. Anderson, X.-L. Li, and T. u. l. Adali, "Blind source separation by entropy rate minimization," *IEEE Trans. Signal Processing*, vol. 62, no. 16, pp. 4245–4255, 2014.

[8] X.-L. Li and T. Adali, "Blind separation of noncircular correlated sources using Gaussian entropy rate," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2969–2975, 2011.

[9] C. Krumme, A. Llorente, M. Cebrian, E. Moro *et al.*, "The predictability of consumer visitation patterns," *Scientific Reports*, vol. 3, p. 1645, 2013.

[10] W. Xiong, H. Li, T. Adali, Y.-O. Li, and V. D. Calhoun, "On entropy rate for the complex domain and its application to i.i.d. sampling," *IEEE Trans. Signal Processing*, vol. 58, no. 4, pp. 2409–2414, 2010.

[11] J. Lai and J. J. Ford, "Relative entropy rate based multiple hidden markov model approximation," *IEEE Trans. Signal Processing*, vol. 58, no. 1, pp. 165–174, 2010.

[12] 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, no. 6, pp. 1250–1258, 1989.

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

[14] S. Kamath and S. Verdú, "Estimation of entropy rate and Rényi entropy rate for Markov chains," in *Proc. IEEE Intl. Symposium on Inform. Theory*, Barcelona, Spain, Jul 2016, pp. 685–689.

[15] Y. Han, J. Jiao, C.-Z. Lee, T. Weissman, Y. Wu, and T. Yu, "Entropy rate estimation for Markov chains with large state space," in *Advances in Neural Information Processing Systems*, 2018, pp. 9781–9792.

[16] 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. ACM Symposium on Theory of Computing*, San Jose, CA, Jun 2011, pp. 685–694.

[17] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3702–3720, Mar 2016.

[18] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, Oct 1987.

[19] A. B. Tsybakov and E. Van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scandinavian Journal of Statistics*, pp. 75–83, Mar 1996.

[20] G. Biau and L. Devroye, *Lectures on the nearest neighbor method*. Springer, 2015.

[21] S. Singh and B. Póczos, "Analysis of k-nearest neighbor distances with application to entropy estimation," *arXiv preprint arXiv:1603.08578*, 2016.

[22] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *arXiv preprint arXiv:1606.00304*, 2016.

[23] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Trans. Inform. Theory*, Feb 2018.

[24] P. Zhao and L. Lai, "Analysis of kNN information estimators for smooth distributions," *IEEE Trans. Inform. Theory*, vol. 66, no. 6, pp. 3798–3826, Jun 2020.

[25] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, June 2004.

[26] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec 2018, pp. 3156–3167.

[27] P. Zhao and L. Lai, "Minimax optimal estimation of KL divergence for continuous distributions," *IEEE Trans. Inform. Theory*, vol. 66, no. 12, pp. 7787 – 7811, Dec. 2020.