Efficient Approximations for Matrix-Based Rényi's Entropy on Sequential Data

Yuxin Dong, Tieliang Gong, Hong Chen, and Chen Li

***Abstract*— The matrix-based Rényi's entropy (MBRE) has recently been introduced as a substitute for the original Rényi's entropy that could be directly obtained from data samples, avoiding the expensive intermediate step of density estimation. Despite its remarkable success in a broad of information-related tasks, the computational cost of MBRE, however, becomes a bottleneck for large-scale applications. The challenge, when facing sequential data, is further amplified due to the requirement of large-scale eigenvalue decomposition on multiple dense kernel matrices constructed by sliding windows in the region of interest, resulting in $O(mn^3)$ overall time complexity, where m and n denote the number and the size of windows, respectively. To overcome this issue, we adopt the static MBRE estimator together with a variance reduction criterion to develop randomized approximations for the target entropy, leading to high accuracy with substantially lower query complexity by utilizing the historical estimation results. Specifically, assuming that the changes of adjacent sliding windows are bounded by $\beta \ll 1$, which is a trivial case in domains, e.g., time-series analysis, we lower the complexity by a factor of $\sqrt{\beta}$. Polynomial approximation techniques are further adopted to support arbitrary α orders. In general, our algorithms achieve $O(mn^2\sqrt{\beta}st)$ total computational complexity, where $s, t \ll n$ denote the number of vector queries and the polynomial degrees, respectively. Theoretical upper and lower bounds are established in terms of the convergence rate for both s and t, and large-scale experiments on both simulation and real-world data are conducted to validate the effectiveness of our algorithms. The results show that our methods achieve promising speedup with only a trivial loss in performance.**

***Index Terms*— Information theory, matrix-based Rényi's entropy (MBRE), randomized numerical linear algebra, signal processing.**

Manuscript received 16 September 2022; revised 24 June 2023; accepted 7 September 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0110700; in part by the National Natural Science Foundation of China under Grant 62106191, Grant 12071166, Grant 62192781, and Grant 61721002; in part by the Innovation Research Team of Ministry of Education under Grant IRT_17R86; in part by the Project of China Knowledge Center for Engineering Science and Technology; and in part by the Project of Chinese Academy of Engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China." *(Corresponding author: Tieliang Gong.)*

Yuxin Dong, Tieliang Gong, and Chen Li are with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: dongyuxin@stu.xjtu.edu.cn; adidasgtl@gmail.com; cli@xjtu.edu.cn).

Hong Chen is with the College of Science, Huazhong Agriculture University, Wuhan 430070, China (e-mail: chenh@mail.hzau.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3314089.

Digital Object Identifier 10.1109/TNNLS.2023.3314089

I. Introduction

THE classical Rényi entropy covers a family of different entropy measures through a hyperparameter α ($\alpha > 0$ and $\alpha \neq 1$), including Shannon entropy ($\alpha \to 1$), min entropy ($\alpha \to \infty$), and collision entropy ($\alpha = 2$), making it widely adopted in machine learning and statistical inference tasks. However, exact access to Rényi entropy requires complete knowledge about the underlying data distribution, which is usually prohibitive in real-world applications especially in high-dimensional cases. As a substitute, the recently developed matrix-based Rényi's entropy (MBRE) [1] receives considerable attention. It is defined on the eigenspectrum of projected data points in kernel space, which enables us to directly calculate entropy values from given data, avoiding the expensive density estimations. This intriguing property makes it widely adopted in various data science applications, including classical statistical tasks [2], [3], [4] and advanced deep learning algorithms [5], [6], [7], e.g., the information bottleneck principle [8], [9], [10].

Besides static scenarios, applications of entropy on sequential data have also gained considerable interest recently, e.g., regularity quantification in time-series analysis [11], [12], [13] and adaptive filtering based on minimum error entropy (MEE) criterion [14], [15]. Albeit attaining elegant performance, calculating MBRE requires $O(n^3)$ (n is the number of data points) time complexity through traditional eigenvalue decomposition techniques, e.g., CUR decomposition and QR factorization, leading to unacceptable computational costs for large-scale applications. Especially, when dealing with sequential data, consecutively calculating MBRE of the sliding windows in regions of interest becomes computationally prohibitive. Several attempts have been made on this topic: a very recent work [16] developed static approximation algorithms based on random numerical linear algebra, lowering the computational cost to $O(n^2 s)$ ($s \ll n$) with optimal statistical guarantees. However, it still remains a challenge to meet the real-time requirements.

Inspired by the recent advancement in dynamic implicit trace estimation [17], we utilize historical approximation results to reduce the computational cost, and adopt a shrinkage factor γ to control error accumulation. This strategy further improves the overall time complexity by a factor of $\sqrt{\beta}$, where $\beta \ll 1$ upper bounds the relative changes of adjacent kernel matrices in sequential scenarios. We further adopt polynomial approximation techniques to build approximations for noninteger-order Rényi's entropy. A theoretical analysis in

terms of approximation error is provided for all established algorithms, which are then validated by large-scale simulation and real data experiments. We summarize the main contributions of this work as follows.

1) We develop efficient approximations for sequential MBRE with variance reduction and polynomial approximation techniques. Our algorithms further improve the efficiency by a factor of $\sqrt{\beta}$ compared with directly applying [16], lowering the overall complexity from $O(mn^3)$ to $O(mn^2\sqrt{\beta}st)$, where m is the number of kernel matrices, $s \ll n$ is the number of random queries for each kernel matrix, and $t \ll n$ is the polynomial degree.
2) We theoretically analyze the performance of all established algorithms and provide both upper and lower bounds in terms of estimation error. Our results indicate that the achieved convergence rate coefficients $O(\sqrt{\beta}/\epsilon|1-\alpha|)$ are nearly optimal up to a logarithmic factor.
3) Experimental results on both simulation studies and real-world fault detection tasks of sequential data show that the proposed approximation methods achieve promising speedup with only a trivial loss in performance.

II. Preliminaries

The MBRE was first proposed in [1]. Unlike the original Rényi's entropy defined on the probability distribution of random variables, this matrix-based variation could be directly obtained from sampled data points, avoiding the expensive density estimation operation while retraining the elegant performance and scalability of the original one.

Definition 1 [1]: Let $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a real-valued positive kernel that is also infinitely divisible [18]. Given $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$, each $\mathbf{x}_i$ being a real-valued scalar or vector, and the Gram matrix K obtained from $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, a matrix-based analog to Rényi's α-entropy can be defined as follows:

$$\mathbf{S}_\alpha\left(\{\mathbf{x}_i\}_{i=1}^n, \kappa\right) \triangleq \mathbf{S}_\alpha(A) = \frac{1}{1-\alpha}\log(\mathrm{tr}(A^\alpha)) = \frac{1}{1-\alpha}\log\left[\sum_{i=1}^n \lambda_i^\alpha(A)\right]$$

where $A_{ij} = K_{ij}/n(K_{ii}K_{jj})^{1/2}$ is a normalized kernel matrix and $\lambda_i(A)$ is the ith eigenvalue of A.

The constructed matrix A is symmetric semipositive definite (SPD) and normalized, i.e., the eigenvalues are in $[0, 1]$ and satisfy $\sum_{i=1}^n \lambda_i(A) = \mathrm{tr}(A) = 1$. Let the maximum and minimum eigenvalues be $\mu \in [1/n, 1]$ and $\nu \in [0, 1/n]$, respectively, and the condition number of A is then $\kappa = \mu/\nu$. Note that both μ and ν could be estimated by various numerical approaches, e.g., power iteration and restarted Lanczos algorithm with complexity far less than $O(n^3)$. We formulate the sequential entropy estimation problem as follows.

Problem 1 (Sequential Entropy Estimation): Let $\mathbf{x}_1, \ldots, \mathbf{x}_N, \ldots$ be sequential data with stationary distribution, $\mathbf{X}_j = \{\mathbf{x}_j, \ldots, \mathbf{x}_{j+n-1}\}$ be overlapping sliding windows, and $I_1, \ldots, I_m$ be indexes that satisfy $I_{i+1} - I_i \in [1, \beta_0]$ for all $i \in [1, m-1]$. Let $A_1, \ldots, A_m \in \mathbb{R}^{n\times n}$ be normalized kernel matrices constructed by $\mathbf{X}_{I_1}, \ldots, \mathbf{X}_{I_m}$ with kernel κ, respectively. The goal is to compute approximations $\tilde{S}_\alpha(A_1), \ldots, \tilde{S}_\alpha(A_m)$ for $S_\alpha(A_1), \ldots, S_\alpha(A_m)$, such that for each $i \in [1, m]$

$$\mathbb{P}\left[\left|\tilde{S}_\alpha(A_i) - S_\alpha(A_i)\right| \geq \epsilon \cdot S_\alpha(A_i)\right] \leq \delta.$$

This problem arises when the dynamic behavior of entropy value in the region of interest is requested with both high accuracy and fine granularity. For example, in the bearing fault detection task [19], one expects to detect suspicious rises in entropy value as soon as possible. Assuming that the maximum interval of adjacent window positions is at most β_0, we expect to approximate sequential entropy values efficiently by utilizing the similarity of adjacent kernel matrices. Typically, we have $\beta_0 \ll n$ for common situations.

III. Approximating MBRE for Sequential Data

In this work, we seek for efficient solutions of Problem 1 under the implicit matrix computation model. Given access to an oracle for computing $A_{c_1}r_1, \ldots, A_{c_s}r_s$, where $r_1, \ldots, r_s$ and $c_1, \ldots, c_s$ are the possibly adaptively chosen vectors and indexes of matrices, respectively, our goal is to compute approximations to the objectives $S_\alpha(A_1), \ldots, S_\alpha(A_m)$ with minimum number of vectors s. In this section, we provide both integer and noninteger α-order algorithms for sequential approximation of MBRE.

A. Randomized Approximation

The previous work [16] developed efficient algorithms for approximating static MBRE by stochastic trace estimation techniques, where both cases of integer and noninteger-order α are considered. It argued that the following relative error guarantee holds with high probability if the number of random queries is chosen by $s = O(1/\epsilon)$:

$$\left|\tilde{S}_\alpha(A) - S_\alpha(A)\right| \leq \epsilon \cdot S_\alpha(A). \tag{1}$$

Intuitively, a straightforward solution for Problem 1 could be derived by simply applying the static approximation algorithms to every kernel matrix A_i, $i \in [1, m]$. To achieve the same error bound as (1) for all m kernel matrices, $O(m/\epsilon)$ matrix-vector multiplications are required in total. However, this approach does not take the bounded window intervals β_0 assumption into consideration, which could be utilized to further reduce the overall complexity.

Recently, [17] investigated the dynamic implicit trace estimation problem: given access to a matrix-vector multiplication oracle for a dynamically changing matrix with changes bounded in nuclear norm, maintain an approximation to its trace. By noticing the linearity of the trace operator, they adopted a variance reduction scheme and successfully utilize historical approximations to reduce the computational burden, achieving a $\sqrt{\beta}$ factor improvement of the upper bound, as shown in Algorithm 1 and Lemma 1.

Lemma 1 [17, Th. 4.3]: Let $A_1, \ldots, A_m$ be positive semidefinite matrices that satisfy

$$\|A_i\|_* \leq 1, \quad \text{for all } i \in [1, m]$$
$$\|A_{i+1} - A_i\|_* \leq \beta, \quad \text{for all } i \in [1, m-1]$$

Algorithm 1 Dynamic Trace Estimation Algorithm [17]

Input: Normalized kernel matrices $A_1, \ldots, A_m$, positive matrix function $f(A)$, number of random vectors s_0 and s, shrinkage factor $\gamma \in [0, 1]$.
Output: Approximations to $\mathrm{tr}(f(A_1)), \ldots, \mathrm{tr}(f(A_m))$.
Denote $h_s(\cdot)$ as the output of Hutch++ algorithm [20] with s random vectors.
Initialize $\tilde{\mathrm{tr}}(f(A_1)) \leftarrow h_{s_0}(f(A_1))$.
for $j \leftarrow 2, \ldots, m$ **do**
  Calculate $\tilde{\mathrm{tr}}(f(A_j)) \leftarrow \gamma \cdot h_s(f(A_j)) + (1-\gamma)\big(\tilde{\mathrm{tr}}(f(A_{j-1})) + h_s(f(A_j) - f(A_{j-1}))\big)$.
end for
Return: $\tilde{\mathrm{tr}}(f(A_1)), \ldots, \tilde{\mathrm{tr}}(f(A_m))$.

where $\|\cdot\|_*$ is the nuclear norm. For any ϵ, δ, $\beta \in (0, 1)$, let $\tilde{\mathrm{tr}}(A_1), \ldots, \tilde{\mathrm{tr}}(A_m)$ be the outputs of Algorithm 1 with $\gamma = \beta$, $s_0 = O((1/\delta)^{1/2}/\epsilon)$, and $s = O((\beta/\delta)^{1/2}/\epsilon)$; then, for all $i \in [1, m]$, with probability at least $1 - \delta$, $\big|\tilde{\mathrm{tr}}(A_i) - \mathrm{tr}(A_i)\big| \le \epsilon \cdot \mathrm{tr}(A_i)$.

It should be noted that the dynamic trace estimation problem proposed in [17] mainly concerns trace approximation of given matrix sequences, while Problem 1 aims to approximate the matrix-based Rényi's entropy for sequential data. Beyond that, the overlapping assumption in Problem 1 makes the bounded difference assumption between two adjacent matrices naturally satisfied.

Theorem 1: Let $A_1, \ldots, A_m$ be kernel matrices defined in Problem 1; then, for integer $\alpha \ge 2$

$$\big\|A_{i+1}^\alpha - A_i^\alpha\big\|_F \le O\left(L^\alpha\left[\left(1 + \sqrt{\beta_0/n}\right)^\alpha - 1\right]\right)$$
$$\big\|A_{i+1}^\alpha - A_i^\alpha\big\|_* \le O\left(\left[1 + \beta_0 L/\sqrt{n}\right]^\alpha - 1\right)$$

where $L = \max_{x,y\in\mathcal{X}} \kappa(x, y)/(\kappa(x, x)\kappa(y, y))^{1/2}$ is a constant that depends only on the kernel κ.

Remark 1: Theorem 1 establishes the connection between Problem 1 and the dynamic trace estimation problem. It shows that the bounded changes requisite of Lemma 1 is naturally satisfied given our assumption on sliding window intervals, which further allows us exploiting the variance reduction strategy of Algorithm 1 for entropy approximation. Note that for most kernel functions, e.g., radial basis function (RBF) kernels and polynomial kernels, we have finite L to establish the upper bounds. Otherwise, L can be specified from the given dataset D: $L = \max_{x,y\in D} \kappa(x, y)/(\kappa(x, x)\kappa(y, y))^{1/2}$.

However, this adoption is not straightforward, since the concentration results with respect to trace estimation cannot be directly applied due to the nonlinear nature of MBRE. To establish provable statistical accuracy guarantees, we adopt randomized numerical linear algebra techniques together with matrix permutation theories to obtain both the upper and lower bounds for integer and noninteger α orders in Problem 1.

B. Integer-Order Approach

When $\alpha \in \mathbb{N}^+$, the matrix-vector multiplications $f(A) \cdot \mathbf{v} = A^\alpha \cdot \mathbf{v}$ could be directly calculated given arbitrary vector $\mathbf{v}$ by continuously multiplying A with the previous result for α times. An algorithm for integer-order entropy estimation is then derived accordingly, as shown in Algorithm 2.

Algorithm 2 Sequential Approximation of Integer-Order MBRE

Input: Normalized kernel matrices $A_1, \ldots, A_m$, integer order $\alpha \ge 2$, number of random vectors s_0 and s, shrinkage factor $\gamma \in [0, 1]$.
Output: Approximations to $S_\alpha(A_1), \ldots, S_\alpha(A_m)$.
Run algorithm 1 with s_0, s, γ, $A_1, \ldots, A_m$ and $f(A) = A^\alpha$.
Return: $\tilde{S}_\alpha(A_i) = \frac{1}{1-\alpha} \log \tilde{\mathrm{tr}}(A_i^\alpha)$ for $i = 1, \ldots, m$.

Theorem 2: Let $A_1, \ldots, A_m$ be normalized kernel matrices that satisfy

$$\big\|A_{i+1}^\alpha - A_i^\alpha\big\|_* \le \beta \cdot \max_j \mathrm{tr}\big(A_j^\alpha\big), \quad \text{for all } i \in [1, m-1]$$

where $\|\cdot\|_*$ is the nuclear norm. Let $\tilde{S}_\alpha(A_1), \ldots, \tilde{S}_\alpha(A_m)$ be the outputs of Algorithm 2 with

$$\gamma = \beta, \quad s_0 = O\left(\frac{\rho\sqrt{1/\delta}}{\epsilon}\right), \quad s = O\left(\frac{\rho\sqrt{\beta/\delta}}{\epsilon}\right)$$

where $\rho = \max_j \mathrm{tr}(A_j^\alpha)/\min_j \mathrm{tr}(A_j^\alpha)$; then, for all $i \in [1, m]$, with probability at least $1-\delta$, $\big|\tilde{S}_\alpha(A_i) - S_\alpha(A_i)\big| \le \epsilon \cdot S_\alpha(A_i)$. In total, it requires

$$O\left(m\alpha\rho \cdot \frac{\sqrt{\beta/\delta}}{\epsilon}\right)$$

matrix-vector multiplications involving $A_1, \ldots, A_m$.

Remark 2: Theorem 2 establishes the main quality-of-approximation result for Algorithm 2 that $O(m\sqrt{\beta}/\epsilon)$ random queries in total are sufficient to guarantee the approximation error for all matrices A_i, $i \in [1, m]$, with high probability. Comparing with the straightforward approach $O(m/\epsilon)$, it is improved substantially by a factor of $\sqrt{\beta}$. Dharangutte and Musco [17] also provided a parameter-free version of the DeltaShift++ algorithm, which automatically selects the shrinkage factor γ in each step through a minimal variance criterion.

C. Noninteger-Order Approach

Besides integer orders, fractional α is also frequently encountered in real-world applications. As pointed out by Yu et al. [4], α taking values less than 2 or even 1 could help to improve the performance in information-based feature selection tasks. However, it is not straightforward to extend Algorithm 2 to noninteger circumstances, since the matrix-vector multiplication $f(A) \cdot \mathbf{v}$ is no longer directly acquirable. To tackle this issue, we further introduce polynomial approximation techniques. Among multiple classical polynomial series, e.g., Taylor series, Legendre series, and so on, Chebyshev series usually achieves the fastest convergence rate and yields competitive accuracy compared with the optimal solution [21]. Given an analytic function f defined on $[-1, 1]$, it is defined by

$$f(x) = \frac{c_0}{2} + \sum_{k=1}^{\infty} c_k T_k(x), \quad x \in [-1, 1]$$

Algorithm 3 Sequential Approximation for Noninteger-Order MBRE

1: **Input:** Normalized kernel matrices $A_1, \ldots, A_m$, integer order $\alpha \geq 2$, number of random vectors s_0 and s, polynomial order t, eigenvalue upper and lower bounds μ, ν, shrinkage factor $\gamma \in [0, 1]$.
2: **Output:** Approximations to $S_\alpha(A_1), \ldots, S_\alpha(A_m)$.
3: Run algorithm 1 with s_0, s, γ, $A_1, \ldots, A_m$ and $f(A) = \sum_{k=0}^{t} c_k \hat{T}_k(A)$.
4: **Return:** $\tilde{S}_\alpha(A_i) = \frac{1}{1-\alpha} \log \tilde{\mathrm{tr}}(\sum_{k=0}^{t} c_k \hat{T}_k(A))$ for $i = 1, \ldots, m$.

where the Chebyshev polynomials T_k are defined by $T_0(x) = 1$, $T_1(x) = x$, and $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$ when $k \geq 1$. In the sequential entropy estimation problem, the function to approximate is the power function $f(\lambda) = \lambda^\alpha$ combined with a linear mapping g: $[-1, 1] \to [\nu, \mu]$. Taking the first t major terms, the coefficients c_k, $k \in [0, t]$, could be numerically calculated as follows:

$$c_k = \frac{2}{t+1} \sum_{i=0}^{t} f(x_i) T_k(x_i)$$

where $x_i = \cos(\pi(i + 1/2)/(t + 1))$. We are now able to approximate the matrix-vector multiplications for arbitrary vector $\mathbf{v}$ by calculating $f(A) \cdot \mathbf{v} = c_0\mathbf{v}/2 + \sum_{k=1}^{t} c_k \hat{T}_k(A) \cdot \mathbf{v}$, where $\hat{T}_k = T_k \circ g^{-1}$ for all $k \in [1, t]$.

Theorem 3: Let $A_1, \ldots, A_m$ be normalized kernel matrices that satisfy

$$\left\| A_{i+1}^\alpha - A_i^\alpha \right\|_* \leq \beta \cdot \max_j \mathrm{tr}\left(A_j^\alpha\right), \quad \text{for all } i \in [1, m-1]$$

where $\|\cdot\|_*$ is the nuclear norm. Let $\tilde{S}_\alpha(A_1), \ldots, \tilde{S}_\alpha(A_m)$ be the outputs of Algorithm 3 with

$$\gamma = \beta, \quad s_0 = O\left(\frac{\rho\sqrt{1/\delta}}{\epsilon|1-\alpha|}\right)$$
$$s = O\left(\frac{\rho\sqrt{\beta/\delta}}{\epsilon|1-\alpha|}\right), \quad t = O\left(\sqrt{\kappa}\log\left(\frac{\kappa}{\epsilon|1-\alpha|}\right)\right)$$

where $\kappa = \mu/\nu$ and $\rho = \max_j \mathrm{tr}(A_j^\alpha)/\min_j \mathrm{tr}(A_j^\alpha)$; then, for all $i \in [1, m]$, with probability at least $1 - \delta$, $\left|\tilde{S}_\alpha(A_i) - S_\alpha(A_i)\right| \leq \epsilon \cdot S_\alpha(A_i)$. In total, it requires

$$O\left(m\rho \cdot \frac{\sqrt{\beta\kappa/\delta}}{\epsilon|1-\alpha|}\log\left(\frac{\kappa}{\epsilon|1-\alpha|}\right)\right)$$

matrix-vector multiplications involving $A_1, \ldots, A_m$.

Remark 3: Theorem 3 establishes the relative error bound for Algorithm 3, where explicit orders of s_0, s, and t are given to guarantee the approximation accuracy. Comparing with Theorem 2, there is an additional factor $1/|1-\alpha|$ for s_0 and s, which indicates that it is more difficult for MBRE estimation when $\alpha \approx 1$. Moreover, the order of t mainly depends on κ but not ϵ.

It is worthwhile to note that the Lanczos method [22], [23] is introduced as an alternative approach for matrix function approximation: given implicit matrix $f(A)$ and arbitrary vector b, an approximation to $f(A) \cdot b$ can be calculated by a linear interpolation in Krylov subspace $\{b, A \cdot b, \ldots, A^t \cdot b\}$. This approach could be regarded as an adaptive polynomial approximation technique, where the polynomial coefficients are dynamically chosen according to the properties of A and b, and is shown to outperform explicit approximations, such as the Chebyshev series. However, this approach is not applicable in our scenario, as an unbiased trace estimator is required to control the error accumulation. We leave it for future research to adopt Lanczos methods in sequential entropy estimation.

IV. Lower Bounds

In this section, we prove a lower bound showing that the main coefficients $1/\epsilon$ and $\sqrt{\beta}$ in Theorems 2 and 3 are optimal. The proof is under a finite precision assumption, where the elements in each vector query have bounded precision of b bits. We first establish the lower bound for the general trace estimation problem and then the lower bound for Rényi's entropy through complexity reduction.

Theorem 4: Let $A_1, \ldots, A_m$ be positive semidefinite matrices that satisfy

$$\mathrm{tr}(A_i) \leq 1, \quad \text{for all } i \in [1, m]$$
$$|\mathrm{tr}(A_{i+1}) - \mathrm{tr}(A_i)| \leq \beta, \quad \text{for all } i \in [1, m-1].$$

Any algorithm that accesses $A_1, \ldots, A_m$ via matrix-vector multiplication queries $A_{c_1}r_1, \ldots, A_{c_s}r_s$, where $c_1, \ldots, c_s \in [1, m]$ are indexes of chosen matrices and $r_1, \ldots, r_s$ are possibly adaptively chosen vectors with integer entries in $\{-2^b, \ldots, 2^b\}$, requires

$$s = \Omega\left(\frac{\sqrt{m\beta}}{\epsilon(\log(1/m\beta\epsilon) + b)}\right)$$

such queries to output estimates $\tilde{\mathrm{tr}}(A_1), \ldots, \tilde{\mathrm{tr}}(A_m)$, so that with probability at least (2/3), $\left|\tilde{\mathrm{tr}}(A_i) - \mathrm{tr}(A_i)\right| \leq \epsilon \cdot \mathrm{tr}(A_i)$ for all $i = 1, \ldots, m$.

Remark 4: Note that the prerequisites on matrix traces are weaker than that in Lemma 1, since the nuclear norm is always larger than the trace. The coefficients $s = \Omega(\sqrt{\beta}/\epsilon)$ match the previous result Lemma 1 up to a logarithm factor in limited precision computation models, where b is a constant. However, the mismatch of coefficient m indicates that this lower bound could be further improved.

Corollary 1: Let $A_1, \ldots, A_m$ be normalized $n \times n$ kernel matrices that satisfy

$$\left|\mathrm{tr}\left(A_{i+1}^\alpha\right) - \mathrm{tr}\left(A_i^\alpha\right)\right| \leq \beta \cdot \max_j \mathrm{tr}\left(A_j^\alpha\right), \quad \text{for all } i \in [1, m-1].$$

Any algorithm that accesses $A_1, \ldots, A_m$ via matrix-vector multiplication queries $A_{c_1}r_1, \ldots, A_{c_s}r_s$, where $c_1, \ldots, c_s \in [1, m]$ are indexes of chosen matrices and $r_1, \ldots, r_s$ are possibly adaptively chosen vectors with integer entries in $\{-2^b, \ldots, 2^b\}$, requires

$$s = \Omega\left(\frac{\sqrt{m\beta}}{\epsilon|1-\alpha|\log n(\log(1/m\beta\epsilon|1-\alpha|\log n) + b)}\right)$$

such queries to output estimates $\tilde{S}_\alpha(A_1), \ldots, \tilde{S}_\alpha(A_m)$, so that with probability at least (2/3), $\left|\tilde{S}_\alpha(A_i) - S_\alpha(A_i)\right| \leq \epsilon \cdot S_\alpha(A_i)$ for all $i = 1, \ldots, m$.

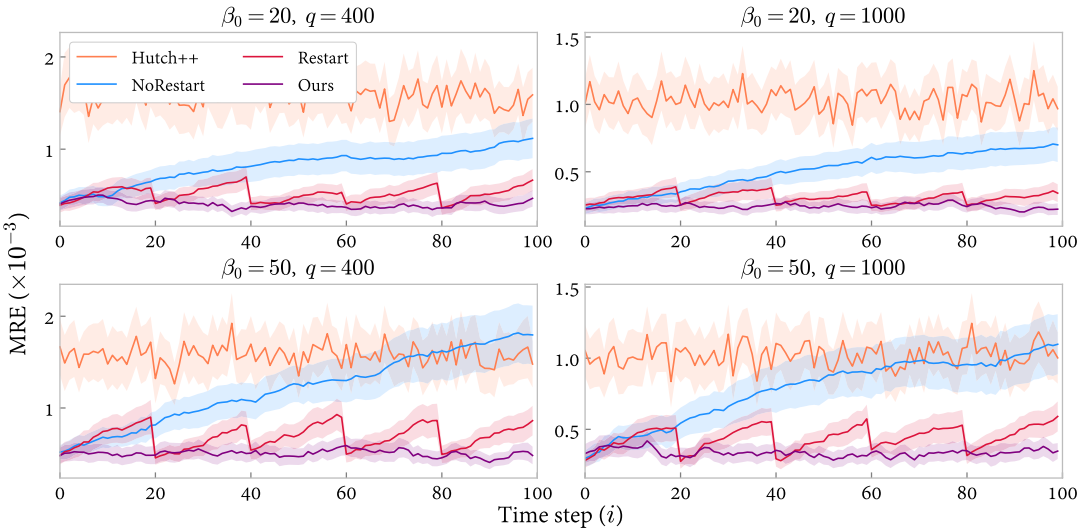


Fig. 1. Comparison of different approaches for integer-order Rényi's entropy estimation.

Remark 5: Corollary 1 presents the lower bounds for both integer and noninteger-order sequential Rényi's entropy estimation. The coefficients $s = O\left(\sqrt{\beta}/\epsilon|1-\alpha|\right)$ match our previous upper bounds in Theorems 2 and 3 up to a logarithm factor.

V. Experimental Results

In this section, we demonstrate the performance of our approximation methods on both synthetic and real-world datasets. The algorithms are implemented in C++ with linear algebra library Eigen, and the experiments are conducted with an Intel i7-10700 (2.90 GHz) CPU and 64 GB of RAM. We adopt the adaptive strategy in [17] to automatically select the hyperparameter γ.

A. Synthetic Data

For synthetic data, we compare our algorithms with three other alternative approaches, namely, the following hold.

1) *Hutch++:* The straightforward approach estimates each $S_\alpha(A_i)$, $i \in [1, m]$, independently.
2) *NoRestart:* Noticing that $\mathrm{tr}(A_{i+1}^\alpha) = \mathrm{tr}(A_i^\alpha) + \mathrm{tr}(A_{i+1}^\alpha - A_i^\alpha)$ for all $i \in [1, m-1]$, we estimate $\mathrm{tr}(A_0^\alpha)$ with s_0 random vectors and each $\mathrm{tr}(A_{i+1}^\alpha - A_i^\alpha)$, $i \in [1, m-1]$, with s random vectors.
3) *Restart:* Similar to NoRestart, but after every r time steps, we restart the progress to alleviate error accumulation by estimating the next matrix with s_0 random vectors.

The synthetic data are generated by mixture of Gaussian distribution $(1/2)N(-1, I_d) + (1/2)N(1, I_d)$ with $n = 5000$ and $d = 10$, where I_d is the $d \times d$ identity matrix. Gaussian kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/2\sigma^2)$ with $\sigma = 1$ is adopted to build kernel matrices in Rényi's entropy. We run each test for $K = 100$ times and report the mean relative error (MRE) curves, where the shaded area indicates $\pm 1/4$ standard deviation (SD). We set number of time steps $m = 100$. It takes around 1.4 h for exactly computing MBRE on all 100 kernel matrices through eigenvalue decomposition techniques.

1) Integer Orders: We first evaluate the performance of Algorithm 2 for integer-order entropy estimation. For each method, we fix the total number of random vectors q during all estimation steps $[1, m]$ and determine the actual values of s and s_0 according to the criterion $s_0 = s(n/\beta_0)^{1/2}$. Two levels of perturbations are considered: 1) $\beta_0 = 20$ and 2) $\beta_0 = 50$, and the MRE versus time step curves are reported in Fig. 1 for $q = 400$ and $q = 1000$. As expected, the straightforward approach Hutch++ brings the highest error. NoRestart achieves low estimation error at the beginning, but loses its advantage after some steps because of error accumulation. Restart expels the accumulated error at each restarting point and is able to control the estimation error to a reasonable range. Among all approaches, Algorithm 2 can completely avoid error accumulation and, thus, yields the best performance with the same number of random queries.

2) Noninteger Orders: We then apply Algorithm 3 for noninteger-order entropy estimation. Unless otherwise noted, we keep the same settings above and set $\alpha = 1.5$, $q = 1000$, and $t = 10$. To begin with, we test the convergence rate of Chebyshev polynomial approximation in Fig. 2, in which the polynomial order is selected in $t \in \{4, 10\}$. As can be seen, a small $t = 10$ is enough to achieve high level of approximation accuracy.

Next, we evaluate the performance of Algorithm 3 for different α entropy orders. As indicated by Theorem 1, we have $\beta = O(\beta_0 L\alpha/\sqrt{n})$ when $\beta_0 \ll n$, which suggests that the difficulty of Problem 1 scales linearly with the magnitude of α. This theoretical finding is verified by our experimental results in Fig. 3, where larger α values correspond to faster error accumulation. Furthermore, the approximation error is significantly larger when $\alpha \approx 1$, which verifies the $(1/(|1-\alpha|))$ factor in our main theorems.

We then test the impact of the Gaussian kernel width σ, where different σ values correspond to different distributions

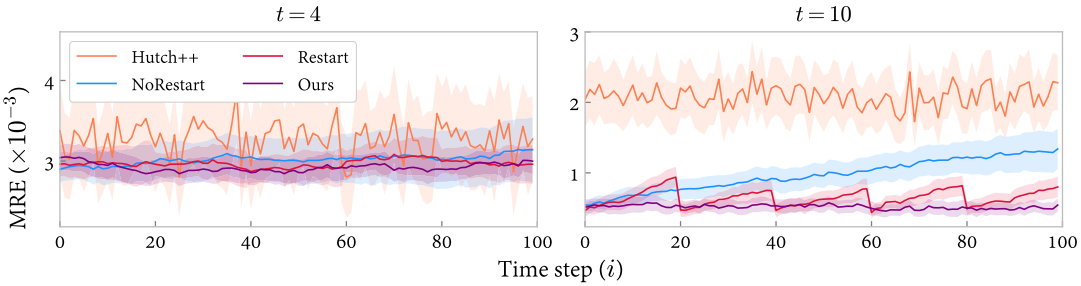


Fig. 2. Comparison of different approaches for noninteger-order Rényi's entropy estimation with different polynomial orders (t).

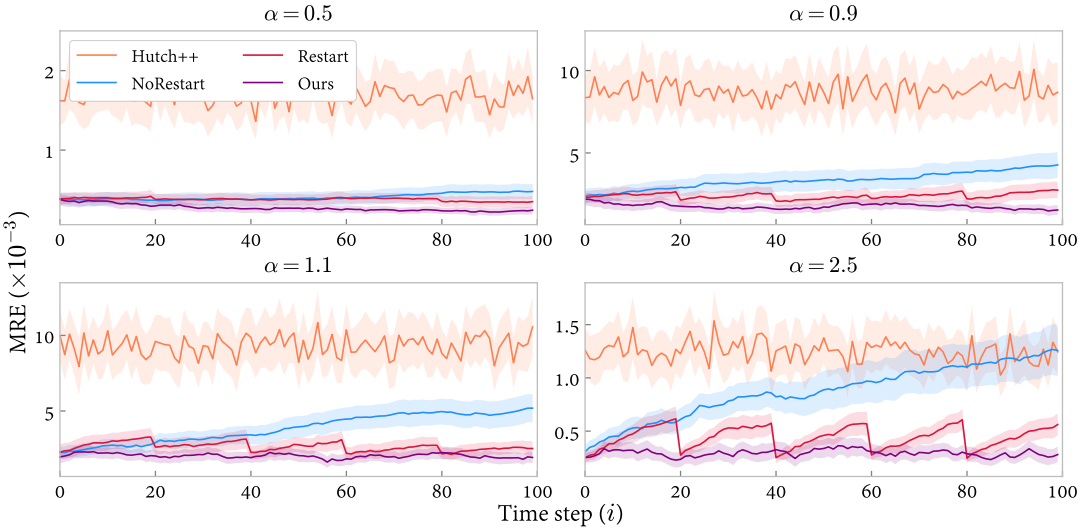


Fig. 3. Comparison of different approaches for noninteger-order Rényi's entropy estimation with different α orders.

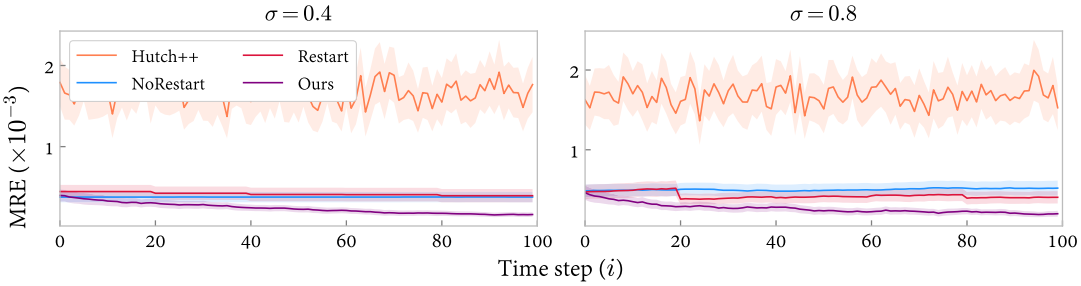


Fig. 4. Comparison of different approaches for noninteger-order Rényi's entropy estimation with different kernel widths (σ).

of the eigenspectrum. Intuitively, a larger σ usually results in faster eigenvalue decay rates. While the convergence rates of s and s_0 are independent with the eigenspectrum, the polynomial order t is dependent on the condition number κ and, thus, results in higher approximation error for larger σ. As shown in Fig. 4, our approximation algorithm can adapt to various eigenspectrum distributions.

We further apply our approximation algorithm on rank-deficient matrices, where the minimum eigenvalue $\nu = 0$. Note that this violates Theorem 3, since the condition number $\kappa = \infty$. However, we demonstrate that our algorithm can still converge in such circumstances. We use the polynomial kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + r)^p$ with $p = 2$ and $r = 1$ to construct rank-deficient kernel matrices and set the dimension of data points as $d = 98$, such that roughly 1% of the eigenvalues are zero. As shown in Fig. 5, Algorithm 3 still achieves high-level approximation accuracies for such extreme cases.

B. Real-World Data

For real-world scenarios, we apply our algorithms to the change detection and diagnosis (CDD) task in rolling element bearing (REB) operations. The goal is to detect change points in given vibrating time-series data, which often indicate the occurrence of serious bearing failures and could lead to further expensive damages to the operating machine.

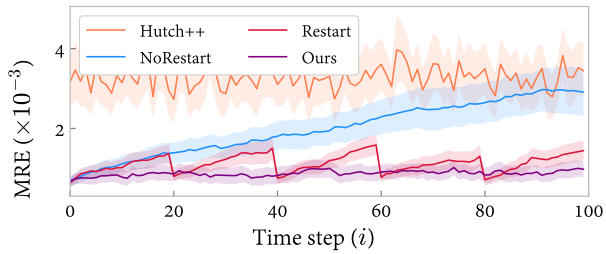


Fig. 5. Comparison of different approaches for noninteger-order Rényi's entropy estimation with rank-deficient kernel matrices.

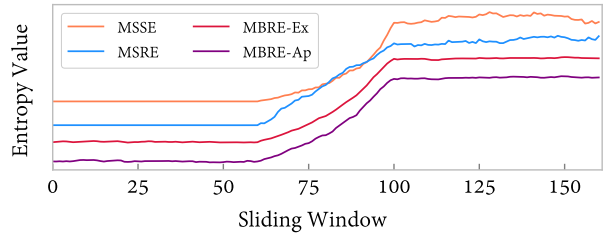


Fig. 6. Entropy of sliding windows evaluated for concatenated data $[x_0(t), x_{11}(t)]$. The curves are scaled and shifted in the y-axis for comparison.

Bearing faults data are provided by the Bearing Data Center, Case Western Reserve University [24]. In this experiment, we follow the experiment settings used in [19], in which 12 signals of drive end accelerometer data are used, namely, $x_i(t)$, $i \in [0, 11]$, each containing 5000 samples. $x_0(t)$ is recorded in a fault-free setting, while other signals $x_1(t), \ldots, x_{11}(t)$ cover multiple bearing fault types: $F0$ (no fault), $F1$ (inner race), $F2$ (ball), and $F3$ (outer race), and fault sizes: $S0$ (0.000″), $S1$ (0.007″), $S2$ (0.014″), $S3$ (0.021″), and $S4$ (0.028″). We compare our methods with two state-of-the-art entropy-based CDD algorithms [19]: multiscale sample entropy (MSSE) and multiscale Rényi's entropy (MSRE). For MBRE, we consider two approaches: exactly computing all entropy values through eigenvalue decomposition (Ex) and our approximation algorithm (Ap) with $s_0 = 100$ and $s = 20$.

1) Single-Fault Detection: In this experiment, the fault-free data $x_0(t)$ are concatenated with $x_1(t), \ldots, x_{11}(t)$, respectively, resulting in 11 signals of length 10 000 that constitute our CDD benchmark. We set the order $\alpha = 3$ in Rényi's entropy and the number of scales $\tau = 10$ in MSSE and MSRE. We use the Gaussian kernel with kernel width σ adaptively selected as the average $k = 10$ nearest Euclidean distances over all samples.

To detect the occurrence of bearing faults, we divide the input signal into overlapping sliding windows and calculate the entropy $e(t)$ for each sliding window. In general, a larger window size n allows us to acquire more accurate entropy values, and a smaller sliding interval β_0 helps detect change points to a higher granularity. Note that both these two factors will result in high-computational cost, since exact access to MBRE requires n^3/β_0. We will show that with our approximation techniques, the time cost of approximated MBRE is no longer unacceptable, while the performance remains high compared with other entropy measures.

TABLE I
COMPARISON OF DIFFERENT ENTROPY MEASURES FOR SINGLE-FAULT DETECTION

Method	Time (s)	IoU	W
MSSE	15.2	0.386	127.8
MSRE	127	0.745	205.4
MBRE-Ex	435	0.934	328.8
MBRE-Ap	108	0.930	253.8

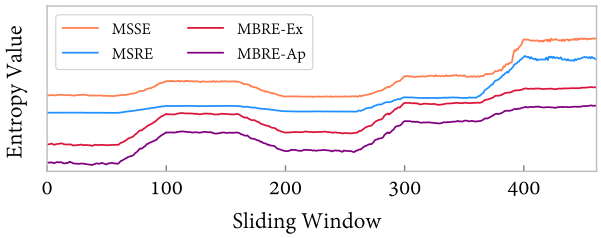


Fig. 7. Entropy of sliding windows evaluated for concatenated data $[x_0(t), x_1(t), x_4(t), x_7(t), x_{10}(t)]$. The curves are scaled and shifted in the y-axis for comparison.

We set $n = 2000$ and $\beta_0 = 50$ in this experiment. To our observation, signals containing bearing faults usually result in higher entropy, so the evaluated entropy values are expected to increase dramatically once the fault occurs, as shown in Fig. 6. Two key properties influence most for the detection accuracy: 1) stability, i.e., entropy values should be stable and do not vary much in the same signal and 2) distinguishability, i.e., the changes of entropy values should be significant enough to be identified when the signal changes. We then design the evaluation criteria as follows.

1) The intersection over union (IoU) score between the intervals that entropy values increase or decrease, and the interval of sliding windows that contain the change point $t = 5000$ (i.e., $t \in [60, 100]$ in Fig. 6).
2) The test statistic W in hypothesis test for the mean between the entropy values of the two signals: $W = |M_1 - M_2|/((V_1 + V_2)/w)^{1/2}$, where M_1, M_2, V_1, and V_2 are the mean and variance of $e(i)$, $i \in [0, 60]$, and $e(j)$, $j \in [100, 160]$, respectively. $w = 40$ is the number of sliding windows.

We use the dynamic programming algorithm in Python package ruptures to detect the increasing or decreasing interval of entropy series, and the final results are reported in Table I. As can be seen, the exactly calculated MBRE achieves significantly higher performance over other methods. Our approximated MBRE achieves more than four times speedup compared with exact computation, while only introducing a minimal drop in IoU.

2) Multiple Fault Detection: We additionally consider signals with multiple faults. Following the benchmark settings in [25], we concatenate signals of the same type of fault but different sizes.

1) $s_1(t) = [x_0(t), x_1(t), x_4(t), x_7(t), x_{10}(t)]$.
2) $s_2(t) = [x_0(t), x_2(t), x_5(t), x_8(t), x_{11}(t)]$.
3) $s_3(t) = [x_0(t), x_3(t), x_6(t), x_9(t)]$.

We keep the experiment settings above and compare the performance of different information measures on these three

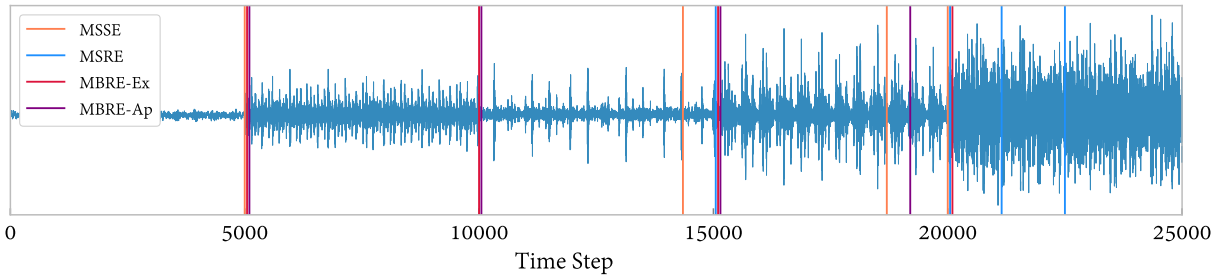


Fig. 8. Multiple fault detection result of different methods on signal $s_1(t)$.

TABLE II
COMPARISON OF DIFFERENT ENTROPY MEASURES FOR MULTIPLE FAULT DETECTION

Method	Time (s)	IoU	W
MSSE	41.9	0.471	79.4
MSRE	351	0.512	92.3
MBRE-Ex	1182	0.796	172.0
MBRE-Ap	284	0.740	146.9

signals in terms of IoU and the minimum W statistic for all change points. The entropy curves of signal $s_1(t)$ are shown in Fig. 7, for example, and the results are reported in Table II. It can be seen that the MBRE still outperforms other methods in terms of both IoU and W. The change points of signal $s_1(t)$ detected by different entropy measures are shown in Fig. 8, and it is worth noting that MBRE-Ex achieves IoU $= 1$ for this signal.

VI. Conclusion

In this article, we develop efficient approximating methods for MBRE on sequential data. By utilizing historical approximation results following a variance reduction criterion, our method achieves substantially higher performance than the previous approach. Through the further adaptation of Chebyshev polynomials, we support arbitrary α orders in MBRE. We establish statistical upper and lower bounds in terms of approximating accuracy for our algorithms, which show that the main convergence coefficients $\sqrt{\beta}/\epsilon$ are nearly optimal. Both synthetic and real-world experiments are conducted to support our theoretical analysis, showing promising speedup for various machine learning tasks with only trivial loss in performance.

Appendix

A. Proof of Theorem 1

Proof: Note that for all $i \in [1, m]$, the sliding windows I_{i+1} and I_i have at least $n-\beta_0$ overlapping data samples, $\Gamma = A_{i+1} - A_i$ is a low-rank matrix through a proper reorder of the samples in I_{i+1}, and I_i: $\Gamma_{jk} = 0$ if $\min(i, j) > \beta_0$. Combining with the fact that the diagonal elements of the normalized kernel matrix A_i are $1/n$, Γ has at most $2n\beta_0 - \beta_0^2 - \beta_0$ nonzero elements; therefore

$$\|A_{i+1} - A_i\|_F = \|\Gamma\|_F \le L\sqrt{\frac{2n\beta_0 - \beta_0^2 - \beta_0}{n^2}}.$$

From the low-rank property of Γ, we know that the values of Γ have at most $2\beta_0$ nonzero eigenvalues. Considering that the Frobenius norm is the l_2 norm of the eigenvalues, the nuclear norm reaches its maximum when all nonzero eigenvalues of Γ are equal

$$\|A_{i+1} - A_i\|_* = \|\Gamma\|_* \le 2\beta_0 L\sqrt{\frac{2n - \beta_0 - 1}{2n^2}}.$$

We first consider integer α orders. Let $R_k^\alpha(A, B)$ be the sum of all terms in the expansion of $(A+B)^\alpha$ that the power of matrix A is k, for example, $(A+B)^2 = A^2 + AB + BA + B^2$, so $R_1^2(A, B) = AB + BA$. Let $p \in \{F, *\}$ be any of Frobenius norm or nuclear norm. For integer $\alpha \ge 2$, we have

$$\begin{aligned}
\left\|A_{i+1}^\alpha - A_i^\alpha\right\|_p &= \left\|(A_i+\Gamma)^\alpha - A_i^\alpha\right\|_p \\
&= \left\|\sum_{k=0}^{\alpha} R_k^\alpha(A_i, \Gamma) - A_i^\alpha\right\|_p \\
&= \left\|\sum_{k=0}^{\alpha-1} R_k^\alpha(A_i, \Gamma)\right\|_p \\
&\le \sum_{k=0}^{\alpha-1}\left\|R_k^\alpha(A_i, \Gamma)\right\|_p \\
&\le \sum_{k=0}^{\alpha-1}\binom{\alpha}{k}\|A_i\|_p^k\|\Gamma\|_p^{\alpha-k} + \|A_i\|_p^\alpha - \|A_i\|_p^\alpha \\
&= \sum_{k=0}^{\alpha}\binom{\alpha}{k}\|A_i\|_p^k\|\Gamma\|_p^{\alpha-k} - \|A_i\|_p^\alpha \\
&= \left(\|A_i\|_p + \|\Gamma\|_p\right)^\alpha - \|A_i\|_p^\alpha.
\end{aligned}$$

For Frobenius norm, noticing the diagonal elements of A_i are $1/n$ and the other elements are bounded by L/n, we have $\|A_i\|_F \le (L^2 - L^2/n + 1/n)^{1/2}$ and

$$\begin{aligned}
&\left\|A_{i+1}^\alpha - A_i^\alpha\right\|_F \\
&\le \left(\sqrt{L^2 - L^2/n + 1/n} + L\sqrt{\frac{2n\beta_0 - \beta_0^2 - \beta_0}{n^2}}\right)^\alpha \\
&\quad - \left(\sqrt{L^2 - L^2/n + 1/n}\right)^\alpha \\
&= O\left(L^\alpha\left(1 + \sqrt{\beta_0/n}\right)^\alpha - L^\alpha\right).
\end{aligned}$$

For nuclear norm, noticing that $\mathrm{tr}(A_i) = 1$ and its positive definiteness, we have $\|A_i\|_* = 1$ for all $i \in [1, m]$ and

$$\begin{aligned}\left\|A_{i+1}^{\alpha} - A_i^{\alpha}\right\|_* &\le \left(1 + 2\beta_0 L\sqrt{\frac{2n-\beta_0-1}{2n^2}}\right)^{\alpha} - 1 \\ &= O\left(\left[1 + \beta_0 L/\sqrt{n}\right]^{\alpha} - 1\right).\end{aligned}$$

For most kernel functions, the constant L is finite to build our bounds. For example, the following kernel families.

1) *RBF Kernels:* $\kappa(x, y) = \exp(-(\|x - y\|/b))$, $b > 0$.
2) *Polynomial Kernels:* $\kappa(x, y) = (x^\top y + c)^d$, $c, d > 0$.

For RBF kernels, we have $\kappa(x, y) \le \kappa(x, x) = \kappa(y, y) = 1$ for any given vector norm $\|\cdot\|$; therefore, $((\kappa(x, y))/((\kappa(x, x)\kappa(y, y))^{1/2})) = 1$. For polynomial kernels, denoting θ as the angle between vectors x and y, we have

$$\begin{aligned}\left(\sqrt[d]{\kappa(x, y)}\right)^2 &= \left(x^\top y + c\right)^2 = \left(\|x\|_2\|y\|_2\cos\theta + c\right)^2 \\ &= \|x\|_2^2\|y\|_2^2\cos^2\theta + 2c\|x\|_2\|y\|_2\cos\theta + c^2 \\ &\le \|x\|_2^2\|y\|_2^2 + c\left(\|x\|_2^2 + \|y\|_2^2\right) + c^2 \\ &\le \left(\|x\|_2^2 + c\right)\left(\|y\|_2^2 + c\right) \\ &= (x^\top x + c)(y^\top y + c) \\ &= \sqrt[d]{\kappa(x, x)\kappa(y, y)}.\end{aligned}$$

Therefore, $\kappa(x, y) \le (\kappa(x, x)\kappa(y, y))^{1/2}$. ∎

B. Proof of Theorem 2

Lemma 2 [16, Proposition A.1]: Let μ and ν denote the largest and smallest eigenvalue of an $n \times n$ normalized kernel matrix A, respectively; then

$$\mathrm{tr}(A^{\alpha}) \in \begin{cases}[\eta, n^{1-\alpha}], & \text{for } \alpha < 1 \\ [n^{1-\alpha}, \eta], & \text{for } \alpha > 1\end{cases} \tag{2}$$

where

$$\eta = \frac{1 - \nu n}{\mu - \nu}\cdot\mu^{\alpha} + \frac{\mu n - 1}{\mu - \nu}\cdot\nu^{\alpha}.$$

Furthermore, $|\log\eta| = \Omega(|1-\alpha|)$, and $|\log\eta| = O(|1-\alpha|\log n)$.

Lemma 3 [16, Proposition III.1]: For any $\epsilon \in (0, 1)$ and sufficient large n, if a randomized algorithm $\mathcal{A}$ can estimate the trace of any $n \times n$ SPD matrix A to relative error $1 \pm \epsilon$ with success probability at least $1 - \delta$ using s queries, then $\mathcal{A}$ can be used to estimate $S_\alpha(A) = (1/(1-\alpha))\log\mathrm{tr}(A^{\alpha})$ to relative error $1 \pm \epsilon_0$ with the same success probability using s queries, where $\epsilon = 1 - \min(\eta, 1/\eta)^{\epsilon_0}$. Vice versa for $\epsilon = \max(n^{\alpha-1}, n^{1-\alpha})^{\epsilon_0} - 1$.

Proof of Theorem 2:

Proof: Let $v = \max_j \mathrm{tr}(A_j^{\alpha})$, $u = \min_j \mathrm{tr}(A_j^{\alpha})$, $\rho = v/u$, $\epsilon_0 = 1 - \min(\eta, 1/\eta)^{\epsilon}$, $\epsilon_1 = \epsilon_0/\rho$, and $B_i = A_i^{\alpha}/v$ for all $i \in [1, m]$; we have $\|B_i\|_* \le 1$ for all $i \in [1, m]$ and $\|B_{i+1} - B_i\| \le \beta$ for all $i \in [1, m-1]$. By applying Lemma 1 with $s_0 = O\left(((1/\delta)^{1/2}/\epsilon_1)\right)$ and $s = O\left((((\beta/\delta)^{1/2})/\epsilon_1)\right)$ on matrices $B_1, \ldots, B_m$, we have $\left|\tilde{\mathrm{tr}}(B_i) - \mathrm{tr}(B_i)\right| \le \epsilon_1$ for all $i \in [1, m]$, which means

$$\left|\tilde{\mathrm{tr}}\left(A_i^{\alpha}\right) - \mathrm{tr}\left(A_i^{\alpha}\right)\right| \le \epsilon_1 v \le \epsilon_0 u \le \epsilon_0 \mathrm{tr}\left(A_i^{\alpha}\right), \quad i \in [1, m].$$

Combining with Lemma 3, we have that for all $i \in [1, m]$, with probability at least $1 - \delta$, the approximation error $\left|\tilde{S}_\alpha(A_i) - S_\alpha(A_i)\right| \le \epsilon \cdot S_\alpha(A_i)$. By noticing that

$$\begin{aligned}\epsilon_0 &= 1 - \exp(\epsilon\log\min(\eta, 1/\eta)) \\ &= \Omega(-\epsilon\log\min(\eta, 1/\eta)) = \Omega(\epsilon|\log\eta|)\end{aligned}$$

and combine with Lemma 2, we have $s_0 = O\left(((\rho(1/\delta)^{1/2})/(\epsilon|1-\alpha|))\right)$ and $s = O(((\rho(\beta/\delta)^{1/2})/(\epsilon|1-\alpha|)))$. For integer $\alpha \ge 2$, we have $|1-\alpha| = \Theta(1)$, which completes the proof. ∎

C. Proof of Theorem 3

Lemma 4 [16, Proposition A.6]: Let g be the linear mapping $[-1, 1] \to [\nu, \mu]$ for arbitrary $0 < \nu < \mu < 1$, $f(\lambda) = \lambda^{\alpha}$ be the α-power function, and $p_t(\lambda)$ be the Chebyshev series of degree $t = O\left((\mu/\nu)^{1/2}\log((\mu/\nu\epsilon))\right)$ for function $f \circ g$; then, the following inequality holds:

$$\max_{x\in[-1,1]} |(f\circ g)(x) - p_t(x)| = \max_{\lambda\in[\nu,\mu]} \left|f(\lambda) - q_t(\lambda)\right| \le \epsilon\nu^{\alpha}$$

where $q_t = p_t \circ g^{-1}$.

Proof of Theorem 3:

Proof: Similar to the proof of Theorem 2, let $v = \max_j \mathrm{tr}(A_j^{\alpha})$, $u = \min_j \mathrm{tr}(A_j^{\alpha})$, $\rho = v/u$, $\epsilon_0 = 1 - \min(\eta, 1/\eta)^{\epsilon}$, $\epsilon_1 = \epsilon_0/3\rho$, and $\epsilon_2 = (1/2)\epsilon_0$. By applying Lemma 4 with $t = O\left(\sqrt{\kappa}\log((\kappa/\epsilon_2))\right)$, we have

$$\begin{aligned}|\mathrm{tr}(q_t(A)) - \mathrm{tr}(A^{\alpha})| &\le \sum_{i=1}^{n}\left|f(\lambda_i) - q_t(\lambda_i)\right| \\ &\le n\epsilon_2\nu^{\alpha} \le \frac{\epsilon_0}{2}\cdot\mathrm{tr}(A^{\alpha})\end{aligned}$$

for any $A \in \{A_1, \ldots, A_m\}$. By applying Lemma 1 with $s_0 = O\left((((1/\delta)^{1/2})/\epsilon_3)\right)$ and $s = O\left(((((\beta/\delta)^{1/2})/\epsilon_3)\right)$, where $\epsilon_3 = (1/3)\epsilon_1$, we have that for all $i \in [1, m]$, with probability at least $1 - \delta$, the approximation error is bounded by $\left|\tilde{\mathrm{tr}}(q_t(A_i)) - \mathrm{tr}(q_t(A_i))\right| \le \epsilon_3 \cdot \max_j \mathrm{tr}(q_t(A_j))$.

Noticing that $\mathrm{tr}(q_t(A_i)) \le (\epsilon_0/2)\mathrm{tr}(A_i^{\alpha}) + \mathrm{tr}(A_i^{\alpha}) \le (3/2)\mathrm{tr}(A_i^{\alpha})$ and similarly $\mathrm{tr}(q_t(A_i)) \ge (1/2)\mathrm{tr}(A_i^{\alpha})$ for all $i \in [1, m]$, we have $\max_j \mathrm{tr}(q_t(A_j))/\min_j \mathrm{tr}(q_t(A_j)) \le 3\rho$ and

$$\begin{aligned}\left|\tilde{\mathrm{tr}}(q_t(A_i)) - \mathrm{tr}\left(A_i^{\alpha}\right)\right| &\le \left|\tilde{\mathrm{tr}}(q_t(A_i)) - \mathrm{tr}(q_t(A_i))\right| \\ &\quad + \left|\mathrm{tr}(q_t(A_i)) - \mathrm{tr}\left(A_i^{\alpha}\right)\right| \\ &\le \frac{\epsilon_1}{3}\max_j\ \mathrm{tr}(q_t(A_j)) + \frac{\epsilon_0}{2}\mathrm{tr}\left(A_i^{\alpha}\right) \\ &\le \frac{\epsilon_0}{3}\min_j\ \mathrm{tr}(q_t(A_j)) + \frac{\epsilon_0}{2}\mathrm{tr}\left(A_i^{\alpha}\right) \\ &\le \epsilon_0\cdot\mathrm{tr}\left(A_i^{\alpha}\right).\end{aligned}$$

Combining with Lemma 3, we have $\left|\tilde{S}_\alpha(A_i) - S_\alpha(A_i)\right| \le \epsilon \cdot S_\alpha(A_i)$ for all $i \in [1, m]$.

Finally by applying Lemma 2, we have $s_0 = O\left(((\rho(1/\delta)^{1/2})/(\epsilon|1-\alpha|))\right)$, $s = O(((\rho(\beta/\delta)^{1/2})/(\epsilon|1-\alpha|)))$, and $t = O\left(\sqrt{\kappa}\log((\kappa/(\epsilon|1-\alpha|)))\right)$, which completes the proof. ∎

D. Proof of Theorem 4

Problem 2: Let Alice and Bob be communicating parties who hold vectors $\mathbf{x} \in \{-1, 1\}^c$ and $\mathbf{y} \in \{-1, 1\}^c$, respectively. The gap-Hamming problem asks Alice and Bob to return

$$1, \quad \text{if } \langle \mathbf{x}, \mathbf{y} \rangle \geq \sqrt{c} \qquad \text{and} \qquad -1, \quad \text{if } \langle \mathbf{x}, \mathbf{y} \rangle \leq -\sqrt{c}.$$

Lemma 5 [26, Th. 2.6]: The randomized communication complexity for solving Problem 2 with probability at least $(2/3)$ is $\Omega(c)$ bits.

Proof of Theorem 4:

Proof: Consider Problem 2 with $c = mw$ and $w = n\beta/2$. Let $\mathbf{x} = \{x_1, \ldots, x_c\}$ and $\mathbf{y} = \{y_1, \ldots, y_c\}$ be the held vectors; then, by Lemma 5, the lower bound of solving Problem 2 is $\Omega(c) = \Omega(mn\beta)$ bits.

Let $\mathbf{x}_i = \{x_{(w(i-1) \mod c)+1}, x_{(w(i-1)+1 \mod c)+1}, \ldots, x_{(w(i-1)+n \mod c)+1}\}$, $i \in [1, m]$, be cyclic sliding windows of $\mathbf{x}$ and define $\mathbf{y}_i$ in the same way. Let $X_i \in \{-1, 1\}^{\sqrt{n} \times \sqrt{n}}$ and $Y_i \in \{-1, 1\}^{\sqrt{n} \times \sqrt{n}}$ contain the entries of $\mathbf{x}_i$ and $\mathbf{y}_i$ rearranged into matrices. Let $Z_i = X_i + Y_i$ and let $A_i = Z_i^\top Z_i$. Similarly, we construct X–Z by entries of $\mathbf{x}$ and $\mathbf{y}$. Then, A and A_i are positive semidefinite for all $i \in [1, m]$ and

$$\begin{aligned}
\operatorname{tr}(A_i) &= \|Z_i\|_F^2 = \|\mathbf{x}_i + \mathbf{y}_i\|_2^2 \\
&= \|\mathbf{x}_i\|_2^2 + \|\mathbf{y}_i\|_2^2 + 2\langle \mathbf{x}_i, \mathbf{y}_i \rangle \\
&= 2n + 2\langle \mathbf{x}_i, \mathbf{y}_i \rangle \leq 4n \\
|\operatorname{tr}(A_{i+1}) - \operatorname{tr}(A_i)| &= |2\langle \mathbf{x}_{i+1}, \mathbf{y}_{i+1} \rangle - 2\langle \mathbf{x}_i, \mathbf{y}_i \rangle| \leq 8w \\
\operatorname{tr}(A) &= \|Z\|_F^2 = \|\mathbf{x} + \mathbf{y}\|_2^2 \\
&= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle \\
&= 2mw + 2\langle \mathbf{x}, \mathbf{y} \rangle.
\end{aligned}$$

Normalizing the matrices $A_1, \ldots, A_m$ with $4n$, we have

$$\begin{aligned}
\operatorname{tr}(A_i/4n) &\leq 1 \\
|\operatorname{tr}(A_{i+1}/4n) - \operatorname{tr}(A_i/4n)| &\leq 8w/4n = \beta.
\end{aligned}$$

Noticing that we can get $\operatorname{tr}(A)$ by summing up $\operatorname{tr}(A_i)$

$$\begin{aligned}
\sum_{i=1}^{m} \operatorname{tr}(A_i) &= 2nm + 2\sum_{i=1}^{m} \langle \mathbf{x}_i, \mathbf{y}_i \rangle \\
&= 2nm + \frac{2m}{w} \langle \mathbf{x}, \mathbf{y} \rangle = \operatorname{tr}(A)/\beta.
\end{aligned}$$

If $\langle \mathbf{x}, \mathbf{y} \rangle \geq \sqrt{mw}$, we have $\operatorname{tr}(A) \geq 2mw + 2\sqrt{mw}$, and if $\langle \mathbf{x}, \mathbf{y} \rangle \leq -\sqrt{mw}$, we have $\operatorname{tr}(A) \leq 2mw - 2\sqrt{mw}$. So, if Alice and Bob can approximate all $\operatorname{tr}(A_i/4n)$ to relative error $1 \pm 1/\sqrt{mw}$ with s matrix-vector multiplications, they can approximate $\operatorname{tr}(A)$ to the same relative error and solve Problem 2. It is proven in [20] that they can do so with $O(s \cdot \sqrt{n}(\log n + b))$ bits of communication if the possibly adaptively chosen vectors have integer entries bounded by 2^b. Combining with the $\Omega(mw)$ lower bound for Problem 2, we have that $s = \Omega\big((\sqrt{m\beta}/(\epsilon(\log(1/m\beta\epsilon) + b)))\big)$ queries are needed to approximate all $\operatorname{tr}(A_i/4n)$ to accuracy $1 \pm \epsilon$ for $\epsilon = 1/\sqrt{mw}$, with probability at least $(2/3)$. ∎

E. Proof of Corollary 1

Proof: Let $\epsilon_0 = \max(n^{\alpha-1}, n^{1-\alpha})^\epsilon - 1 \leq \epsilon|1 - \alpha| \log n$. From Lemma 3, we know that if we can approximate all $S_\alpha(A_i)$ to relative error ϵ with s matrix-vector multiplications, then we can approximate all $\operatorname{tr}(A_i)$ to relative error ϵ_0. Combining with Theorem 4, we get the final lower bound

$$\begin{aligned}
s &= \Omega\left(\frac{\sqrt{m\beta}}{\epsilon_0(\log(1/m\beta\epsilon_0) + b)}\right) \\
&= \Omega\left(\frac{\sqrt{m\beta}}{\epsilon|1 - \alpha| \log n(\log(1/m\beta\epsilon|1 - \alpha| \log n) + b)}\right).
\end{aligned}$$

∎

REFERENCES

[1] L. G. Sanchez Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 535–548, Jan. 2015.

[2] A. J. Brockmeier, T. Mu, S. Ananiadou, and J. Y. Goulermas, "Quantifying the informativeness of similarity measurements," *J. Mach. Learn. Res.*, vol. 18, pp. 1–61, Jul. 2017.

[3] A. M. Alvarez-Meza, J. A. Lee, M. Verleysen, and G. Castellanos-Dominguez, "Kernel-based dimensionality reduction using Rényi's α-entropy measures of similarity," *Neurocomputing*, vol. 222, pp. 36–46, Jan. 2017.

[4] S. Yu, L. G. S. Giraldo, R. Jenssen, and J. C. Principe, "Multivariate extension of matrix-based Rényi $\alpha\alpha$-order entropy functional," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2960–2966, Nov. 2020.

[5] S. Yu, F. Alesiani, X. Yu, R. Jenssen, and J. Principe, "Measuring dependence with matrix-based entropy functional," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 12, pp. 10781–10789.

[6] C. H. Sarvani, M. Ghorai, S. R. Dubey, and S. H. S. Basha, "HRel: Filter pruning based on high relevance between activation maps and class labels," *Neural Netw.*, vol. 147, pp. 186–197, Mar. 2022.

[7] R. Miles, A. L. Rodriguez, and K. Mikolajczyk, "Information theoretic representation distillation," 2021, *arXiv:2112.00459*.

[8] B. C. Geiger, "On information plane analyses of neural network classifiers—A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7039–7051, Dec. 2022.

[9] M. Xu, T. Zhang, Z. Li, and D. Zhang, "InfoAT: Improving adversarial training using the information bottleneck principle," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 22, 2022, doi: 10.1109/TNNLS.2022.3183095.

[10] P. Zhai and S. Zhang, "Adversarial information bottleneck," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 20, 2022, doi: 10.1109/TNNLS.2022.3172986.

[11] M. Costa and A. Goldberger, "Generalized multiscale entropy analysis: Application to quantifying the complex volatility of human heartbeat time series," *Entropy*, vol. 17, no. 3, pp. 1197–1203, Mar. 2015.

[12] Y. Yin, P. Shang, and G. Feng, "Modified multiscale cross-sample entropy for complex time series," *Appl. Math. Comput.*, vol. 289, pp. 98–110, Oct. 2016.

[13] L. Montesinos, R. Castaldo, and L. Pecchia, "On the use of approximate entropy and sample entropy with centre of pressure time-series," *J. NeuroEng. Rehabil.*, vol. 15, no. 1, pp. 1–15, Dec. 2018.

[14] S. Peng, W. Ser, B. Chen, L. Sun, and Z. Lin, "Robust constrained adaptive filtering under minimum error entropy criterion," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 8, pp. 1119–1123, Aug. 2018.

[15] G. Wang, B. Peng, Z. Feng, X. Yang, J. Deng, and N. Wang, "Adaptive filtering based on recursive minimum error entropy criterion," *Signal Process.*, vol. 179, Feb. 2021, Art. no. 107836.

[16] Y. Dong, T. Gong, S. Yu, and C. Li, "Optimal randomized approximations for matrix-based Rényi's entropy," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4218–4234, Jul. 2023.

[17] P. Dharangutte and C. Musco, "Dynamic trace estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.

[18] R. Bhatia, "Infinitely divisible matrices," *Amer. Math. Monthly*, vol. 113, no. 3, pp. 221–235, Mar. 2006.

[19] D. Aiordachioaie, T. D. Popescu, and S. Marius Pavel, "On change detection in the complexity of the time series with multiscale Rényi entropy processing," in *Proc. 24th Int. Conf. Syst. Theory, Control Comput. (ICSTCC)*, Oct. 2020, pp. 927–932.

[20] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff, "Hutch++: Optimal stochastic trace estimation," in *Proc. Symp. Simplicity Algorithms (SOSA)*, 2021, pp. 142–155.

[21] S. Xiang, X. Chen, and H. Wang, "Error bounds for approximation in Chebyshev points," *Numerische Math.*, vol. 116, no. 3, pp. 463–491, Sep. 2010.
[22] C. Musco, C. Musco, and A. Sidford, "Stability of the Lanczos method for matrix function approximation," in *Proc. 29th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: SIAM, 2018, pp. 1605–1624.
[23] S. Ubaru, J. Chen, and Y. Saad, "Fast estimation of tr(F(A)) via stochastic Lanczos quadrature," *SIAM J. Matrix Anal. Appl.*, vol. 38, no. 4, pp. 1075–1099, 2017.
[24] (2017). *Case Western Reserve University Bearing Data Center*. [Online]. Available: https://engineering.case.edu/bearingdatacenter
[25] T. D. Popescu and D. Aiordachioaie, "Fault detection of rolling element bearings using optimal segmentation of vibrating signals," *Mech. Syst. Signal Process.*, vol. 116, pp. 370–391, Feb. 2019.
[26] A. Chakrabarti and O. Regev, "An optimal lower bound on the communication complexity of gap-Hamming-distance," *SIAM J. Comput.*, vol. 41, no. 5, pp. 1299–1317, Jan. 2012.

Yuxin Dong received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology.

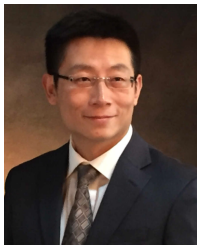
His research interests include information theory, statistical learning theory, and bioinformatics.

Tieliang Gong received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2018.

From September 2018 to October 2020, he was a Post-Doctoral Researcher with the Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada. He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. His current research interests include statistical learning theory, machine learning, and information theory.

Hong Chen received the B.S., M.S., and Ph.D. degrees from Hubei University, Wuhan, China, in 2003, 2006, and 2009, respectively.

From February 2016 to August 2017, he was a Post-Doctoral Researcher with the Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX, USA. He is currently a Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan. His current research interests include machine learning, statistical learning theory, and approximation theory.



Chen Li received the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2014.

From June 2014 to March 2016, he was a Post-Doctoral Researcher with the Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. His research interests include natural language processing, biological text mining, digital pathology, and bioinformatics.