Technical Appendix: Robust and Fast Measure of Information via Low-rank Representation

Supplementary Experimental Results

Parameter Settings of Robustness Experiment

In the first simulation study, we set $n = 100$, $d = 400$, $\varepsilon = 1/n = 0.01$ and use the linear kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$ to generate the kernel matrices. The data samples $\{\mathbf{x}_i\}_{i=1}^n$ are generated by i.i.d Gaussian distribution $N(0, 0.1^2)$. The noise distributions $\mathcal{E}$ are selected following the criterion that $\mathbf{E}[\mathcal{E}] = 0$ and $\mathbf{Var}[\mathcal{E}] = 1$. It can be seen that we get similar results under different noise settings, which further verifies our analysis that when ε is small, the variance of $\mathbf{S}_\alpha^k(\mathbf{A})$ mainly depends on the variance of random perturbations of data samples.

Additional Results of Approximation Algorithms

Additionally, we evaluate the impact of α and c on approximation accuracy. We keep the previous $n = 8192$ parameter settings and set $k = 64$. The results of MRE curves with $c = 1.0$ and varying α are reported in Figure 1. For SGS, we set the sparsity hyper-parameter $p = 2$. For Lanczos, we randomly select the initial vector $\mathbf{q}$ from standard Gaussian. It can be seen that GRP achieves the lowest MRE amongst all random projection algorithms. When α is small, all MRE curves exhibit similar behavior and grow with the increase of α. This behavior starts to differ when α gets larger. For random projection algorithms, MRE keeps at the same level; for Lanczos algorithm, MRE starts to decrease when $\alpha > 2$. Recall that the larger eigenvalues of $\mathbf{A}$ take the main role in the calculation of $\mathbf{S}_\alpha^k(\mathbf{A})$ for large α, this phenomenon indicates that Lanczos algorithm achieves higher precision for larger eigenvalues than smaller ones (as shown in our proof, it requires only $\mathcal{O}(i + \log(1/\epsilon))$ steps to approximate λ_i to relative error $1 \pm \epsilon$), while random projection achieves similar level of precision for all of the k eigenvalues.

We then evaluate the impact of EDR (c) on approximation accuracy. In Figure 2, we report the MRE curves for $\alpha = 2$ while c varies from 0 (flat) to 2 (steep). It is interesting that the behavior of the two types of methods is entirely different. For random projections, the MRE curves grow slowly (GRP) or keep unchanged (SRHT, IST and SGS) when c is small, and start to increase at a constant rate when c gets large. This is because $\mathbf{S}_\alpha^k(\mathbf{A})$ is decreasing along with the increase of c, whose slope is low at first and high when c gets large (see Figure 2). This results in the slow to fast increasing behavior in relative error since the absolute error is upper bounded (Theorem 2). For Lanczos method, the ratio λ_1/λ_r in Theorem 3 increases fast when c is small, resulting in the increase of MRE. When c gets large, λ_1 gradually reaches its upper bound $\lambda_1 \leq 1$, while the intervals between adjacent eigenvalues also increase and result in a higher R (Theorem 3). Moreover, recall that Lanczos algorithm approximates larger eigenvalues to higher precision, these reasons together explain the increase and then decrease MRE of the Lanczos approach.

Next, we conduct large-scale experiments to evaluate the approximation algorithms. The kernel matrices are generated by $\mathbf{A} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top$, where $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$ is a random orthogonal matrix, $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ is a diagonal matrix such that $\mathbf{\Sigma}_{ii} = i^{-c}$ for $i \in [1, n]$, and c is a constant that controls the EDR. We set size of the kernel matrix $n = 8192$. For random projection methods, we make s vary from 100 to 1000; while for Lanczos algorithm, s varies from 64 to 110. The mean relative error (MRE) and $\pm\frac{1}{4}$ standard deviation (SD) are reported in Figure 3 for each test after 100 trials with $\alpha = 1.5$ and $c \in \{1.5, 1.0, 0.5\}$, which correspond to high, medium and low EDR respectively. For comparison, the trivial eigenvalue decomposition approach requires 134 seconds. It can be seen that all random projection methods yield similar approximation accuracy, in which GRP achieves slightly lower MRE when $c = 0.5$ while IST & SGS bring the highest speedup. The Lanczos method achieves the highest accuracy with significantly lower s values but requires longer running time. Generally, we recommend IST or SGS for medium precision approximation, and Lanczos when high precision is required. These methods achieve more than 25 times speedup compared to the trivial approach for an 8192×8192 kernel matrix.

Additional Results of Feature Selection

The hyper-parameter selection result of α and k for matrix-based Rényi's entropy and low-rank Rényi's entropy in feature selection experiment via cross-validation are shown in Table 1. As can be seen, $k = 100$ is already suitable for most circumstances. We perform a Nemenyi's post-hoc test (Demvsar 2006) to give the significant level, in which the confidence that method i significantly outperforms method

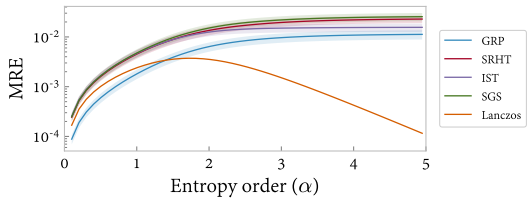

Figure 1: α versus MRE curves for entropy approximation.

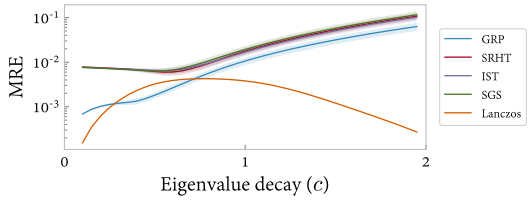

Figure 2: c versus MRE curves for entropy approximation.

j is calculated as:

$$p_{ij} = \mathbf{\Phi}\left((R_j - R_i) \Big/ \sqrt{\frac{M(M+1)}{6N}}\right),$$

where $\mathbf{\Phi}$ is the CDF of standard normal distribution, R_i is the average rank of method i, M is the number of methods and N is the number of datasets. For our case, we have $M = 8$, $N = 8$ and the value of R_i are given in the last column of table 2. The confidence level of different methods is shown in Figure 4. It can be seen that under significance level $p = 0.05$, LRMI significantly outperforms all Shannon's entropy-based methods, while the confidence of MRMI outperforming CMIM is not significant enough. In Figure 5, we report the classification accuracy achieved by different feature selection methods for the first 10 features. It can be seen that classification error tends to stabilize after selecting the 10 most informative features.

Dataset	α	k
Breast	2.0	100
Semeion	1.01	400
Madelon	2.0	100
Krvskp	1.01	200
Spambase	2.0	100
Waveform	2.0	100
Optdigits	1.01	200
Statlog	0.6	100

Table 1: Hyper-parameter selection results of α and k in feature selection experiment.

Proof of Main Results

Proof of Proposition 1

Proof. For (a): Let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigenvalue decomposition of $\mathbf{A}$, then $\mathbf{PU}$ is a unitary matrix and $\lambda_i(\mathbf{A}) = \lambda_i(\mathbf{PAP}^\top)$ for all $i \in [1, k]$.

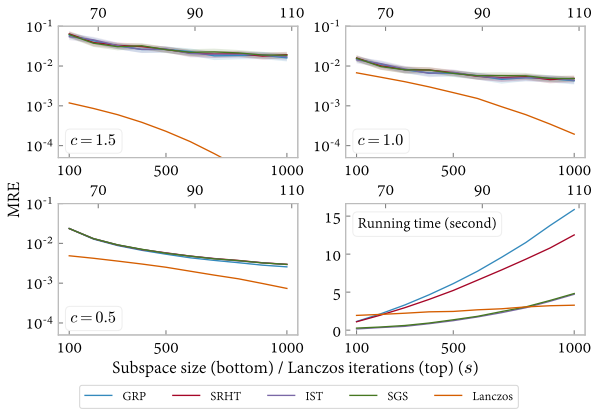

Figure 3: s versus MRE curves for entropy approximation. The first three sub-figure correspond to different c values, while the last sub-figure show the running time.

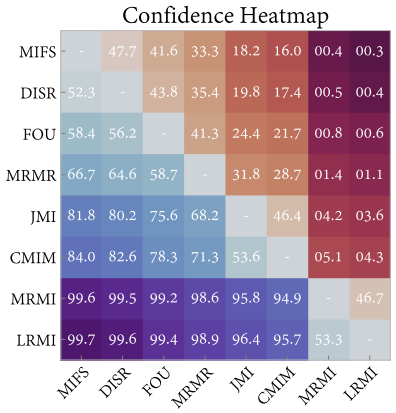

Figure 4: Confidence of significant outperforming (%) for different feature selection methods.

For (b): When $p > 0$, $\mathbf{tr}((pL_k(\mathbf{A}))^\alpha) = p^\alpha \cdot \mathbf{tr}(L_k^\alpha(\mathbf{A})) > 0$, then (b) follows by the continuity of the logarithm function.

For (c): Notice that $\mathbf{S}_\alpha^k(\mathbf{A}) = \mathbf{S}_\alpha(L_k(\mathbf{A}))$, where $\mathbf{tr}(L_k(\mathbf{A})) = 1$ and $\lambda_i(L_k(\mathbf{A})) \in [0, 1]$ for all $i \in [1, n]$. Then we have $\mathbf{tr}(L_k^\alpha(\mathbf{A})) \geq 1$ when $\alpha \in (0, 1)$ and $\mathbf{tr}(L_k^\alpha(\mathbf{A})) \leq 1$ when $\alpha > 1$, which further implies that $\mathbf{S}_\alpha^k(\mathbf{A}) \geq 0$.

Let $f(x) = x^\alpha$, it is obvious that f is concave when $\alpha \in (0, 1)$ and convex when $\alpha > 1$. Then by Jensen's inequality, $\mathbf{tr}(f(L_k(\mathbf{A}))) \leq \mathbf{tr}(f(\frac{1}{n}I))$ when $\alpha \in (0, 1)$ and otherwise the opposite, which further implies that $\mathbf{S}_\alpha^k(\mathbf{A}) \leq \mathbf{S}_\alpha^k(\frac{1}{n}I)$.

Moreover, it is straightforward to show that $\mathbf{S}_\alpha^k(\frac{1}{n}I) = \mathbf{S}_\alpha(\frac{1}{n}I) = \log_2(n)$.

For (d): From Proposition 4.1 in (Sanchez Giraldo, Rao, and Principe 2014) we have that $\mathbf{S}_\alpha(\mathbf{A} \otimes \mathbf{B}) = \mathbf{S}_\alpha(\mathbf{A}) + \mathbf{S}_\alpha(\mathbf{B})$, therefore $\mathbf{S}_\alpha(L_k(\mathbf{A}) \otimes L_k(\mathbf{B})) = \mathbf{S}_\alpha(L_k(\mathbf{A})) + \mathbf{S}_\alpha(L_k(\mathbf{B}))$. Notice that the smaller $(n-k)^2$ eigenvalues

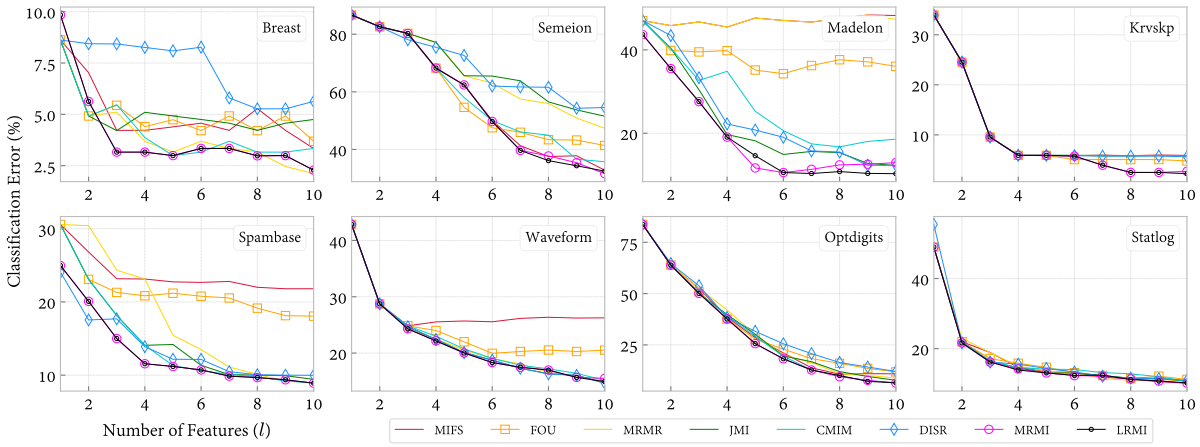

Figure 5: Number of Features (l) versus Classification Error (%) curves for different feature selection methods.

of $L_k(\mathbf{A}) \otimes L_k(\mathbf{B})$ are equal to $\lambda_r(\mathbf{A})\lambda_r(\mathbf{B})$, we have $\mathbf{S}_\alpha^{n^2-(n-k)^2}(L_k(\mathbf{A})\otimes L_k(\mathbf{B})) = \mathbf{S}_\alpha^k(L_k(\mathbf{A}))+\mathbf{S}_\alpha^k(L_k(\mathbf{B}))$.

For (e): From Proposition 4.1 in (Sanchez Giraldo, Rao, and Principe 2014) we have that $\mathbf{S}_\alpha(t\mathbf{A}_k + (1-t)\mathbf{B}_k) = g^{-1}\big(tg(\mathbf{S}_\alpha(\mathbf{A})) + (1-t)g(\mathbf{S}_\alpha(\mathbf{B}))\big)$. Notice that $\mathbf{A}_k = \mathbf{A}$ when $\mathbf{tr}(\mathbf{A}_k) = 1$, we have $\mathbf{S}_\alpha^k(t\mathbf{A} + (1-t)\mathbf{B}) = g^{-1}\big(tg(\mathbf{S}_\alpha^k(\mathbf{A})) + (1-t)g(\mathbf{S}_\alpha^k(\mathbf{B}))\big)$.

For (f): From the proof of Proposition 4.1 in (Sanchez Giraldo, Rao, and Principe 2014) we have that

$$\sum_{i=1}^{t}\lambda_i(\mathbf{A}\circ\mathbf{B}) \le \frac{1}{n}\sum_{i=1}^{t}\lambda_i(\mathbf{B}),$$

where t is any integer in $[1,n]$. Therefore

$$\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{A}\circ\mathbf{B})) \le \frac{1}{n}\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{B})), \quad \forall t\in[1,k],$$
$$\sum_{i=1}^{n}\lambda_i(L_k(\mathbf{A}\circ\mathbf{B})) = \frac{1}{n}\sum_{i=1}^{n}\lambda_i(L_k(\mathbf{B})) = \frac{1}{n},$$

From the case $t=k$ we know that $\lambda_r(\mathbf{A}\circ\mathbf{B}) \ge \lambda_r(\mathbf{B})/n$, therefore for any $t\in[k+1,n]$, we have

$$\begin{aligned}\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{A}\circ\mathbf{B})) &= \frac{1}{n} - (n-t)\lambda_r(\mathbf{A}\circ\mathbf{B})\\ &\le \frac{1}{n} - \frac{n-t}{n}\lambda_r(\mathbf{B})\\ &= \frac{1}{n}\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{B})).\end{aligned}$$

Then we can prove that

$$\begin{aligned}\mathbf{S}_\alpha^k\left(\frac{\mathbf{A}\circ\mathbf{B}}{\mathbf{tr}(\mathbf{A}\circ\mathbf{B})}\right) &= \mathbf{S}_\alpha\left(L_k\left(\frac{\mathbf{A}\circ\mathbf{B}}{\mathbf{tr}(\mathbf{A}\circ\mathbf{B})}\right)\right)\\ &\ge \mathbf{S}_\alpha(L_k(\mathbf{B})) = \mathbf{S}_\alpha^k(\mathbf{B})\end{aligned}$$

following the proof in (Sanchez Giraldo, Rao, and Principe 2014).

For (g): From the proof of Proposition 4.1 in (Sanchez Giraldo, Rao, and Principe 2014), when $\mathbf{A} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\mathbf{A} = \frac{1}{n}I$, we have

$$\sum_{i=1}^{t}\lambda_i(\mathbf{A}\circ\mathbf{B}) \le \frac{1}{n}\sum_{i=1}^{t}\lambda_i(\mathbf{B}) \quad \text{and}$$
$$\frac{1}{n}\sum_{i=1}^{t}\lambda_i(\mathbf{A}\circ\mathbf{B}) \le \frac{1}{n}\sum_{i=1}^{t}\lambda_i(\mathbf{B})$$

respectively, where t is any integer in $[1,n]$. Similar with the proof of (f), for these two extreme cases we can prove that

$$\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{A}\circ\mathbf{B})) \le \frac{1}{n}\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{B})) \quad \text{and}$$
$$\frac{1}{n}\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{A}\circ\mathbf{B})) \le \frac{1}{n}\sum_{i=1}^{t}\lambda_i(L_k(\mathbf{B}))$$

respectively. These inequalities imply that

$$\mathbf{S}_\alpha\left(L_k\left(\frac{\mathbf{A}\circ\mathbf{B}}{\mathbf{tr}(\mathbf{A}\circ\mathbf{B})}\right)\right) \le \mathbf{S}_\alpha(L_k(\mathbf{A})) + \mathbf{S}_\alpha(L_k(\mathbf{B}))$$

following the proof in (Sanchez Giraldo, Rao, and Principe 2014). □

Proof of Theorem 1

Proof. Without loss of generality, we assume $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_n$. Note that $\lambda_i, i\in[1,n]$ may not be monotonically decreasing. By the definition of information potential, we have

$$\mathbf{IP}_\alpha(\mathbf{B}) = \sum_{i=1}^{n}\mu_i^\alpha,$$

$$\mathbf{IP}_\alpha^k(\mathbf{B}) = \sum_{i=1}^{k}\mu_i^\alpha + (n-k)\mu_r^\alpha,$$

$$\mu_r = \frac{1}{n-k}\left(1 - \sum_{i=1}^{k}\mu_i\right).$$

When ν_i is small, we have the following first-order approximation:

$$\begin{aligned}\mu_i^\alpha &= (\lambda_i+\nu_i)^\alpha \\ &= \lambda_i^\alpha + \alpha\lambda_i^{\alpha-1}\nu_i + \frac{\alpha(\alpha-1)}{2}\lambda_i^{\alpha-2}\nu_i^2 + \cdots \\ &= \lambda_i^\alpha + \alpha\lambda_i^{\alpha-1}\nu_i + o(\nu_i).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{Var}[\mathbf{IP}_\alpha(\mathbf{B})] &= \mathbf{Var}[\mathbf{IP}_\alpha(\mathbf{B}) - \mathbf{IP}_\alpha(\mathbf{A})] \\ &= \mathbf{Var}\left[\sum_{i=1}^n \mu_i^\alpha - \lambda_i^\alpha\right] \\ &= \mathbf{Var}\left[\sum_{i=1}^n \alpha\lambda_i^{\alpha-1}\nu_i + o(\nu_i)\right] \\ &\approx \alpha^2\sum_{i=1}^n \mathbf{Var}\left[\lambda_i^{\alpha-1}\nu_i\right] \\ &= \alpha^2\sum_{i=1}^n \sigma_i^2\lambda_i^{2(\alpha-1)}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbf{Var}[\mathbf{IP}_\alpha^k(\mathbf{B})] &= \mathbf{Var}[\mathbf{IP}_\alpha^k(\mathbf{B}) - \mathbf{IP}_\alpha^k(\mathbf{A})] \\ &= \mathbf{Var}\left[\sum_{i=1}^k(\mu_i^\alpha - \lambda_i^\alpha) + (n-k)(\mu_r^\alpha - \lambda_r^\alpha)\right] \\ &= \mathbf{Var}\Bigg[\sum_{i=1}^k\big(\alpha\lambda_i^{\alpha-1}\nu_i + o(\nu_i)\big) \\ &\qquad - \alpha(n-k)\lambda_r^{\alpha-1}\cdot\frac{1}{n-k}\sum_{i=1}^k\nu_i\Bigg].\end{aligned}$$

When $\alpha\in(0,1)$, i.e. $\alpha-1<0$, we have $\lambda_r^{\alpha-1}\geq\lambda_i^{\alpha-1}$ for $i\in[1,k]$. Therefore

$$\begin{aligned}\mathbf{Var}[\mathbf{IP}_\alpha^k(\mathbf{B})] &\leq \mathbf{Var}\left[\alpha\lambda_r^{\alpha-1}\sum_{i=1}^k\nu_i\right] \\ &= \alpha^2\lambda_r^{2(\alpha-1)}\sum_{i=1}^k\sigma_i^2 \\ &\leq \alpha^2\sum_{i=k+1}^n\sigma_i^2\lambda_i^{2(\alpha-1)}\frac{\sum_{i=1}^k\sigma_i^2}{\sum_{i=k+1}^n\sigma_i^2} \qquad (1) \\ &\leq \mathbf{Var}[\mathbf{IP}_\alpha(\mathbf{B})].\end{aligned}$$

(1) follows by Jensen's inequality using the fact that $\lambda_r=\frac{1}{n-k}\sum_{i=k+1}^n\lambda_i$ and σ_i are non-negative, since the function $f(x)=x^{2(\alpha-1)}$ is convex.

Otherwise when $\alpha>1$, we have $\lambda_r^{\alpha-1}\leq\lambda_i^{\alpha-1}$ for $i\in[1,k]$. Therefore

$$\begin{aligned}\mathbf{Var}[\mathbf{IP}_\alpha^k(\mathbf{B})] &\leq \mathbf{Var}\left[\alpha\sum_{i=1}^k\lambda_i^{\alpha-1}\nu_i\right] \\ &= \alpha^2\sum_{i=1}^k\sigma_i^2\lambda_i^{2(\alpha-1)} \\ &\leq \mathbf{Var}[\mathbf{IP}_\alpha(\mathbf{B})].\end{aligned}$$

This completes the proof. □

Uniqueness of Low-rank Rényi's Entropy

Let $\mathbf{S}_\alpha^k(\mathbf{A})$ be a measure of entropy defined on the largest k eigenvalues of $\mathbf{A}$. Then $\mathbf{S}_\alpha^k(\mathbf{A})$ must adopt some strategy to build a probability distribution upon known eigenvalues, i.e. let the summation of all eigenvalues be exactly 1, otherwise $\mathbf{S}_\alpha^k(\mathbf{A})$ will not be continuous at $\alpha=1$. One choice is to adopt some strategy to complement the missing eigenvalues. Let $L_k(\mathbf{A})$ be the complemented matrix, we have $\lambda_i(L_k(\mathbf{A}))=\lambda_i(\mathbf{A}),\forall i\in[1,k]$, $\lambda_n(L_k(\mathbf{A}))\leq\cdots\leq\lambda_{k+1}(L_k(\mathbf{A}))\leq\lambda_k(\mathbf{A})$ and $\mathbf{tr}(L_k(\mathbf{A}))=1$.

Let $F_\mathbf{A}(t)$ be the CDF of $\mathbf{A}$: $F_\mathbf{A}(t)=\sum_{i=1}^t\lambda_i(\mathbf{A})$, and let $\lambda_r(\mathbf{A})=\frac{1}{n-k}(1-\sum_{i=1}^k\lambda_i(\mathbf{A}))$. Then we have

$$F_{L_k(\mathbf{A})}(t)\geq\sum_{i=1}^k\lambda_i(\mathbf{A})+(t-k)\lambda_r(\mathbf{A}) \qquad (2)$$

for all $t\in[k+1,n]$ since the function $F_\mathbf{A}$ is always concave. Let $\mathbf{B}\in\mathbb{R}^{n\times n}$ be a PSD matrix satisfying

$$\sum_{i=1}^t\lambda_i(\mathbf{A})\leq\sum_{i=1}^t\lambda_i(\mathbf{B})$$

for all $t\in[1,n]$, then in order to maintain the triangle inequality (axiom (f) and (g)), The function $F_\mathbf{A}$ must satisfy $F_{L_k(\mathbf{A})}(t)\leq F_{L_k(\mathbf{B})}(t)$, $\forall t\in[k+1,n]$. Construct $\mathbf{B}$ by letting $\lambda_1(\mathbf{B})=1-(n-1)\lambda_r(\mathbf{A})$ and $\lambda_2(\mathbf{B})=\cdots=\lambda_n(\mathbf{B})=\lambda_r(\mathbf{A})$, then combining with the fact that $\lambda_i(L_k(\mathbf{B}))\leq\lambda_k(\mathbf{B})$, $\forall i\in[k+1,n]$, we have

$$F_{L_k(\mathbf{A})}(t)\leq F_{L_k(\mathbf{B})}(t)=\sum_{i=1}^k\lambda_i(\mathbf{A})+(t-k)\lambda_r(\mathbf{A}). \qquad (3)$$

Eq. (2) and (3) together imply that taking $\lambda_i(L_k(\mathbf{A}))=\lambda_r(\mathbf{A})$ for all $i\in[k+1,n]$ is the only choice that fulfills all axioms in Proposition 1.

Another reasonable choice to normalize the probability distribution is to scale the known largest k eigenvalues:

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

or

$$\mathbf{S}_\alpha^k(\mathbf{A})=\frac{1}{1-\alpha}\log_2\left(\frac{\sum_{i=1}^k\lambda_i^\alpha(\mathbf{A})}{\sum_{i=1}^n\lambda_i(\mathbf{A})}\right).$$

However, these methods do not fulfill the triangle inequality, i.e. we cannot infer $\mathbf{S}_\alpha^k(\mathbf{A})\geq\mathbf{S}_\alpha^k(\mathbf{B})$ from the condition that $F_\mathbf{A}(t)\leq F_\mathbf{B}(t)$, $\forall t\in[1,k]$. This results in violations of axioms (f) and (g).

Proof of Theorem 2

We first present the ℓ_2 embedding results for RGP, SRHT, IST and SGS in Lemma 1, 2, 3 and 4 respectively, where the dimension of embedding subspace is given to guarantee the ϵ error. Lemma 5 presents the permutation bound for symmetric positive definite matrix. All these theoretical results are helpful to our proof.

Lemma 1. *(Foucart and Rauhut 2013) Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$ and $\mathbf{P} \in \mathbb{R}^{n\times s}$ constructed by GRP. Then, with probability at least $1-\delta$,*

$$\left\|\mathbf{U}^\top\mathbf{P}\mathbf{P}^\top\mathbf{U} - \mathbf{I}_k\right\|_2 \le \epsilon,$$

by setting $s = \mathcal{O}\left(k + \log(1/\delta)/\epsilon^2\right)$.

Lemma 2. *(Drineas et al. 2012) Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$ and $\mathbf{P} \in \mathbb{R}^{n\times s}$ constructed by SRHT. Then, with probability at least 0.9,*

$$\left\|\frac{n}{k}\mathbf{U}^\top\mathbf{P}\mathbf{P}^\top\mathbf{U} - \mathbf{I}_k\right\|_2 \le \epsilon,$$

by setting $s = \mathcal{O}\left((k+\log n)\frac{\log k}{\epsilon^2}\right)$.

Lemma 3. *(Woodruff 2014) Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$ and $\mathbf{P} \in \mathbb{R}^{n\times s}$ constructed by IST. Then, with probability at least 0.9,*

$$\left\|\mathbf{U}^\top\mathbf{P}\mathbf{P}^\top\mathbf{U} - \mathbf{I}_k\right\|_2 \le \epsilon,$$

by setting $s = \mathcal{O}\left(k^2/\epsilon^2\right)$.

Lemma 4. *(Hu et al. 2021) Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$ and $\mathbf{P} \in \mathbb{R}^{n\times s}$ constructed by SGS. Then, with probability at least $1-\delta$,*

$$\left\|\mathbf{U}^\top\mathbf{P}\mathbf{P}^\top\mathbf{U} - \mathbf{I}_k\right\|_2 \le \epsilon,$$

by setting

$$s = \mathcal{O}\left(k\log(k/\delta\epsilon)/\epsilon^2\right),$$
$$p = \mathcal{O}\left(\log(k/\delta\epsilon)/\epsilon\right).$$

Lemma 5. *(Demmel and Veselić 1992) Let $\mathbf{DGD}$ be a symmetric positive definite matrix such that $\mathbf{D}$ is a diagonal matrix and $\mathbf{G}_{ii} = 1$ for all i. Let $\mathbf{DED}$ be a permutation matrix such that $\|\mathbf{E}\|_2 < \lambda_{\min}(\mathbf{G})$. Let λ_i be the i-th eigenvalue of $\mathbf{DGD}$ and $\hat{\lambda}_i$ be the i-th eigenvalue of $\mathbf{D}(\mathbf{G}+\mathbf{E})\mathbf{D}$. Then, for all i,*

$$|\lambda_i - \hat{\lambda}_i| \le \frac{\|\mathbf{E}\|_2}{\lambda_{\min}(\mathbf{G})}.$$

The following proposition shows that if we can bound each eigenvalue of $\mathbf{A}$ to absolute error ϵ, we have an absolute bound for $\mathbf{S}_\alpha(\mathbf{A})$.

Proposition 1. *Let $\mathbf{A}$ and $\hat{\mathbf{A}}$ be positive definite matrices with eigenvalues λ_i and $\hat{\lambda}_i$, $i \in [1,n]$ respectively, such that for each $i \in [1,n]$, $|\lambda_i - \hat{\lambda}_i| \le \epsilon$, then*

$$|\mathbf{S}_\alpha(\mathbf{A}) - \mathbf{S}_\alpha(\hat{\mathbf{A}})| \le \left|\frac{\alpha}{1-\alpha}\log_2\left(1 - \frac{\epsilon}{\lambda_n}\right)\right|.$$

Proof. Let $\lambda_n > 0$ be the smallest eigenvalue of $\mathbf{A}$ and let $\epsilon_0 = \epsilon/\lambda_n$, then we have $|\lambda_i - \hat{\lambda}_i| \le \epsilon_0\lambda_i$ for each $i \in [1,n]$. Observe that when $\alpha < 1$,

$$\begin{aligned}
\mathbf{S}_\alpha(\hat{\mathbf{A}}) &= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^n \hat{\lambda}_i^\alpha\right) \\
&\ge \frac{1}{1-\alpha}\log_2\left((1-\epsilon_0)^\alpha\sum_{i=1}^n \lambda_i^\alpha\right) \\
&= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^n \lambda_i^\alpha\right) + \frac{\alpha}{1-\alpha}\log_2(1-\epsilon_0) \\
&= \mathbf{S}_\alpha(\mathbf{A}) + \frac{\alpha}{1-\alpha}\log_2(1-\epsilon_0).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbf{S}_\alpha(\hat{\mathbf{A}}) &= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^n \hat{\lambda}_i^\alpha\right) \\
&\le \frac{1}{1-\alpha}\log_2\left((1+\epsilon_0)^\alpha\sum_{i=1}^n \lambda_i^\alpha\right) \\
&= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^n \lambda_i^\alpha\right) + \frac{\alpha}{1-\alpha}\log_2(1+\epsilon_0) \\
&= \mathbf{S}_\alpha(\mathbf{A}) + \frac{\alpha}{1-\alpha}\log_2(1+\epsilon_0).
\end{aligned}$$

We can get the same results for the other case when $\alpha > 1$, which finishes the proof. □

Proof of Theorem 2. Note that $\lambda_{\min}(\mathbf{G})$ in the Lemma 5 is a real, strictly positive number since $\mathbf{G}$ is positive definite and the fact $0 \le \|E\|_2\lambda_{\min}(\mathbf{G})$. Now consider the matrix $\mathbf{A}\mathbf{P}\mathbf{P}^\top\mathbf{A}^\top$, we will show that the singular values of $\mathbf{A}\mathbf{P}\mathbf{P}^\top\mathbf{A}$ are sufficient approximation to that of $\mathbf{A}\mathbf{A}^\top$ by the permutation theory presented in Lemma 5.

Let λ_i, $i \in [1,n]$ be the eigenvalues of the positive definite kernel matrix $\mathbf{A}$, $\hat{\lambda}_i$ be their approximations and $\mathbf{A} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top$ be the eigenvalue decomposition $\mathbf{A}$. Since $\mathbf{\Phi}$ is an orthogonal matrix, we have that the eigenvalues of $\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{P}\mathbf{P}^\top\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top$ are equal to the eigenvalues of $\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{P}\mathbf{P}^\top\mathbf{\Phi}\mathbf{\Sigma}$. Let $\mathbf{\Sigma}_k$ be the $k\times k$ diagonal matrix containing the k largest eigenvalues of $\mathbf{A}$ and $\mathbf{\Phi}_k$ be the matrix containing the corresponding eigenvectors, then λ_i^2, $i \in [1,k]$ are the eigenvalues of matrix $\mathbf{\Sigma}_k\mathbf{I}_k\mathbf{\Sigma}_k$, and $\hat{\lambda}_i^2$, $i \in [1,k]$ are the eigenvalues of matrix $\mathbf{\Sigma}_k\mathbf{\Phi}_k^\top\mathbf{P}\mathbf{P}^\top\mathbf{\Phi}_k\mathbf{\Sigma}_k$ (since the first k singular values of $\mathbf{\Sigma}_k\mathbf{\Phi}_k^\top\mathbf{P}$ are equal to those of $\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{P} = \mathbf{A}\mathbf{P}$). Let $\mathbf{E} = \mathbf{\Phi}_k^\top\mathbf{P}\mathbf{P}^\top\mathbf{\Phi}_k - \mathbf{I}_k$, we know from Lemma 1 (or Lemma 2, 3 and 4) that $\|\mathbf{E}\|_2 \le \epsilon_0$ with high probability. It meets the condition of Lemma 5 since $\lambda_{\min}(\mathbf{I}_k) = 1$. Hence, we have

$$|\lambda_i^2 - \hat{\lambda}_i^2| \le \epsilon_0, \quad \forall i \in [1,k],$$

which then implies that

$$\lambda_i - \sqrt{\lambda_i^2 - \epsilon_0} \le |\hat{\lambda}_i - \lambda_i| \le \sqrt{\lambda_i^2 + \epsilon_0} - \lambda_i.$$

Since λ_k is the smallest eigenvalue amongst λ_i, $i \in [1,k]$, we have

$$\begin{aligned}|\hat{\lambda}_i - \lambda_i| &\leq \lambda_k - \sqrt{\lambda_k^2 - \epsilon_0} \\ &= \lambda_k\left(1 - \sqrt{1 - \frac{\epsilon_0}{\lambda_k^2}}\right) \\ &\leq \lambda_k \frac{\epsilon_0}{\lambda_k^2} = \frac{\epsilon_0}{\lambda_k}.\end{aligned}$$

Combining with $k \leq n/2$, we have

$$\begin{aligned}|\hat{\lambda}_r - \lambda_r| &= \left|\sum_{i=1}^{k}\hat{\lambda}_i - \sum_{i=1}^{k}\lambda_i\right| \Big/ (n-k) \\ &\leq \frac{\epsilon_0}{\lambda_k} \cdot \frac{k}{n-k} \leq \frac{\epsilon_0}{\lambda_k}.\end{aligned}$$

Let $\epsilon_0 = \epsilon\lambda_k\lambda_r$ and $\mathbf{B}$ be a positive definite matrix with the first k eigenvalues equal to $\hat{\lambda}_i$, $i \in [1,k]$ and the other $n-k$ eigenvalues equal to $\hat{\lambda}_r$. Recall that λ_r is the smallest eigenvalue of $L_k(\mathbf{A})$, by applying Proposition 1, we have

$$\begin{aligned}|\mathbf{S}_\alpha^k(\mathbf{A}) - \hat{\mathbf{S}}_\alpha^k(\mathbf{A})| &= |\mathbf{S}_\alpha(L_k(\mathbf{A})) - \mathbf{S}_\alpha(\mathbf{B})| \\ &\leq \left|\frac{\alpha}{1-\alpha}\log_2(1-\epsilon)\right|.\end{aligned}$$

□

A Potential Improvement

The upper bound of s in Theorem 2 relies on λ_r, which grows large if the kernel matrix $\mathbf{A}$ is ill-posed and λ_r is small. Alternatively, we derive an upper bound for s in terms of n and $\mathbf{tr}(\mathbf{A}^\alpha)$, which is tighter for such kernel matrices.

Proposition 2. *Under the same conditions as Proposition 1, we have*

$$|\mathbf{S}_\alpha(\mathbf{A}) - \mathbf{S}_\alpha(\hat{\mathbf{A}})| \leq \begin{cases}\left|\frac{1}{1-\alpha}\log\left(1 - \frac{n\alpha\epsilon}{1+n\epsilon}\right)\right| & \text{if } \alpha > 1, \\ \left|\frac{1}{1-\alpha}\log\left(1 - \frac{n\epsilon^\alpha}{tr(A^\alpha)}\right)\right| & \text{if } \alpha < 1.\end{cases}$$

Proof.

$$\begin{aligned}S_\alpha(\tilde{\mathbf{A}}) - S_\alpha(\mathbf{A}) &= \left|\frac{1}{1-\alpha}\log\left(1 - \frac{\sum_{i=1}^n \hat{\lambda}_i^\alpha - \sum_{i=1}^n \lambda_i^\alpha}{\sum_{i=1}^n \hat{\lambda}_i^\alpha}\right)\right| \\ &\leq \left|\frac{1}{1-\alpha}\log(1-\beta)\right|,\end{aligned}$$

where

$$\begin{aligned}\beta &= \left|\frac{\sum_{i=1}^n(\lambda_i+\epsilon)^\alpha - \sum_{i=1}^n \lambda_i^\alpha}{\sum_{i=1}^n(\lambda_i+\epsilon)^\alpha}\right| \\ &\leq \left|\frac{\sum_{i=1}^n(\lambda_i+\epsilon)^\alpha - \sum_{i=1}^n \lambda_i^\alpha}{\sum_{i=1}^n \lambda_i^\alpha}\right|.\end{aligned}$$

When $\alpha > 1$, we have

$$\beta \leq \alpha\epsilon\left|\frac{\sum_{i=1}^n(\lambda_i+\epsilon)^{\alpha-1}}{\sum_{i=1}^n(\lambda_i+\epsilon)^\alpha}\right| \leq \frac{n\alpha\epsilon}{1+n\epsilon},$$

where the last step takes equality if and only if $\lambda_1 = \cdots = \lambda_n = \frac{1}{n}$. Otherwise when $\alpha < 1$,

$$\beta \leq \left|\frac{\sum_{i=1}^n \epsilon^\alpha}{\sum_{i=1}^n \lambda_i^\alpha}\right| = \frac{n\epsilon^\alpha}{tr(A^\alpha)}.$$

One can upper bound $S_\alpha(A) - S_\alpha(\tilde{A})$ through the same strategy, which finishes the proof. □

Proof of Theorem 3

The following lemma gives the convergence rate of the Lanczos algorithm:

Lemma 6. *(Saad 1980) Let $\mathbf{q}$ be the initial vector, λ_i be the i-th largest eigenvalue of $\mathbf{A}$ with associated eigenvector $\boldsymbol{\phi}_i$ such that $\langle\boldsymbol{\phi}_i, \mathbf{q}\rangle \neq 0$, $\hat{\lambda}_i$ be the corresponding approximation of λ_i after s steps of Lanczos iteration, and assume that $\hat{\lambda}_{i-1} > \lambda_i$. Let*

$$\gamma_i = 1 + 2\frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_n},$$

$$K_i = \begin{cases}\prod_{j=1}^{i-1}\frac{\hat{\lambda}_j - \lambda_n}{\hat{\lambda}_j - \lambda_i}, & i > 1 \\ 1, & i = 1\end{cases},$$

then

$$0 \leq \lambda_i - \hat{\lambda}_i \leq (\lambda_i - \lambda_n) \cdot \left(\frac{K_i}{T_{s-i}(\gamma_i)}\tan\langle\boldsymbol{\phi}_i, \mathbf{q}\rangle\right)^2,$$

where $T_i(x) = \frac{1}{2}\left[(x+\sqrt{x^2-1})^i + (x-\sqrt{x^2-1})^i\right]$ is the Chebyshev polynomial of the first kind of degree i.

Proof of Theorem 3. It is easy to see that K_i is monotonically increasing with the increase of i. Let

$$\begin{aligned}\gamma &= \min_{i\in[1,k]}\gamma_i, \\ \theta &= \max_{i\in[1,k]}\tan\langle\boldsymbol{\phi}_i, \mathbf{q}\rangle, \\ R &= \gamma + \sqrt{\gamma^2 - 1},\end{aligned}$$

then $\forall i \in [1,k]$,

$$\begin{aligned}\lambda_i - \hat{\lambda}_i &\leq \lambda_i \cdot \left(\frac{2\theta K_i}{R^{s-i} + R^{-(s-i)}}\right)^2 \\ &\leq \lambda_i \cdot 4\theta^2 K_i^2 R^{-2(s-i)} \\ &\leq \lambda_i \cdot 4\theta^2 K_k^2 R^{-2(s-k)}.\end{aligned}$$

By selecting $s = \left\lceil k + \frac{\log(4\theta^2 K_k^2/\epsilon_0)}{2\log R}\right\rceil$, we have that $\forall i \in [1,k]$,

$$|\lambda_i - \hat{\lambda}_i| \leq \epsilon_0\lambda_i.$$

Similarly, by combining with $k \leq n/2$ we have

$$|\hat{\lambda}_r - \lambda_r| \leq \epsilon_0\lambda_1.$$

Let $\epsilon_0 = \epsilon\lambda_r/\lambda_1$, by applying Proposition 1, we have

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

□

References

Demmel, J.; and Veselić, K. 1992. Jacobi's method is more accurate than QR. *SIAM Journal on Matrix Analysis and Applications*, 13(4): 1204–1245.

Demvsar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7: 1–30.

Drineas, P.; Magdon-Ismail, M.; Mahoney, M. W.; and Woodruff, D. P. 2012. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1): 3475–3506.

Foucart, S.; and Rauhut, H. 2013. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel.

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.

Saad, Y. 1980. On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM Journal on Numerical Analysis*, 17(5): 687–706.

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.

Woodruff, D. P. 2014. Sketching as a Tool for Numerical Linear Algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2): 1–157.