Robust and Fast Measure of Information via Low-Rank Representation

Yuxin Dong[1], Tieliang Gong[1]*, Shujian Yu[2], Hong Chen[3], Chen Li[1]

[1]Xi'an Jiaotong University, China
[2]Vrije Universiteit Amsterdam
[3]Huazhong Agricultural University, China
dongyuxin@stu.xjtu.edu.cn, adidasgtl@gmail.com, yusj9011@gmail.com, chenh@mail.hzau.edu.cn, cli@xjtu.edu.cn

Abstract

The matrix-based Rényi's entropy allows us to directly quantify information measures from given data, without explicit estimation of the underlying probability distribution. This intriguing property makes it widely applied in statistical inference and machine learning tasks. However, this information theoretical quantity is not robust against noise in the data, and is computationally prohibitive in large-scale applications. To address these issues, we propose a novel measure of information, termed low-rank matrix-based Rényi's entropy, based on low-rank representations of infinitely divisible kernel matrices. The proposed entropy functional inherits the specialty of of the original definition to directly quantify information from data, but enjoys additional advantages including robustness and effective calculation. Specifically, our low-rank variant is more sensitive to informative perturbations induced by changes in underlying distributions, while being insensitive to uninformative ones caused by noises. Moreover, low-rank Rényi's entropy can be efficiently approximated by random projection and Lanczos iteration techniques, reducing the overall complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 s)$ or even $\mathcal{O}(ns^2)$, where n is the number of data samples and $s \ll n$. We conduct large-scale experiments to evaluate the effectiveness of this new information measure, demonstrating superior results compared to matrix-based Rényi's entropy in terms of both performance and computational efficiency.

Introduction

The practical applications of traditional entropy measures e.g. Shannon's entropy (Shannon 1948) and Rényi's entropy (Rényi 1961) have long been hindered by their heavy reliance on the underlying data distributions, which are extremely hard to estimate or even intractable in high-dimensional spaces (Fan and Li 2006). Alternatively, the matrix-based Rényi's entropy proposed by (Sanchez Giraldo, Rao, and Principe 2014) treats the entire eigenspectrum of a normalized kernel matrix as a probability distribution, thus allows direct quantification from given data samples by projecting them in reproducing kernel Hilbert spaces (RKHS) without the exhausting density estimation. This intriguing property makes matrix-based Rényi's entropy and its multivariate extensions (Yu et al. 2019) successfully applied in various data science applications, ranging from classical dimensionality reduction and feature selection (Brockmeier et al. 2017; Álvarez-Meza et al. 2017) problems to advanced deep learning problems such as network pruning (Sarvani et al. 2021) and knowledge distillation (Miles, Rodríguez, and Mikolajczyk 2021).

Despite the empirical success of matrix-based Rényi's entropy, it has been shown to be not robust against noises in the data (Yu et al. 2019), because it cannot distinguish them from linear combinations of informative features in high-dimensional scenarios. Moreover, the exact calculation requires $\mathcal{O}(n^3)$ time complexity with traditional eigenvalue decomposition techniques e.g. CUR decomposition and QR factorization (Mahoney and Drineas 2009; Watkins 2008), greatly hampering its application in large scale tasks due to the unacceptable computational cost.

Inspired by the success of min-entropy which uses the largest outcome solely as a measure of information (Wan et al. 2018; Konig, Renner, and Schaffner 2009), we seek for a robust information quantity by utilizing low-rank representations of kernel matrices. Our new definition, termed low-rank matrix-based Rényi's entropy (abbreviated as low-rank Rényi's entropy), fulfills the entire set of axioms provided by Rényi (Rényi 1961) that a function must satisfy to be considered a measure of information. Compared to the original matrix-based Rényi's entropy, our low-rank variant is more sensitive to informative perturbations caused by variation of the underlying probability distribution, while being more robust to uninformative ones caused by noises in the data samples. Moreover, our low-rank Rényi's entropy can be efficiently approximated by random projection and Lanczos iteration techniques, achieving substantially lower time complexity than the trivial eigenvalue decomposition approach. We theoretically analyze the quality of approximation results, and conduct large-scale experiments to evaluate the effectiveness of low-rank Rényi's entropy as well as the approximation algorithms. The main contributions of this work are summarized as follows:

- We extend Giraldo et al.'s definition and show that a measure of entropy can be built upon the low-rank representation of the kernel matrix. Our low-rank definition can be naturally extended to measure the interactions between multiple random variables, including joint entropy,

*Corresponding author.

conditional entropy, and mutual information.

- Theoretically, we show that low-rank Rényi's entropy is more insensitive to random perturbations of the data samples under mild assumptions. We also give empirical examples of low-rank Rényi's entropy achieving higher discriminability for different eigenspectrum distributions through a proper choice of the hyper-parameter k.
- We develop efficient algorithms to approximate low-rank Rényi's entropy through random projection and Lanczos iteration techniques, enabling fast and accurate estimations respectively. The overall complexity is reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 s)$ or even $\mathcal{O}(ns^2)$ for some $s \ll n$, leading to a significant speedup compared to the original matrix-based Rényi's entropy.
- We evaluate the effectiveness of low-rank Rényi's entropy on large-scale synthetic and real-world datasets, demonstrating superior performance compared to the original matrix-based Rényi's entropy while bringing tremendous improvements in computational efficiency.

Related Work

Matrix-based Rényi's Entropy

Given random variable $\mathbf{X}$ with probability density function (PDF) $p(\mathbf{x})$ defined in a finite set $\mathcal{X}$, the α-order Rényi's entropy ($\alpha > 0, \alpha \neq 1$) $\mathbf{H}_\alpha(\mathbf{X})$ is defined as

$$\mathbf{H}_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \log_2 \int_{\mathcal{X}} p^\alpha(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

where the limit case $\alpha \to 1$ yields the well-known Shannon's entropy. It is easy to see that Rényi's entropy relies heavily on the distribution of the underlying variable $\mathbf{X}$, preventing its further adoption in data-driven science, especially for high-dimensional scenarios. To alleviate this issue, an alternative measure namely matrix-based Rényi's entropy was proposed (Sanchez Giraldo, Rao, and Principe 2014):

Definition 1. *Let $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be an infinitely divisible positive kernel (Bhatia 2006). 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 analogue to Rényi's α-entropy can be defined as:*

$$\mathbf{S}_\alpha(\mathbf{A}) = \frac{1}{1-\alpha} \log_2 \Big(\sum_{i=1}^n \lambda_i^\alpha(\mathbf{A}) \Big),$$

where $\mathbf{A}_{ij} = \frac{1}{n} \frac{K_{ij}}{\sqrt{K_{ii} K_{jj}}}$ is a normalized kernel matrix and $\lambda_i(\mathbf{A})$ is the i-th largest eigenvalue of $\mathbf{A}$.

The kernel matrix $\mathbf{A}$ is positive semi-definite (PSD) and satisfies $\mathbf{tr}(\mathbf{A}) = 1$, therefore $\lambda_i \in [0, 1]$ for all $i \in [1, n]$. With this setting, one can similarly define matrix notion of Rényi's conditional entropy $\mathbf{S}_\alpha(\mathbf{A}|\mathbf{B})$, mutual information $\mathbf{I}_\alpha(\mathbf{A}; \mathbf{B})$, and their multivariate extensions (Yu et al. 2019).

Approximating Matrix-based Rényi's Entropy

Exactly calculating $\mathbf{S}_\alpha(\mathbf{A})$ requires $\mathcal{O}(n^3)$ time complexity in general with traditional eigenvalue decomposition techniques. Recently, several attempts have been made towards accelerating the computation of $\mathbf{S}_\alpha(\mathbf{A})$ from the perspective of randomized numerical linear algebra (Gong et al. 2021; Dong et al. 2022). Although we also develop fast approximations, the motivation and technical solutions are totally different: we aim to propose a new measure of information that is robust to noise in data and also enjoys fast computation, whereas Gong and Dong et al. only accelerate the original matrix-based Rényi's entropy. Moreover, in terms of adopted mathematical tools, we mainly focus on random projection and Lanczos iteration algorithms, rather than stochastic trace estimation and polynomial approximation techniques used in their works. As a result, the corresponding theoretical error bounds are also different.

A Low-rank Definition of Rényi's Entropy

Our motivations root in two observations. Recall that the min-entropy (Konig, Renner, and Schaffner 2009), defined by $\mathbf{H}_{\min}(\mathbf{X}) = -\log_2 \max_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x})$, measures the amount of information using solely the largest probability outcome. In terms of quantum statistical mechanics, it is the largest eigenvalue of the quantum state ρ which is PSD and has unit trace (Ohya and Petz 2004). On the other hand, the eigenvalues with the maximum magnitude characterize the main properties of a PSD matrix. Inspired by these observations, we develop a robust information theoretical quantity by exploiting the low-rank representation:

Definition 2. *Let $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be an infinitely divisible kernel. Given $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and integer $k \in [1, n-1]$, the low-rank Rényi's α-order entropy is defined as:*

$$\mathbf{S}_\alpha^k(\mathbf{A}) = \frac{1}{1-\alpha} \log_2 \Big(\sum_{i=1}^k \lambda_i^\alpha(\mathbf{A}) + (n-k) \lambda_r^\alpha(\mathbf{A}) \Big),$$

where $\mathbf{A}$ is the normalized kernel matrix constructed from $\{\mathbf{x}_i\}_{i=1}^n$ and κ, $\lambda_i(\mathbf{A})$ is the i-th largest eigenvalues of $\mathbf{A}$ and $\lambda_r(\mathbf{A}) = \frac{1}{n-k}\big(1 - \sum_{i=1}^k \lambda_i(\mathbf{A})\big)$.

Let $\mathbf{A}_k$ be the best rank-k approximation of $\mathbf{A}$ and $L_k(\mathbf{A})$ be the matrix constructed by replacing the smaller $n-k$ eigenvalues in $\mathbf{A}$ to $\lambda_r(\mathbf{A})$. It is easy to verify that $\mathbf{S}_\alpha^k(\mathbf{A}) = \mathbf{S}_\alpha^k(\mathbf{A}_k) = \mathbf{S}_\alpha^k(L_k(\mathbf{A})) = \mathbf{S}_\alpha(L_k(\mathbf{A}))$. Definition 2 complements the smaller eigenvalues through a uniform distribution, which is the **unique method** that fulfills all axioms below (the uniqueness is discussed in the appendix[1]).

Proposition 1. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ be arbitrary normalized kernel matrices, then*

(a) $\mathbf{S}_\alpha^k(\mathbf{P}\mathbf{A}\mathbf{P}^\top) = \mathbf{S}_\alpha^k(\mathbf{A})$ *for any orthogonal matrix* $\mathbf{P}$.
(b) $\mathbf{S}_\alpha^k(p\mathbf{A})$ *is a continuous function for* $0 < p \leq 1$.
(c) $0 \leq \mathbf{S}_\alpha^k(\mathbf{A}) \leq \mathbf{S}_\alpha^k(\frac{1}{n}\mathbf{I}) = \log_2(n)$.
(d) $\mathbf{S}_\alpha^{2nk-k^2}\big(L_k(\mathbf{A}) \otimes L_k(\mathbf{B})\big) = \mathbf{S}_\alpha^k(\mathbf{A}) + \mathbf{S}_\alpha^k(\mathbf{B})$.
(e) If $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} = \mathbf{0}$ *and* $\mathbf{tr}(\mathbf{A}_k) = \mathbf{tr}(\mathbf{B}_k) = 1$, *then for* $g(x) = 2^{(1-\alpha)x}$ *and* $t \in [0, 1]$, *we have* $\mathbf{S}_\alpha^{2k}\big(t\mathbf{A} + (1-t)\mathbf{B}\big) = g^{-1}\big(tg(\mathbf{S}_\alpha^k(\mathbf{A})) + (1-t)g(\mathbf{S}_\alpha^k(\mathbf{B}))\big)$.
(f) $\mathbf{S}_\alpha^k\Big(\frac{\mathbf{A} \circ \mathbf{B}}{\mathbf{tr}(\mathbf{A} \circ \mathbf{B})}\Big) \geq \max\big(\mathbf{S}_\alpha^k(\mathbf{A}), \mathbf{S}_\alpha^k(\mathbf{B})\big)$.
(g) $\mathbf{S}_\alpha^k\Big(\frac{\mathbf{A} \circ \mathbf{B}}{\mathbf{tr}(\mathbf{A} \circ \mathbf{B})}\Big) \leq \mathbf{S}_\alpha^k(\mathbf{A}) + \mathbf{S}_\alpha^k(\mathbf{B})$.

[1] https://github.com/Gamepiaynmo/LRMI

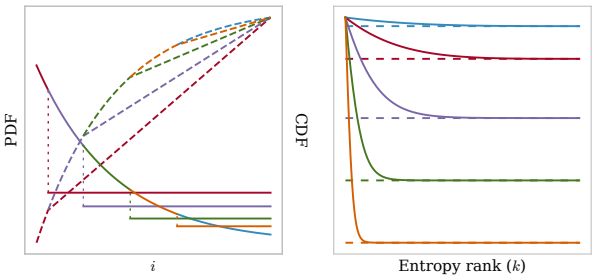

Figure 1: Left: PDF (solid) and CDF (dashed) of the altered eigenspectrum for different ranks k. Right: The convergence behavior of $\mathbf{S}_\alpha^k(\mathbf{A})$ (solid) to $\mathbf{S}_\alpha(\mathbf{A})$ (dashed) with the increase of rank k for different EDR (r).

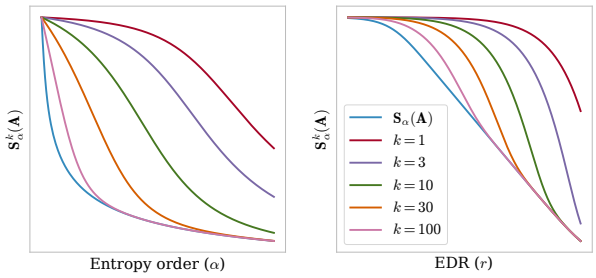

Figure 2: Left: The behavior of $\mathbf{S}_\alpha^k(\mathbf{A})$ when the entropy order α varies from 0 to 2. Right: The behavior of $\mathbf{S}_\alpha^k(\mathbf{A})$ when the EDR of $\mathbf{A}$ varies from flat to steep.

Remark 1. *Proposition 1 characterizes the basic properties of low-rank Rényi's entropy, in which (a)-(e) are the set of axioms provided by Rényi (Rényi 1961) that a function must satisfy to be a measure of information. Additionally, (f) and (g) together imply a definition of joint entropy which is also compatible with the individual entropy measures:*

$$\mathbf{S}_\alpha^k(\mathbf{A},\mathbf{B}) = \mathbf{S}_\alpha^k\left(\frac{\mathbf{A}\circ\mathbf{B}}{\mathbf{tr}(\mathbf{A}\circ\mathbf{B})}\right).$$

This further allows us to define the low-rank conditional entropy $\mathbf{S}_\alpha^k(\mathbf{A}|\mathbf{B})$ and mutual information $\mathbf{I}_\alpha^k(\mathbf{A};\mathbf{B})$, whose positiveness is guaranteed by (f) and (g) respectively:

$$\mathbf{S}_\alpha^k(\mathbf{A}|\mathbf{B}) = \mathbf{S}_\alpha^k(\mathbf{A},\mathbf{B}) - \mathbf{S}_\alpha^k(\mathbf{B}),$$
$$\mathbf{I}_\alpha^k(\mathbf{A};\mathbf{B}) = \mathbf{S}_\alpha^k(\mathbf{A}) + \mathbf{S}_\alpha^k(\mathbf{B}) - \mathbf{S}_\alpha^k(\mathbf{A},\mathbf{B}).$$

An intuitive overview of the comparative behavior between $\mathbf{S}_\alpha(\mathbf{A})$ and $\mathbf{S}_\alpha^k(\mathbf{A})$ for $n = 1000$ is reported in Figure 1 and 2, where we evaluate the impact of k, α and eigenspectrum decay rate (EDR) r respectively. The eigenvalues are initialized by $\lambda_i = e^{-ri/n}$ and then normalized. It can be observed from Figure 1 that $\mathbf{S}_\alpha^k(\mathbf{A})$ is always larger than $\mathbf{S}_\alpha(\mathbf{A})$ since the uncertainty of the latter $n-k$ outcomes are maximized. Moreover, $\mathbf{S}_\alpha^k(\mathbf{A})$ quickly converges to $\mathbf{S}_\alpha(\mathbf{A})$ with the increase of k, especially in extreme cases when the eigenspectrum of $\mathbf{A}$ is flat or steep. From Figure 2, we can see that for small k, $\mathbf{S}_\alpha^k(\mathbf{A})$ decreases slow with the increase of α when $\alpha < 1$ and fast otherwise. This behavior is the opposite when k becomes large. Furthermore, we can see that EDR directly influences the value of entropy, as a flat eigenspectrum indicates higher uncertainty and steep the opposite. As can be seen, $\mathbf{S}_\alpha^k(\mathbf{A})$ monotonically decreases with the increase of r, and decreases faster than $\mathbf{S}_\alpha(\mathbf{A})$ in a certain range which varies according to the choice of k, indicating higher sensitivity to informative distribution changes when the hyper-parameter k is selected properly.

Moreover, consider the case that the data samples $\{\mathbf{x}_i\}_{i=1}^n$ are randomly perturbed, i.e. $\mathbf{y}_i = \mathbf{x}_i + \varepsilon\mathbf{p}_i$, where $\mathbf{p}_i$ are random vectors comprised of i.i.d. entries with zero expectation and unit variance. Let $\mathbf{A}$ and $\mathbf{B}$ be kernel matrices constructed from $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ respectively, and let $\{\lambda_i\}_{i=1}^n$, $\{\mu_i\}_{i=1}^n$ be their eigenvalues. Then it satisfies that $\mu_i \approx \lambda_i + \mathbf{u}_i^\top(\mathbf{B}-\mathbf{A})\mathbf{u}_i$ (Ngo 2005), where $\mathbf{u}_i$ is the corresponding eigenvector of λ_i. When ε is small, the entries as well as the eigenvalues of $\mathbf{A}$ are nearly independently perturbed. The following theorem shows that $\mathbf{S}_\alpha^k(\mathbf{A})$ is more robust against small noises in data compared to $\mathbf{S}_\alpha(\mathbf{A})$:

Theorem 1. *Let $\{\nu_i\}_{i=1}^n$ be independent random variables with zero mean and variance $\{\sigma_i^2\}_{i=1}^n$. Let $\mathbf{A}$ and $\mathbf{B}$ be PSD matrices with eigenvalues λ_i and $\mu_i = \lambda_i + \nu_i$ respectively. If $\sum_{i=1}^k \sigma_i^2 \le \sum_{i=k+1}^n \sigma_i^2$ or $\alpha > 1$, there exists $\epsilon > 0$ such that when all $|\nu_i| \le \epsilon$, we have $\mathbf{Var}[\mathbf{IP}_\alpha^k(\mathbf{B})] \le \mathbf{Var}[\mathbf{IP}_\alpha(\mathbf{B})]$, where $\mathbf{IP}$ is the information potential (Gokcay and Principe 2000) defined as $\mathbf{IP}_\alpha(\mathbf{B}) = 2^{(1-\alpha)\mathbf{S}_\alpha(\mathbf{B})}$ and $\mathbf{IP}_\alpha^k(\mathbf{B}) = 2^{(1-\alpha)\mathbf{S}_\alpha^k(\mathbf{B})}$.*

Remark 2. *Theorem 1 indicates that $\mathbf{IP}_\alpha^k(\mathbf{B})$ enables lower variance than $\mathbf{IP}_\alpha(\mathbf{B})$ against random perturbation of the eigenvalues under mild conditions, which is easy to be satisfied since in most cases we have $k \ll n$. Combining with our discussion above, the low-rank Rényi's entropy is more sensitive to informative variations in probability distributions which will surely induce an increase or decrease in entropy, while being insensitive to uninformative perturbations caused by noises in the data samples.*

Extending to Multivariate Scenarios

Following Definition 2 and Proposition 1, the low-rank variant of multivariate Rényi's joint entropy, in virtue of the Venn diagram relation for Shannon's entropy (Yeung 1991), could be naturally derived:

Definition 3. *Let $\{\kappa_i\}_{i=1}^L : \mathcal{X}^i \times \mathcal{X}^i \mapsto \mathbb{R}$ be positive infinitely divisible kernels and $\{\mathbf{x}_i^1, \cdots, \mathbf{x}_i^L\}_{i=1}^n \subset \mathcal{X}^1 \times \cdots \times \mathcal{X}^L$, the low-rank Rényi's joint entropy is defined as:*

$$\mathbf{S}_\alpha^k(\mathbf{A}_1, \cdots, \mathbf{A}_L) = \mathbf{S}_\alpha^k\left(\frac{\mathbf{A}_1\circ\cdots\circ\mathbf{A}_L}{\mathbf{tr}(\mathbf{A}_1\circ\cdots\circ\mathbf{A}_L)}\right),$$

where $\mathbf{A}_1, \cdots, \mathbf{A}_L$ are normalized kernel matrices constructed from $\{\mathbf{x}_i^1\}_{i=1}^n, \cdots, \{\mathbf{x}_i^L\}_{i=1}^n$ respectively and $\circ$ denotes the Hadamard product.

This joint entropy definition enables further extension to multivariate conditional entropy and mutual information:

$$\mathbf{S}_\alpha^k(\mathbf{A}_1, \cdots, \mathbf{A}_k|\mathbf{B}) = \mathbf{S}_\alpha^k(\mathbf{A}_1, \cdots, \mathbf{A}_k, \mathbf{B}) - \mathbf{S}_\alpha^k(\mathbf{B}),$$
$$\mathbf{I}_\alpha^k(\mathbf{A}_1, \cdots, \mathbf{A}_k; \mathbf{B}) = \mathbf{S}_\alpha^k(\mathbf{A}_1, \cdots, \mathbf{A}_k) + \mathbf{S}_\alpha^k(\mathbf{B}) - \mathbf{S}_\alpha^k(\mathbf{A}_1, \cdots, \mathbf{A}_k, \mathbf{B}),$$

Algorithm 1: Approximation via Random Projection

Input: Integers n, $k \in [1, n/2], s \geq k$, kernel matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, order $\alpha > 0$.
Output: Approximation to $\mathbf{S}_\alpha^k(\mathbf{A})$;
Construct a random projection matrix $\mathbf{P} \in \mathbb{R}^{n\times s}$.
Calculate $\hat{\mathbf{A}} = \mathbf{AP} \in \mathbb{R}^{n\times s}$.
Calculate the largest k singular values $\hat{\lambda}_i$, $i \in [1, k]$ of $\hat{\mathbf{A}}$ through singular value decomposition.
Calculate $\hat{\lambda}_r = \frac{1}{n-k}\Big(1 - \sum_{i=1}^k \hat{\lambda}_i\Big)$.
Return: $\hat{\mathbf{S}}_\alpha^k(\mathbf{A}) = \frac{1}{1-\alpha}\log_2\Big(\sum_{i=1}^k \hat{\lambda}_i^\alpha + (n-k)\hat{\lambda}_r^\alpha\Big)$.

where $\mathbf{A}_1, \cdots, \mathbf{A}_L$ and $\mathbf{B}$ are normalized kernel matrices constructed from the variables $\{\mathbf{x}_i^1\}_{i=1}^n, \cdots, \{\mathbf{x}_i^L\}_{i=1}^n$ and the target label $\{\mathbf{y}_i\}_{i=1}^n$ respectively. Their positiveness can be guaranteed through a reduction to axiom (f) and (g). These multivariate information quantities enable much more widespread applications e.g. feature selection, dimension reduction and information-based clustering.

Approximating Low-rank Rényi's Entropy

Although only the largest eigenvalues are accessed by our entropy definition, one still needs to calculate the full eigenspectrum of the PSD matrix $\mathbf{A}$ through eigenvalue decomposition algorithms, resulting in $\mathcal{O}(n^3)$ overall time cost. To alleviate the computational burden, we design fast approximations by leveraging random projection and Lanczos iteration techniques for low-rank Rényi's entropy.

Random Projection Approach

Random projection offers a natural way to approximate the low-rank representation of kernel matrices. The core idea is to project the $n \times n$ PSD matrix $\mathbf{A}$ into a $n \times s$ subspace, and then use the largest k singular value of the projected matrix as approximations of the largest k eigenvalues, as summarized in Algorithm 1. In this way, the main computation cost is reduced to $\mathcal{O}(n^2 s)$ or even $\mathcal{O}(ns^2)$, ($s \ll n$), substantially lower than the original $\mathcal{O}(n^3)$ approach. Based on this fact, we develop efficient approximation algorithms by exploring different random projection techniques, in which the construction of $\mathbf{P}$ varies depending on the practical applications, ranging from simple but effective Gaussian distributions to advanced random orthogonal projections.

Gaussian Random Projection
As one of the most widely used random projection techniques, Gaussian random projection (GRP) admits a simple but elegant solution for eigenvalue approximation:

$$\mathbf{P} = \sqrt{n/s} \cdot \mathbf{G},$$

where the columns of $\mathbf{G} \in \mathbb{R}^{n\times s}$ are initialized by i.i.d random standard Gaussian variables and then orthogonalized. The time complexity of GRP is $\mathcal{O}(n^2 s)$.

Subsampled Randomized Hadamard Transform
SRHT (Lu et al. 2012; Tropp 2011) is a simplification of the fast Johnson-Lindenstrauss transform (Ailon and Chazelle 2009) which preserves the geometry of an entire subspace of vectors compared to GRP. In our settings, the $n \times s$ SRHT matrix is constructed by

$$\mathbf{P} = \sqrt{1/s} \cdot \mathbf{DHS},$$

where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with random $\{\pm 1\}$ entries, $\mathbf{H} \in \mathbb{R}^{n\times n}$ is a Walsh-Hadamard matrix, $\mathbf{S} \in \mathbb{R}^{n\times s}$ is a subsampling matrix whose columns are a uniformly chosen subset of the standard basis of $\mathbb{R}^n$.

Two key ingredients make SRHT an efficient approximation strategy: first, it takes only $\mathcal{O}(n^2 \min(\log(n), s))$ time complexity to calculate the projected matrix $\hat{\mathbf{A}}$; second, the orthonormality between the columns of $\mathbf{A}$ can be preserved after projection, thus is more likely to achieve lower approximation error compared to GRP.

Input-Sparsity Transform
Similar to SRHT, input-sparsity transform (IST) (Mahoney 2011; Woodruff and Zandieh 2020) utilizes the fast John-Lindenstrauss transform to reduce time complexity for least-square regression and low-rank approximation:

$$\mathbf{P} = \sqrt{n/s} \cdot \mathbf{DS},$$

where $\mathbf{D}$ and $\mathbf{S}$ are constructed in the same way as SRHT. The complexity of calculating $\hat{\mathbf{A}}$ using IST is $\mathcal{O}(\text{nnz}(\mathbf{A}))$, where nnz denotes the number of non-zero entries, resulting in a total complexity of $\mathcal{O}(\min(\text{nnz}(\mathbf{A}), ns^2))$.

Sparse Graph Sketching
The idea of using sparse graphs as sketching matrices is proposed in (Hu et al. 2021). It is shown that the generated bipartite graphs by uniformly adding edges enjoy elegant theoretical properties known as the Expander Graph or Magical Graph with high probability, and thus serve as an effective random projection strategy:

$$\mathbf{P} = \sqrt{1/p} \cdot \mathbf{G},$$

where $p \in \mathbb{N}$ is the hyper-parameter that controls the sparsity, and each column $\mathbf{g}$ of $\mathbf{G}$ is constructed independently by uniformly sampling $c \subset [n]$ with $|c| = p$, and then setting $\mathbf{g}_i = \{\pm 1\}$ randomly for $i \in c$ and $\mathbf{g}_i = 0$ for $i \notin c$. Similar to IST, sparse graph sketching (SGS) also utilizes the sparsity of input matrices and achieves $\mathcal{O}(\text{nnz}(\mathbf{A})p)$ computational complexity to calculate the projected matrix.

Theoretical Results
Next, we provide the main theorem on characterizing the quality-of-approximation for low-rank Rényi's entropy:

Theorem 2. *Let* $\mathbf{A}$ *be positive definite and*

$$s = \begin{cases} \mathcal{O}(k + \log(1/\delta)/\epsilon_0^2), & \textit{for GRP} \\ \mathcal{O}((k + \log n)\log k/\epsilon_0^2), & \textit{for SRHT} \\ \mathcal{O}(k^2/\epsilon_0^2), & \textit{for IST} \\ \mathcal{O}(k\log(k/\delta\epsilon_0)/\epsilon_0^2), & \textit{for SGS} \end{cases}$$

$$p = \mathcal{O}(\log(k/\delta\epsilon_0)/\epsilon_0), \qquad \textit{for SGS}$$

where $\epsilon_0 = \epsilon\lambda_k\lambda_r$*, then for* $k \leq n/2$*, with confidence at least* $1 - \delta$*, the output of Algorithm 1 satisfies*

$$|\lambda_i^2 - \hat{\lambda}_i^2| \leq \epsilon$$

Algorithm 2: Approximation via Lanczos Iteration

Input: Integers n, $k \in [1, n/2], s \geq k$, kernel matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, order $\alpha > 0$, initial vector $\mathbf{q}$.
Output: Approximation to $\mathbf{S}_\alpha^k(G)$.
Set $\mathbf{q}_0 = 0, \beta_0 = 0, \mathbf{q}_1 = \mathbf{q}/\|\mathbf{q}\|$.
for $j = 1, 2, \cdots, s$ **do**
  $\hat{\mathbf{q}}_{j+1} = \mathbf{A}\mathbf{q}_j - \beta_{j-1}\mathbf{q}_{j-1}, \gamma_j = \langle \hat{\mathbf{q}}_{j+1}, \mathbf{q}_j \rangle$.
  $\hat{\mathbf{q}}_{j+1} = \hat{\mathbf{q}}_{j+1} - \gamma_j \mathbf{q}_j$.
  Orthogonalize $\hat{\mathbf{q}}_{j+1}$ against $\mathbf{q}_1, \cdots, \mathbf{q}_{j-1}$.
  $\beta_j = \|\hat{\mathbf{q}}_{j+1}\|, \mathbf{q}_{j+1} = \hat{\mathbf{q}}_{j+1}/\beta_j$.
end for
Calculate the largest k eigenvalues $\hat{\lambda}_i$, $i \in [1, k]$ of

$$\mathbf{T} = \begin{bmatrix} \gamma_1 & \beta_1 & & 0 \\ \beta_1 & \gamma_2 & & \\ & & \ddots & \beta_{s-1} \\ 0 & & \beta_{s-1} & \gamma_s \end{bmatrix}.$$

Calculate $\hat{\lambda}_r = \frac{1}{n-k}\Big(1 - \sum_{i=1}^k \hat{\lambda}_i\Big)$.
Return: $\hat{\mathbf{S}}_\alpha^k(\mathbf{A}) = \frac{1}{1-\alpha}\log_2\Big(\sum_{i=1}^k \hat{\lambda}_i^\alpha + (n-k)\hat{\lambda}_r^\alpha\Big)$.

for all $i \in [1, k]$ eigenvalues of $\mathbf{A}$ and

$$|\mathbf{S}_\alpha^k(\mathbf{A}) - \hat{\mathbf{S}}_\alpha^k(\mathbf{A})| \leq |\tfrac{\alpha}{1-\alpha}\log_2(1-\epsilon)|.$$

Remark 3. *Theorem 2 provides the accuracy guarantees for low-rank Rényi's entropy approximation via random projections. It can be observed that the approximation error grows with the increase of α when α is small. Note that although the error bound is additive in nature, it can be further reduced to a relative error bound under mild condition $\mathbf{S}_\alpha^k(G) \geq \sqrt{\epsilon}$. In general, Theorem 2 requires $s = \mathcal{O}(k + 1/\epsilon^2)$ to achieve $1 \pm \epsilon$ absolute accuracy, which is consistent with the complexity results of least squares and low rank approximations (Mahoney 2011).*

Lanczos Iteration Approach

Besides random projection, the Lanczos algorithm is also widely adopted to find the k extreme (largest or smallest in magnitude) eigenvalues and the corresponding eigenvectors of an $n \times n$ Hermitian matrix $\mathbf{A}$. Given an initial vector $\mathbf{q}$, the Lanczos algorithm utilizes the Krylov subspace spanned by $\{\mathbf{q}, \mathbf{A}\mathbf{q}, \cdots, \mathbf{A}^s\mathbf{q}\}$ to construct an tridiagonalization of $\mathbf{A}$ whose eigenvalues converge to those of $\mathbf{A}$ along with the increase of s, and are satisfactorily accurate even for $s \ll n$. As shown in Algorithm 2, the main computation cost is the $\mathcal{O}(n^2 s)$ matrix-vector multiplications in the Lanczos process, which could be further reduced to $\mathcal{O}(\mathrm{nnz}(\mathbf{A})s)$ when $\mathbf{A}$ is sparse. The computational cost of reorthogonalization can be further alleviated by explicit or implicit restarting Lanczos methods. The following theorem establishes the accuracy guarantee of Algorithm 2:

Theorem 3. *Let $\mathbf{A}$ be positive definite, $\mathbf{q}$ be the initial vector, $\{\boldsymbol{\phi}_i\}_{i=1}^k$ be the corresponding eigenvectors and*

$$s = \Big\lceil k + \tfrac{1}{2\log R}\log\Big(\tfrac{4\theta^2 K^2 \lambda_1}{\epsilon\lambda_r}\Big)\Big\rceil,$$

where

$$R = \gamma + \sqrt{\gamma^2 - 1}, \quad \gamma = 1 + 2\min_{i\in[1,k]} \tfrac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_n},$$

$$\theta = \max_{i\in[1,k]} \tan\langle \boldsymbol{\phi}_i, \mathbf{q}\rangle, \quad K = \prod_{j=1}^{k-1} \tfrac{\hat{\lambda}_j - \lambda_n}{\hat{\lambda}_j - \lambda_k},$$

then for $k \leq n/2$, the output of Algorithm 2 satisfies

$$0 \leq \lambda_i - \hat{\lambda}_i \leq \epsilon\lambda_i$$

for all $i \in [1, k]$ eigenvalues of $\mathbf{A}$ and

$$|\mathbf{S}_\alpha^k(\mathbf{A}) - \hat{\mathbf{S}}_\alpha^k(\mathbf{A})| \leq |\tfrac{\alpha}{1-\alpha}\log_2(1-\epsilon)|.$$

Remark 4. *Theorem 3 provides the accuracy guarantee for the Lanczos algorithm. The relationship between approximation error and α is similar to those in Theorem 2. Algorithm 2 achieves a much faster convergence rate compared to Algorithm 1 while achieving the same level of absolute precision. When ϵ is small, R, θ and K can be regarded as constants that depends only on the eigenspectrum of $\mathbf{A}$ and the initial vector $\mathbf{q}$, so that $s = \mathcal{O}(k + \log(1/\epsilon))$ is enough to guarantee a $1 \pm \epsilon$ accuracy. In practice, $\mathbf{q}$ is suggested to be generated by random Gaussian in order to avoid a large θ with high probability (Urschel 2021).*

Experimental Results

In this section, we evaluate the proposed low-rank Rényi's entropy and the approximation algorithms under large-scale experiments. Our experiments are conducted on an Intel i7-10700 (2.90GHz) CPU and an RTX 2080Ti GPU with 64GB of RAM. The algorithms are implemented in C++ with the Eigen library and in Python with the Pytorch library.

Simulation Studies

We first test the robustness of $\mathbf{S}_\alpha^k(\mathbf{A})$ against noises in the data. As indicated by Theorem 1, low-rank Rényi's entropy achieves lower variance under mild conditions in terms of the information potential. We consider the case that the input data points are randomly perturbed, i.e. $\mathbf{y}_i = \mathbf{x}_i + \varepsilon\mathbf{p}_i$ for $i \in [1, n]$, where $\mathbf{p}_i$ is comprised of i.i.d. random variables. Let $\{\lambda_i\}_{i=1}^n$, $\{\mu_i\}_{i=1}^n$ denote the eigenvalues of normalized kernel matrices constructed from $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ respectively. We test the following noise distributions: Standard Gaussian $N(0, 1)$, Uniform $U(-\sqrt{3}, \sqrt{3})$, Student-t $t(3)/\sqrt{3}$ and Rademacher $\{\pm 1\}$ with $n = 100$ (detailed settings are given in the appendix). The examples of variation in eigenvalues ($\mu_i - \lambda_i$) and the standard deviation (multiplied by n) of entropy values after 100 trials are reported in Figure 3. It verifies our analysis that when ε is small, the eigenvalues μ_i are nearly independently perturbed. Moreover, our low-rank definition achieves lower variance than matrix-based Rényi's entropy under different choices of α, in which smaller k corresponds to higher robustness.

Real Data Examples

In this section, we demonstrate the great potential of applying our low-rank Rényi's entropy functional and its multivariate extensions in two representative real-world information-related applications, which utilize the mutual information (information bottleneck) and multivariate mutual information (feature selection) respectively.

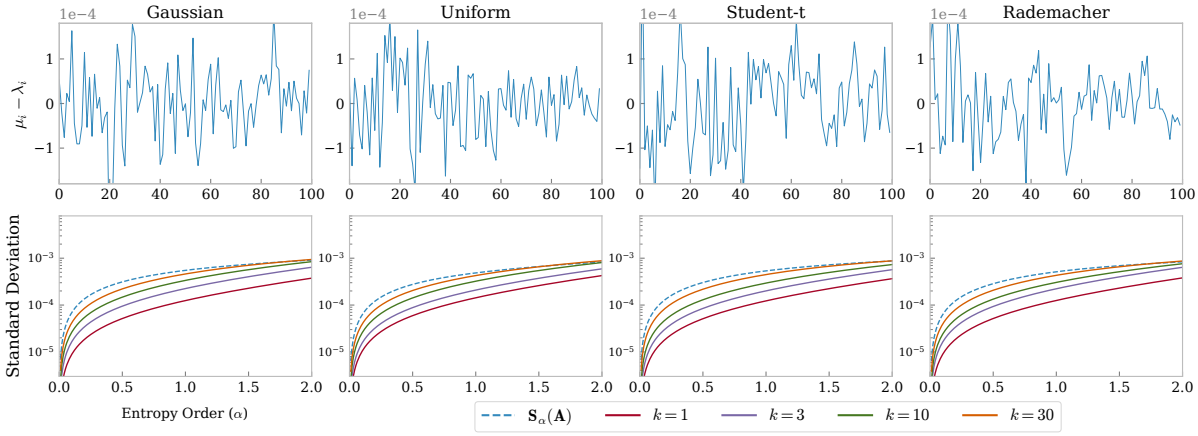

Figure 3: Upper: perturbation of the eigenvalues, i.e. $\mu_i - \lambda_i$. Lower: standard deviation of matrix-based Rényi's entropy and low-rank Rényi's entropy against random perturbations of the data samples for different values of α.

Objective	Accuracy (%)	Training Time (minutes)
CE	92.64 ± 0.03	- / 80
VIB	94.08 ± 0.02	4 / 84
NIB	94.01 ± 0.04	7 / 87
MRIB	94.13 ± 0.04	46 / 126
LRIB	$\mathbf{94.16 \pm 0.09}$	15 / 95

Table 1: Classification accuracy and training time of different IB objectives. Left is the time spent on IB calculation and right is the total training time.

Application to Information Bottleneck

The Information Bottleneck (IB) methods recently achieve great success in compressing redundant or irrelevant information in the inputs and preventing overfitting in deep neural networks. Formally, given network input $\mathbf{X}$ and target label $\mathbf{Y}$, the IB approach tries to extract a compressed intermediate representation $\mathbf{Z}$ from $\mathbf{X}$ that maintains minimal yet meaningful information to predict the task $\mathbf{Y}$ by optimizing the following IB Lagrangian:

$$\mathcal{L}_{\text{IB}} = \mathbf{I}(\mathbf{Y}, \mathbf{Z}) - \beta \cdot \mathbf{I}(\mathbf{X}, \mathbf{Z}),$$

where β is the hyper-parameter that balances the trade-off between **sufficiency** (predictive performance of $\mathbf{Z}$ on task $\mathbf{Y}$, quantified by $\mathbf{I}(\mathbf{Y}, \mathbf{Z})$) and **minimality** (the complexity of $\mathbf{Z}$, quantified by $\mathbf{I}(\mathbf{X}, \mathbf{Z})$). In practice, optimizing $\mathbf{I}(\mathbf{Y}, \mathbf{Z})$ is equivalent to the cross-entropy (CE) loss for classification tasks, so our target remains to optimize the latter term $\mathbf{I}(\mathbf{X}, \mathbf{Z})$. However, mutual information estimation is extremely hard or even intractable for high-dimension distributions, which is usually the case in deep learning. To address this issue, there have been efforts on using variational approximations to optimize a lower bound of $\mathbf{I}(\mathbf{X}, \mathbf{Z})$, e.g. Variational IB (VIB) (Alemi et al. 2017) and Nonlinear IB (NIB) (Kolchinsky, Tracey, and Wolpert 2019). We show that with low-rank Rényi's entropy, $\mathbf{I}(\mathbf{X}, \mathbf{Z})$ can be directly optimized by approximating the largest k eigenvalues of the kernel matrix $\mathbf{A}$ constructed by $\mathbf{X}$ and $\mathbf{Z}$. Recall that the Lanczos method constructs an approximation $\mathbf{A} \approx \mathbf{Q}\mathbf{T}\mathbf{Q}^\top$, where $\mathbf{Q} \in \mathbb{R}^{n \times s}$ has orthogonal columns and $\mathbf{T} \in \mathbb{R}^{s \times s}$ is tridiagonal, we have $\hat{\lambda}_i = \lambda_i(\mathbf{Q}^\top \mathbf{A} \mathbf{Q})$ for all $i \in [1, s]$. Let $\sum_{i=1}^{s} \hat{\lambda}_i \mathbf{u}_i \mathbf{u}_i^\top$ be the eigenvalue decomposition of $\mathbf{Q}^\top \mathbf{A} \mathbf{Q}$, we can approximate the gradient of $\mathbf{S}_\alpha^k(\mathbf{A})$ as:

$$\frac{\partial \mathbf{S}_\alpha^k(\mathbf{A})}{\partial \mathbf{A}} \approx \sum_{i=1}^{k} \frac{\partial \hat{\mathbf{S}}_\alpha^k(\mathbf{A})}{\partial \hat{\lambda}_i} \cdot \mathbf{Q}\mathbf{u}_i\mathbf{u}_i^\top\mathbf{Q}^\top.$$

In this experiment, we test the performance of matrix-based Rényi's IB (MRIB) (Yu, Yu, and Principe 2021) and our low-rank variant (LRIB) with variational approximation-based objectives using VGG16 as the backbone and CIFAR10 as the classification task. All models are trained for 300 epochs with 100 batch size and 0.1 initial learning rate which is divided by 10 every 100 epochs. Following the settings in (Yu, Yu, and Principe 2021), we select $\alpha = 1.01$, $\beta = 0.01$, $k = 10$ and $s = 20$. The final results are reported in Table 1. It can be seen that the matrix-based approaches MRIB and LRIB outperform other methods, while our LRIB achieves the highest performance with significantly less training time.

Application to Feature Selection

In practical regression or classification machine learning tasks, many features can be completely irrelevant to the learning target or redundant in the context of others. Given a set of features $\mathbf{S} = \{\mathbf{X}_1, \cdots, \mathbf{X}_L\}$ and the target label $\mathbf{Y}$, we aim to find a subset $\mathbf{S}_{sub} \subset \mathbf{S}$ which leverage the expressiveness and the complexity simultaneously. In the field of information theoretic learning, this target is equivalent to maximizing the multivariate mutual information $\mathbf{I}(\mathbf{S}_{sub}; \mathbf{Y})$, which is computationally prohibitive due to the curse of high dimensionality. As a result, there have been tremendous efforts on approximation techniques that retain only the first or second order interactions and build mutual information estimators upon low-dimensional probability distributions, including Mutual Information-based Feature Selection (MIFS) (R. Battiti 1994), First-Order Util-

Method	Criterion	Breast	Semeion	Madelon	Krvskp	Spambase	Waveform	Optdigits	Statlog	Average
MIFS	$\mathbf{I}(\mathbf{X}_{i_l};\mathbf{Y})-\beta\sum_{j=1}^{l-1}\mathbf{I}(\mathbf{X}_{i_l};\mathbf{X}_{i_j})$	4.8	2.5	6.6	3.8	7.3	6.4	4.5	3.8	4.96
FOU	$\mathbf{I}(\mathbf{X}_{i_l};\mathbf{Y})-\sum_{j=1}^{l-1}[\mathbf{I}(\mathbf{X}_{i_l};\mathbf{X}_{i_j})-\mathbf{I}(\mathbf{X}_{i_l};\mathbf{X}_{i_j}\|\mathbf{Y})]$	5.2	2.5	5.6	1.9	6.2	5.7	4.8	5.7	4.70
MRMR	$\mathbf{I}(\mathbf{X}_{i_l};\mathbf{Y})-\frac{1}{l-1}\sum_{j=1}^{l-1}\mathbf{I}(\mathbf{X}_{i_l};\mathbf{X}_{i_j})$	**2.3**	4.7	6.7	3.7	5.6	3.6	4.8	4.0	4.43
JMI	$\sum_{j=1}^{l-1}\mathbf{I}(\{\mathbf{X}_{i_l},\mathbf{X}_{i_j}\};\mathbf{Y})$	5.1	5.2	3.0	3.7	4.2	2.3	3.8	3.5	3.85
CMIM	$\min_{j=1}^{l-1}\mathbf{I}(\mathbf{X}_{i_l};\mathbf{Y}\|\mathbf{X}_{i_j})$	3.0	2.7	4.5	3.6	3.2	4.7	2.6	5.6	3.74
DISR	$\sum_{j=1}^{l-1}\mathbf{I}(\{\mathbf{X}_{i_l},\mathbf{X}_{i_j}\};\mathbf{Y})/\mathbf{H}(\mathbf{X}_{i_l},\mathbf{X}_{i_j},\mathbf{Y})$	7.3	5.7	4.0	3.2	3.9	2.3	6.9	5.8	4.89
MRMI	$\mathbf{I}_\alpha(\{\mathbf{X}_{i_1},\mathbf{X}_{i_2},\cdots,\mathbf{X}_{i_l}\};\mathbf{Y})$	2.6	1.8	1.2	1.7	**1.5**	1.8	**1.3**	**2.0**	1.74
LRMI	$\mathbf{I}^k_\alpha(\{\mathbf{X}_{i_1},\mathbf{X}_{i_2},\cdots,\mathbf{X}_{i_l}\};\mathbf{Y})$	2.6	**1.4**	**1.1**	**1.6**	**1.5**	**1.5**	**1.3**	2.1	**1.64**

Table 2: Information theoretic feature selection methods and their average rank over different number of features in each dataset. The first and second best performances are marked as bold and underlined respectively.

ity (FOU) (Brown 2009), Maximum-Relevance Minimum-Redundancy (MRMR) (Peng, Long, and Ding 2005), Joint Mutual Information (JMI) (Yang and Moody 1999), Conditional Mutual Information Maximization (CMIM) (Fleuret 2004) and Double Input Symmetrical Relevance (DISR) (Meyer and Bontempi 2006) which achieve state-of-the-art performance in information-based feature selection tasks.

We evaluate the performance of matrix-based Rényi's mutual information (MRMI) and our low-rank variant (LRMI) with these methods on 8 widely-used classification datasets as shown in Table 3, which is chosen to cover a broad variety of instance-feature ratios, number of classes and discreteness. Notice that non-Rényi methods can only handle discrete features, so we discretize them into 5 bins under equal-width criterion as adopted in (Brown et al. 2012). In this experiment, we choose the Support Vector Machine (SVM) algorithm with RBF kernel ($\sigma = 1$) as the classifier for continuous datasets and a 3-NN classifier for discrete datasets. Following the settings of (Yu et al. 2019), we select $\alpha \in \{0.6, 1.01, 2\}$, $k \in \{100, 200, 400\}$ via cross-validation, $s = k + 50$ and use the Gaussian kernel of width $\sigma = 1$ for matrix-based entropy measures. Considering that it is NP-hard to evaluate each subset of $\mathbf{S}$, we adopt a greedy strategy to incrementally select 10 features that maximize our target $\mathbf{I}(\mathbf{S}_{sub};\mathbf{Y})$. That is, in each step, we fix the current subset $\mathbf{S}_{sub} = \{\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_{l-1}}\}$ and add a new feature $\mathbf{X}_{i_l} \in \mathbf{S}/\mathbf{S}_{sub}$ to $\mathbf{S}_{sub}$. The average rank of each method across different number of features and the running time of MRMI and LRMI are reported in Table 2 and Table 3.

As we can see, both MRMI and LRMI significantly outperform other Shannon entropy based methods. Compared to MRMI, LRMI achieves 6 to 27 times speedup, 15 times on average via Lanczos approximation. Furthermore, LRMI outperforms MRMI on 4 datasets in our test benchmark, which verifies our theoretical analysis that low-rank Rényi's entropy enables higher robustness against noises in the data. This demonstrates the great potential of our low-rank Rényi's entropy on information-related tasks.

Conclusion

In this paper, we investigate an alternative entropy measure built upon the largest k eigenvalues of the data kernel matrix. Compared to the original matrix-based Rényi's entropy, our definition enables higher robustness to noises in the data and sensitivity to informative changes in eigenspectrum distribution with a proper choice of hyper-parameter k. Moreover, low-rank Rényi's entropy can be efficiently approximated with $\mathcal{O}(ns^2)$ random projection and $\mathcal{O}(n^2 s)$ Lanczos iteration techniques, substantially lower than the $\mathcal{O}(n^3)$ complexity required to compute matrix-based Rényi's entropy. We conduct large-scale simulation and real-world experiments on information bottleneck and feature selection tasks to validate the effectiveness of low-rank Rényi's entropy, demonstrating elegant performance with significant improvements in computational efficiency.

Dataset	#I	#F	#C	Discrete	Time	Speedup
Breast	569	30	2	No	0.31 / 0.25	1.2
Semeion	1593	256	10	Yes	56 / 44	1.3
Madelon	2600	500	2	Yes	570 / 39	14.4
Krvskp	3196	36	2	Yes	71 / 11	6.6
Spambase	4601	56	2	No	353 / 13	27.2
Waveform	5000	40	3	No	318 / 14	22.5
Optdigits	5620	64	10	Yes	750 / 41	18.4
Statlog	6435	36	6	Yes	600 / 23	25.7

Table 3: Number of instances (#I), features (#F), classes (#C) and discreteness of classification datasets used in feature selection experiments, running time comparison (minutes) of MRMI (left) and LRMI (right), and speedup ratios.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2021ZD0110700), National Natural Science Foundation of China (62106191, 12071166, 62192781, 61721002), the Research Council of Norway (RCN) under grant 309439, Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology and Project of Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for The Belt and Road Training in MOOC China).

References

Ailon, N.; and Chazelle, B. 2009. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1): 302–322.

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep variational information bottleneck. *ICLR*, 1–19.

Álvarez-Meza, A. M.; Lee, J. A.; Verleysen, M.; and Castellanos-Dominguez, G. 2017. Kernel-based dimensionality reduction using Rényi's α-entropy measures of similarity. *Neurocomputing*, 222: 36–46.

Bhatia, R. 2006. Infinitely divisible matrices. *The American Mathematical Monthly*, 113(3): 221–235.

Brockmeier, A. J.; Mu, T.; Ananiadou, S.; and Goulermas, J. Y. 2017. Quantifying the informativeness of similarity measurements. *JMLR*, 18: 1–61.

Brown, G. 2009. A new perspective for information theoretic feature selection. *Journal of Machine Learning Research*, 5(1): 49–56.

Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *JMLR*, 13: 27–66.

Dong, Y.; Gong, T.; Yu, S.; and Li, C. 2022. Optimal Randomized Approximations for Matrix based Renyi's Entropy. *arXiv preprint arXiv:2205.07426*.

Fan, J.; and Li, R. 2006. Statistical challenges with high dimensionality. In *Proceedings of the international Congress of Mathematicians*.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(9).

Gokcay, E.; and Principe, J. C. 2000. A new clustering evaluation function using Renyi's information potential. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 6, 3490–3493. IEEE.

Gong, T.; Dong, Y.; Yu, S.; Chen, H.; Dong, B.; Li, C.; and Zheng, Q. 2021. Computationally Efficient Approximations for Matrix-based Renyi's Entropy. *arXiv preprint arXiv:2112.13720*.

Hu, D.; Ubaru, S.; Gittens, A.; Clarkson, K. L.; Horesh, L.; and Kalantzis, V. 2021. Sparse graph based sketching for fast numerical linear algebra. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3255–3259. IEEE.

Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2019. Nonlinear information bottleneck. *Entropy*, 21(12): 1181.

Konig, R.; Renner, R.; and Schaffner, C. 2009. The operational meaning of min-and max-entropy. *IEEE Transactions on Information theory*, 55(9): 4337–4347.

Lu, Y.; Dhillon, P. S.; Foster, D. P.; et al. 2012. Faster Ridge Regression via the Subsampled Randomized Hadamard Transform. In *NIPS*.

Mahoney, M. W. 2011. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2).

Mahoney, M. W.; and Drineas, P. 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3): 697–702.

Meyer, P. E.; and Bontempi, G. 2006. On the use of variable complementarity for feature selection in cancer classification. In *Workshops on applications of evolutionary computation*, 91–102. Springer.

Miles, R.; Rodríguez, A. L.; and Mikolajczyk, K. 2021. Information Theoretic Representation Distillation. *arXiv preprint arXiv:2112.00459*.

Ngo, K. V. 2005. An approach of eigenvalue perturbation theory. *Applied Numerical Analysis & Computational Mathematics*, 2(1): 108–125.

Ohya, M.; and Petz, D. 2004. *Quantum entropy and its use*. Springer Science & Business Media.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8): 1226–1238.

R. Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4)(4): 537–550.

Rényi, A. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 547–561. University of California Press.

Sanchez Giraldo, L. G.; Rao, M.; and Principe, J. C. 2014. Measures of entropy from data using infinitely divisible kernels. *IEEE TIT*, 61(1): 535–548.

Sarvani, C.; Ghorai, M.; Dubey, S. R.; and Basha, S. S. 2021. HRel: Filter pruning based on high relevance between activation maps and class labels. *Neural Networks*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Tropp, J. A. 2011. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02): 115–126.

Urschel, J. C. 2021. Uniform error estimates for the Lanczos method. *SIAM Journal on Matrix Analysis and Applications*, 42(3): 1423–1450.

Wan, F.; Wei, P.; Jiao, J.; Han, Z.; and Ye, Q. 2018. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1297–1306.

Watkins, D. S. 2008. The QR algorithm revisited. *SIAM review*, 50(1): 133–145.

Woodruff, D.; and Zandieh, A. 2020. Near input sparsity time kernel embeddings via adaptive sampling. In *ICML*, 10324–10333.

Yang, H.; and Moody, J. 1999. Data visualization and feature selection: New algorithms for nongaussian data. *Advances in neural information processing systems*, 12.

Yeung, R. W. 1991. A new outlook on Shannon's information measures. *IEEE transactions on information theory*, 37(3): 466–474.

Yu, S.; Sanchez Giraldo, L. G.; Jenssen, R.; and Principe, J. C. 2019. Multivariate Extension of Matrix-Based Rényi alpha-Order Entropy Functional. *IEEE TPAMI*, 42(11): 2960–2966.

Yu, X.; Yu, S.; and Principe, J. C. 2021. Deep Deterministic Information Bottleneck with Matrix-Based Entropy Functional. *ICASSP*, 3160–3164.