

ASYMPTOTIC OPTIMALITY OF SELF-REPRESENTATIVE LOW-RANK APPROXIMATION AND ITS APPLICATIONS

Anonymous authors

Paper under double-blind review

ABSTRACT

We propose a novel technique for finding representatives from a large, unsupervised dataset. The approach is based on the concept of *self-rank*, defined as the minimum number of samples needed to reconstruct all samples with an accuracy proportional to the rank- K approximation. As the exact computation of self-rank requires a computationally expensive combinatorial search, we propose an efficient algorithm that jointly estimates self-rank and selects the optimal samples. The ratio of obtained projection error using selected samples to the error of best rank- K approximation is called approximation ratio (AR). In this paper, a new upper bound for AR is derived which is tight for two asymptotic cases. The best AR for self-representative low-rank approximation was presented in ICML 2017 Chierichetti et al. (2017), which was further improved by the bound $\sqrt{1+K}$ reported in NeurIPS 2019 Dan et al. (2019). Both of these bounds are obtained by brute force search which is not practical and these bounds depend solely on K , the rank which is targeted. In this paper, for the first time, we present an adaptive AR depending on spectral properties of the original dataset, $\mathbf{A} \in \mathbb{R}^{N \times M}$. In particular, our performance bound is proportional to the condition number $\kappa(\mathbf{A})$ which is a well-known spectral property. Our derived AR is expressed as $1 + (\kappa(\mathbf{A})^2 - 1)/(N - K)$ which approaches 1 in certain asymptotic cases. Our proposed algorithm enjoys linear complexity w.r.t. the size of original dataset which results in filling a historical gap between practical and theoretical methods in finding representatives. In addition to evaluating the proposed algorithm on a synthetic dataset, we show that it can be utilized in real-world applications such as graph node sampling for optimizing the shortest path criterion, and learning a classifier with representative data.

INTRODUCTION

Low-rank approximation is one of the fundamental workhorses of machine learning and optimization. It has applications for dimension reduction and clustering in recommendation systems, text mining, and computer vision. It is a building block for data compression algorithms such as PCA, where the data are transferred into an eigenspace representation to learn the representative samples which are suitable for the whole data. The representations are either in a lower-dimensional space or expressed in terms of pseudo-data not part of the original dataset.¹ Thus, these methods are not applicable for data sampling. However, it was shown recently that spectral methods are useful tools to devise a sampling framework Zaeemzadeh et al. (2019). In spectral-based sampling, the goal is to keep the spectrum of representatives as close as possible to the spectrum of the original data. Shannon sampling Unser (2000), the most theoretically strong sampling scheme also follows the strategy of keeping the spectrum of the original and sampled function equal to each other, and a reconstruction guarantee is established for signal sampling in an infinite-dimensional space based on spectral properties.

Datasets encountered in practice are usually composed of a large number of vectors in a finite-dimensional space. Assume M samples in an N -dimensional space organized in matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$. Let \mathbb{S} be a set of integers between 1 and M , which, when used to index the columns of \mathbf{A} , define a collection of samples. We define a matrix of the form $\mathbf{V} = [v_{mk}]$ which projects back from the sampled to the original data. We consider the following optimization problem for finding the optimal sampling

¹For example, when studying a face dataset, eigenfaces are spectral components found by PCA and are not part of the dataset. This is similar to K-means clustering, where centroids are not members of the original dataset.

and back-projection:

$$\{\mathbb{S}^*, \mathbf{V}\} = \underset{\mathbb{S}, \mathbf{V}}{\operatorname{argmin}} \sum_{m=1}^M \|\mathbf{a}_m - \sum_{k \in \mathbb{S}} \mathbf{a}_k v_{mk}\|_2^2 + \lambda |\mathbb{S}|, \quad (1)$$

where λ is a parameter that regularizes the rate of sampling. We refer the cardinality of the set \mathbb{S}^* as the *self-rank* of matrix \mathbf{A} in this paper. Substituting \mathbf{a}_k in the inner summation with an arbitrary vector $\mathbf{u}_k \in \mathbb{R}^N$ simplifies this problem to the truncated singular value decomposition (SVD). In this case, $|\mathbb{S}|$ is simplified to the conventional definition of rank for matrix \mathbf{A} . However, the conventional spectral decomposition using SVD is not suitable for data sampling since singular vectors are arbitrary vectors not actual samples. Enforcing the principal bases to be from the dataset itself in equation 1 results in a self-representative low-rank approximation. This problem is related to the column subset selection problem (CSSP) Boutsidis et al. (2009), a problem that is known to be NP-hard Shitov (2017); Çivril (2014) and subject of ongoing research Dan et al. (2019); Song et al. (2019b;a).

In this paper, we propose a versatile sampling framework to determine the self-rank and to identify the corresponding samples. Our approach is based on spectral decomposition; we compute the minimum number of samples that cover the K most significant spectral components of the dataset. The main motivation of the present work and the role of proposed method among the vast literature of CSSP are explained in the next section. The main contributions of this paper can be summarized as follows:

- We introduce Spectrum Pursuit with Residual Descent (SP-RD), a constructive algorithm that jointly estimates the self rank and samples corresponding to informative columns.
- We prove an upper bound for the SP-RD projection error, and show that in two special cases, the bound is asymptotically the tightest one.
- We show that the SP-RD algorithm leads to improved performance in a number of applications, including graph node sampling with shortest path criterion and selection of training data for a classifier.

RELATED WORK AND MOTIVATION

There is a large line of work on data selection in machine learning which are mainly studied under context of CSSP and volume sampling (VS). See Bardenet & Maillard (2015); Dieng et al. (2017); Campbell & Broderick (2018); Langberg & Schulman (2010); Dasgupta (2011; 2006); Huggins et al. (2016); Clarkson & Woodruff (2013); J. W. Demmel & Xiang (2015); Gu (2015); Duersch & Gu (2015) for a non-exhaustive list. Motivated by the fact that datasets continue to grow larger and larger over time, a popular approach relies on the modification of the dataset itself, such that its size is shrunk while preserving its original statistical properties. On this observation, (Bardenet & Maillard, 2015) studied the reduction in size of a large dataset using random linear projection. On a similar note, (Huggins et al., 2016; Dieng et al., 2017; Campbell & Broderick, 2018) constructed a weighted subset of a large dataset, the Bayesian coresets, for a wider class of Bayesian models.

Many other efforts have been devoted to data selection and coresets construction algorithms. Examples include pivoted QR factorization and matrix subset selection, which are under the umbrella of column subset selection problem Paul et al. (2015); Dan et al. (2019); Song et al. (2019a;b), that can be considered as a form of unsupervised feature selection or prototype selection. We briefly discuss the key models relevant to our work. One line of work focuses on the randomized matrix algorithms which have been developed in fast least squares Clarkson & Woodruff (2013), sketching algorithms J. W. Demmel & Xiang (2015) and low-rank approximation problems Gu (2015). Authors in Duersch & Gu (2015) develop a randomized QR with column pivoting (RQRC), where random sampling (RNDS) is used for column selection. A comprehensive summary of these algorithms and their theoretical guarantee is available in Table 1 in Christos Boutsidis & Drinea (2009). Most of these algorithms can be roughly categorized into two branches. One branch of algorithms are based on rank-revealing QR (RRQR) decomposition Gu & Eisenstat (1996). It has been proved in Christos Boutsidis & Drinea (2009) that RRQR is nearly optimal in terms of residue norm. On the other hand, sampling based methods Petros Drineas & Muthukrishnan (2008) try to select columns by sampling from certain distributions over all columns of an input matrix. Extension of sampling based methods to general low-rank matrix approximation problems was also investigated in Srinadh Bhojanapalli (2016). Let us denote the best rank- K approximation of matrix \mathbf{A} by \mathbf{A}_K . The data matrix \mathbf{A} is approximated in polynomial-time in Deshpande & Rademacher (2010b) and a lower bound on the required number of samples as $O(K \log K + K/\epsilon)$, where ϵ is the coefficient of the relative error,

Table 1: The comparison of upper bound on the error of some well-known low-rank approximation algorithms for the CSSP. In all methods, K is considered to be the target rank and approaching \mathbf{A}_K is the goal of sampling. Our achieved AR is compared with Brute-force Dan et al. (2019), Improved CSSP Boutsidis et al. (2009), Volume sampling Deshpande & Rademacher (2010a), CUR Decomposition Drineas et al. (2008), and Near-optimal Boutsidis et al. (2014).

Algorithm	selected samples	Upper bound	Complexity
Brute-force	K	$\sqrt{1+K}\ \mathbf{A}-\mathbf{A}_K\ _F$	brute force $\binom{M}{K}$
Improved CSSP	K	$O(K\sqrt{\log K})\ \mathbf{A}-\mathbf{A}_K\ _F$	$O(\min\{MN^2, M^2N\})$
Volume sampling	K	$(K+1)\ \mathbf{A}-\mathbf{A}_K\ _F$	$O(KNM^w \log M)$
CUR Decomposition	$O(K \log(K)/\varepsilon^2)$	$(1+\varepsilon)\ \mathbf{A}-\mathbf{A}_K\ _F$	$O(MK^2 \log(K))$
Near-optimal	$2K/\varepsilon$	$(1+\varepsilon)\ \mathbf{A}-\mathbf{A}_K\ _F$	$O(NMK^2\varepsilon^{-2})$
Ours	K	$(1+\varepsilon)\ \mathbf{A}-\mathbf{A}_K\ _F$	$O(NMK)$

has been obtained in Deshpande & Rademacher (2010b); P. Drineas & Muthukrishnan (2006). In other words, a subset of data is selected such that the projection error of all samples on it is less than $(1+\varepsilon)\|\mathbf{A}-\mathbf{A}_K\|_F$. For a smaller value of ε , we need more selected samples to achieve the desired projection error. A tighter bound for the number of required samples is introduced in Boutsidis et al. (2014). They show that $O(K/\varepsilon)$ columns contain a subspace which approximates \mathbf{A} with coefficient error of $(1+\varepsilon)$. Here, we introduced our approach based on recent advances on data sampling. The reader is invited to see the supplementary document for more elaboration. Our proposed bound is the first work in the literature that targets rank- K approximation using only K samples while ε can approach to 0 asymptotically.

The AR of existing low-rank approximation methods are only a function of K . We refer the reader to Table 1 for a brief survey of upper bounds on various low-rank models for sampling. Methods in Table 1 are either impractical or shown to be working in very limited practical settings in their experiments. This motivates two main questions in our work: *How many samples suffice to represent a dataset?* and *How can we select such data representatives considering the intrinsic structure of the dataset?*. Our proposed method is based on spectrum pursuit (SP) algorithm which is shown to be working efficiently in a wide range of applications Joneidi et al. (2020). In the present paper we will show theoretically that why SP outperforms state-of-the-art methods in sampling. Moreover, an extension of SP algorithm will be presented which is more accurate and provably convergent. As reflected in Table 1, there is no sampling algorithm whose AR is controlled by structural or spectral properties (such as condition number $\kappa(\mathbf{A})$) of the matrix \mathbf{A} . In this paper, a theoretical upper bound is established which depends on spectral properties of \mathbf{A} in addition to data dimensions and the target rank K . Sampling literature includes a vast set of practical methods which function in limited occasions appropriately, however, come short in theoretical guarantee. On the other hand, one can find sampling methods in the literature supported by established theoretical guarantees with no practical applications. In the present work, for the first time we fill a historical gap between practical sampling and theoretical sampling.

JOINT SELF-RANK ESTIMATION AND SAMPLING

We introduce an equivalent problem to equation 1 that is solved efficiently.

$$\mathbb{S}^* = \underset{\mathbb{S}}{\operatorname{argmin}} |\mathbb{S}| \quad \text{s.t.} \quad \sum_{m=1}^M \|\mathbf{a}_m - \sum_{k \in \mathbb{S}} \mathbf{a}_k v_{mk}\|_2^2 \leq E_K \|\mathbf{A}\|_F^2. \quad (2)$$

where $E_K \in [0, 1]$ is the projection error corresponding to the best rank- K approximation of data normalized to $\|\mathbf{A}\|_F^2$. Parameter K is a user specified target rank. As K increases, more samples are needed in set \mathbb{S} in order to reach the desired error. The cardinality of set \mathbb{S}^* for a specified K is denoted by $S_K(\mathbf{A})$ and it is called the self-rank of dataset \mathbf{A} with parameter K . For example, setting $E_K = 0.25$ results in a set of samples such that their combination is able to reconstruct all data with 25% error. It is straightforward to find a target rank to provide the desired error by using truncated SVD. Fig. 1 Left shows the low-rank approximation error of SVD versus an assumed target rank for a subset of Multi-PIE face dataset with 52,000 images. For instance, 9 spectral components are needed to span all data with a normalized error of less than 0.25. Since spectral components in SVD are not among samples of data, we need to solve Problem (2) in order to find the minimum number of actual samples, such that their span provides a span as accurate as that of the first 9 spectral components. Fig. 1 Right

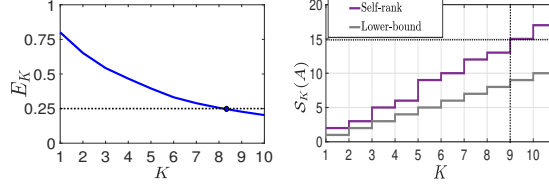


Figure 1: The relation between an assumed target rank and the corresponding self-rank. (a) Reconstruction error of the truncated SVD for 52,000 images from the CMU Multi-PIE face dataset. In order to reach below the normalized error of 0.25 indicated by the dark broken line $K = 9$ spectral components are needed. (b) The computed self-rank by solving Problem (2) versus difference values of K . Setting $K = 9$ ensures an error below 0.25 and it requires at least 15 real samples.



Figure 2: (a) The set of first 9 spectral components a.k.a. Eigen (Fisher) faces for 52,000 images from CMU Multi-PIE dataset. Linear combination of these components is able to reconstruct all 52,000 images with an average error less than 25%. (b) A subset of dataset with 15 images is found such that their span is as accurate as the span of the first 9 spectral components.

shows the computed self-rank versus the target rank. As an example, there is a subset of the dataset with 15 samples such that their span is as accurate as the span of first 9 spectral components as shown in Fig 2. In other words, there are 15 samples in the dataset such that they are able to reconstruct all the dataset with an error less than 25%. The normalized error, E_K , is a user-specified parameter and it can be set to any value between 0 and 1. The target rank of matrix \mathbf{A} corresponding to a desired projection error, E_K , is defined as the minimum integer K that satisfies $(\sum_{n=K+1}^N \sigma_n^2) / \|\mathbf{A}\|_F^2 \leq E_K$. The n^{th} singular value of matrix \mathbf{A} is denoted by σ_n . The self-rank of matrix \mathbf{A} with parameter K is equal to the minimum number of columns in \mathbf{A} such that their span approximates \mathbf{A} by an error less than the error of best rank- K approximation, i.e., the smallest size for set \mathbb{S} that holds

$$\|\mathbf{A} - \mathbf{A}_{\mathbb{S}} \mathbf{A}_{\mathbb{S}}^{\dagger} \mathbf{A}\|_F \leq \frac{\sum_{n=K+1}^N \sigma_n^2}{\|\mathbf{A}\|_F^2}.$$

Matrix $\mathbf{A}_{\mathbb{S}}$ refers to sampled columns of \mathbf{A} , and $\mathbf{A}_{\mathbb{S}}^{\dagger}$ is the Moore–Penrose inverse of $\mathbf{A}_{\mathbb{S}}$.

PRACTICAL ALGORITHM FOR SAMPLING

Solving Problem equation 2 implies a combinatorial search that is not feasible for a massive dataset and a large target rank. However, computing the target rank is tractable via SVD. Since self-rank is lower-bounded by target rank, a practical algorithm should start searching for self-rank values greater than the target rank. Thus, we start with solving the following problem with a fixed $|\mathbb{S}| = K$, and check if the constraint in Problem equation 2 is satisfied for a desired error threshold.

$$\mathbb{S}^* = \underset{\mathbb{S}}{\operatorname{argmin}} \sum_{m=1}^M \|\mathbf{a}_m - \sum_{k \in \mathbb{S}} \alpha_k \mathbf{v}_{mk}\|_2^2 \quad \text{s.t.} \quad |\mathbb{S}| \leq K. \quad (3)$$

The solution for \mathbb{S} does not guarantee that the error in the constraint of Problem equation 2 is less than the desired E_K . The minimum integer, K , that satisfies the constraint in (2) is considered as the self-rank of dataset \mathbf{A} with parameter K . If the error is higher than the desired threshold indicated by E_K , the cardinality of set \mathbb{S} is increased by 1, and Problem (3) is solved with a new value for K .

Spectrum Pursuit (SP) Algorithm is proposed in Joneidi et al. (2020) for solving equation 3. Inspired by SP, a more accurate solver for equation 3 is proposed in this paper which is called Spectrum Pursuit with Residual Descend (SP-RD). Deriving SP-RD Alg from equation 3 and its convergence analysis are studied in the supplementary document. Alg. 1 shows the steps of SP-RD algorithm. In this algorithm P is a parameter that controls the computational burden of the algorithm. The joint problem of self-rank estimation and data sampling is summarized in Alg. 2. The heart of this algorithm is the mentioned SP-RD algorithm. In the experiments we refer Alg. 2 as adaptive SP-RD. In Line 1 of the algorithm K can be initialized using truncated SVD easily. Moreover, $\Pi_{\mathbb{S}}(\mathbf{A})$ is the projection

operator on the subspace spanned by columns of \mathbf{A} indexed by set \mathbb{S} . Mathematically, it is equal to $\mathbf{A}_{\mathbb{S}}(\mathbf{A}_{\mathbb{S}}^T \mathbf{A}_{\mathbb{S}})^{-1} \mathbf{A}_{\mathbb{S}}^T \mathbf{A}$. It is noteworthy that the computational complexity of the SP-RD algorithm is $O(MN + K^2N + MNP)$ per iteration, which depends on the computational burden parameter P . Note that we only need the first singular vector and there are fast methods to get it. For the most relevant cases, large dataset with $K < N < M$, the complexity is dominated by the $O(MNP)$ per iteration and we need to perform $O(K)$ number of iterations. This is a linear complexity w.r.t M .

Algorithm 1: Spectrum Pursuit with Residual Descent (SP-RD)

Require: \mathbf{A} , P and K

Output: \mathbb{S}

1: **Initialization:**

$\mathbb{S} \leftarrow \mathbf{A}$ random subset of $\{1, \dots, M\}$ with $|\mathbb{S}| = K$
 $\{\mathbb{S}_k\}_{k=1}^K \leftarrow$ Partition \mathbb{S} into K singletons.

iter=0

while the stopping criterion is not met

2: $k = \text{mod}(\text{iter}, K) + 1$

3: $\mathbf{U}_{\bar{k}} = \text{normalize column}(\mathbf{A}_{\mathbb{S} \setminus \mathbb{S}_k})$

4: $\mathbf{V}_{\bar{k}} = \mathbf{A}^T \mathbf{U}_{\bar{k}} (\mathbf{U}_{\bar{k}}^T \mathbf{U}_{\bar{k}})^{-1}$

5: $\mathbf{E}_{\bar{k}} = \mathbf{A} - \mathbf{U}_{\bar{k}} \mathbf{V}_{\bar{k}}^T$

6: $\mathbf{u}_k =$ first left singular-vector of $\mathbf{E}_{\bar{k}}$

7: $\Omega \leftarrow P$ most correlated Col. of $\mathbf{E}_{\bar{k}}$ with \mathbf{u}_k

8: $\Omega = \Omega \cup \mathbb{S}_k$

9: $\mathbb{S}_k \leftarrow \arg\min_{k \in \Omega} \|\mathbf{E}_{\bar{k}} - \mathbf{a}_k \mathbf{a}_k^T \mathbf{E}_{\bar{k}}\|_F$

10: $\mathbb{S} \leftarrow \bigcup_{k'=1}^K \mathbb{S}_{k'}$

11: iter=iter+1

end while

Algorithm 2: Joint Self-rank Estimation and Data Sampling via Adaptive SP-RD

Require: dataset \mathbf{A} , P , and the desired sampling error $e \in [0, 1]$

Output: set \mathbb{S} .

1: $K \leftarrow$ smallest K such that $\|\mathbf{A} - \mathbf{A}_K\|_F^2 \leq e \|\mathbf{A}\|_F^2$

$E_K = e$

$\alpha = K$

$\mathbb{S} \leftarrow \text{SP-RD}(\mathbf{A}, P, \alpha)$

While $\|\mathbf{A} - \Pi_{\mathbb{S}}(\mathbf{A})\|_F^2 > E_K \|\mathbf{A}\|_F^2$

2: $\alpha \leftarrow \alpha + 1$

3: $\mathbb{S} \leftarrow \text{SP-RD}(\mathbf{A}, P, \alpha)$

The state-of-the-art upper bound for the minimum required number of samples is introduced as $O(\frac{K}{\varepsilon})$ Boutsidis et al. (2014), in order to guarantee that the projection error of sampling is not worse than the projection error of rank- K approximation. In general, the purpose in CSSP problem is to find an upper bound on the number of sampled data in set \mathbb{S} to ensure that $\|\mathbf{A} - \Pi_{\mathbb{S}}(\mathbf{A})\|_F \leq (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_K\|_F$. In the following theorem, we prove there exists an upper bound for ε when the algorithm SP-RD is used to select one sample. Please note that when $K = 1$, in Line 3, $\mathbf{U}_{\bar{k}}$ is empty and in Line 5, $\mathbf{E}_{\bar{k}} = \mathbf{A}$.

Theorem 1. Let E_1 denote the error achieved by the best first rank-1 approximation of full-rank data matrix \mathbf{A} , i. e., $\|\mathbf{A} - \mathbf{A}_1\|_F^2$. Further, let $\rho(\mathbf{A})$ denote rank-oneness measure of \mathbf{A} defined in Zaeemzadeh et al. (2019), and $\kappa(\mathbf{A})$ denote the condition number of matrix \mathbf{A} . Assume that N is less than M for the given large dataset \mathbf{A} . Also, let $\{\mathbf{s}_1\}$ denote the singleton containing the best sample selected by SP-RD algorithm. Then,

$$\|\mathbf{A} - \pi_{\mathbf{s}_1}(\mathbf{A})\|_F^2 \leq (1 + \varepsilon) E_1, \quad (4)$$

$$\text{where } \varepsilon = \frac{(\kappa(\mathbf{A})^2 - 1)(1 - \rho(\mathbf{A})^2)}{N - 1}.$$

Therefore, the smallest possible ε for the upper bound to be held can be set to the value found in the above equation. Before proceeding to establish an upper bound for K selected samples, we briefly discuss the established bound for one sample in the following remarks clarifying certain properties of the obtained bound in two specific instances,

Remark 1- When $\kappa(\mathbf{A}) = 1$, the data matrix is isometric, meaning that all samples are of equal importance. In this case, the projection error on the span of any sample is equal to the projection error on any spectral component. Thus, $\|\mathbf{A} - \pi_{\mathbf{s}_1}(\mathbf{A})\|_F = \|\mathbf{A} - \mathbf{A}_1\|_F$. In other words, the upper bound holds with correction factor $\varepsilon = 0$.

Remark 2- $\rho(\mathbf{A}) = 1$ means that all the energy of the samples selected is accumulated in the direction of the first eigenvector. In other words, the first sample is oriented towards the best rank-1 approximation of the data matrix, i.e., \mathbf{u}_1 . Therefore, $\varepsilon = 0$, and the upper bound holds tightly. This is trivial as the first sample turns out to be the best rank-1 approximation itself.

Theorem 2. Let $\|\mathbf{A} - \mathbf{A}_K\|_F^2$ denote the minimum error achieved by the best first rank- K approximation of data matrix \mathbf{A} . Further, let $\mathbb{S}_K = \{\mathbf{s}_1, \dots, \mathbf{s}_K\}$ be the set containing K samples selected by Alg 1. Then,

$$\|\mathbf{A} - \pi_{\mathbb{S}_K}(\mathbf{A})\|_F^2 \leq (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_K\|_F^2, \quad \text{for any } \varepsilon \geq \frac{\kappa(\mathbf{A})^2 - 1}{N - K}.$$

All proofs can be found in the supplementary material. The proposed SP-RD is the extended version of SP Joneidi et al. (2020) and iterative projection and matching (IPM) algorithm Zaeemzadeh et al. (2019). Theorem 2 shows us the theoretical reason behind the exhibited success of SP family algorithms for a very wide range of applications due to their simplicity and accuracy. SP and IPM algorithms has no parameter for fine tuning and parameter P in SP-RD does not need fine tuning and it can be set according to our accessible computational power. These favorable properties make the class of SP algorithms problem-independent and dataset-independent.

EXPERIMENTAL RESULTS

In this section, we apply our above theoretical results to perform learning tasks employing the selected samples of datasets based on their complexity, which is reflected in their self-rank. Our aim is to reduce training size of massive data such that accuracy of learning tasks is maintained. Applications of the class of SP algorithms are studied recently Joneidi et al. (2020); Zaeemzadeh et al. (2019). In this section, first a synthetic experiment is designed in order to estimate an oracle self-rank. Then, two new real-world applications are studied.

SELF-RANK ESTIMATION ON SYNTHETIC DATA

Since the computation of self-rank is a combinatorial problem, we need to synthesize a dataset with a known self-rank in order to evaluate sampling algorithms in Alg. 1 and Alg. 2. The known self-rank is assumed as a lower bound on the estimated self-rank using Alg. 2.

A synthetic full-rank dataset is generated with M samples in 200 dimensional space. First, we generate 20 linearly-independent samples whose target rank is equal to 15 with parameter $E_K = 0.1$ ². We call these 20 samples *base samples*. Then, $M - 20$ samples are generated by linear combination of base samples. Finally, all M generated samples are contaminated with noise in the null-space of base samples to result in a full rank dataset. Since base samples can be approximated by a rank-15 subspace, the whole dataset can be approximated by a rank-15 decomposition. However, those 20 base samples correspond to the self-rank of dataset. We test this dataset using different selection algorithms. An efficient algorithm summarizes the dataset to the base samples. Fig. 3 shows the probability of selection from the underlying base samples using different algorithms. We let all selection algorithms to sample 20 data. As can be seen, SP-RD algorithm successfully finds all base samples for up to 1000 generated samples. Increasing P (number of correlated samples) in the SP-RD algorithm results in a more accurate solution to Problem 3. We perform SP-RD with 200 iterations. Note that SP algorithm is equivalent to the SP-RD algorithm with $P = 1$. Table 2 shows the estimated value for the self-rank of the generated synthetic dataset. Setting $P \geq 8$ results in estimating the oracle value for the underlying self-rank. The desired E_K is set to 0.1. It is worthwhile to mention that random selection on average needs 67 samples to have the same accuracy as the projection error of the 20 underlying base samples. While SP-RD with $P = 8$ only needs 20 samples, i.e., base samples are selected intelligently.

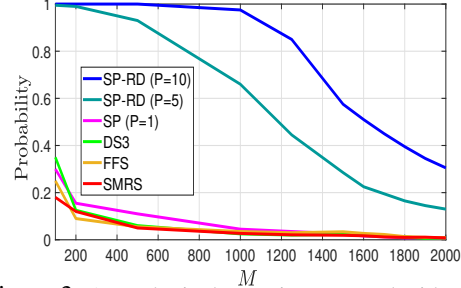


Figure 3: A synthetic dataset is generated with self-rank equal to 20 with $E_K = 0.1$. There are 20 base samples in the dataset which are unknown. This figure shows the probability of capturing the base samples in the sampled subset.

REPRESENTATIVE SELECTION FROM CMU MULTI-PIE DATASET

Here, CMU Multi-PIE Face Database is considered for representative selection Gross et al. (2010). A cropped version of this dataset is available online at ³, which contains 249 subjects with 13 poses, 20 illuminations, and 2 expressions. Thus, there are 520 images for each subject. Fig. 4 compares

²Please refer to Fig. 1 and Fig. 2 for definition of target rank and its relation to self-rank.

³https://drive.google.com/open?id=1QxNCh6vfNSZkod1Rg_zHLI1FM8WyXix4

Table 2: Estimated self rank for the described synthetic data using SP-RD algorithm with different values for parameter P . The expected estimated self-rank using random selection is computed as 67. It means on average 67 samples are needed to span the dataset as accurate as the span of 20 samples found by SP-RD algorithm.

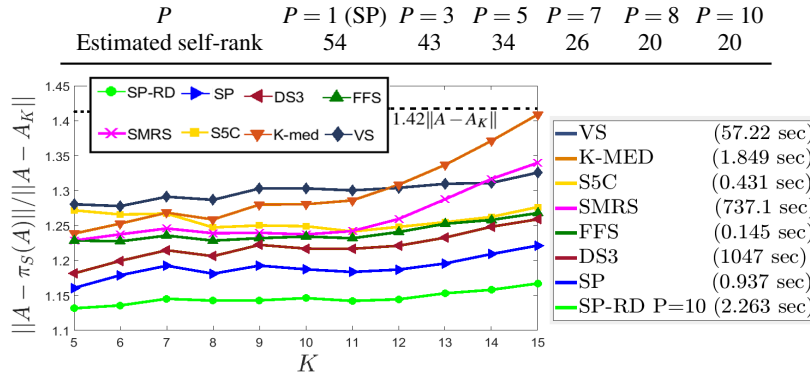


Figure 4: The averaged cost function of CSSP for selecting K samples from each class of Multi-pie face dataset with 130,000 images. The projection error is normalized by the projection error of the best rank- K approximation. SP-RD algorithm is compared with the state-of-the-art methods including with SP Joneidi et al. (2020), DS3 Elhamifar & Vidal (2013), FFS You et al. (2018), SMRS Elhamifar et al. (2012), S5C Matsushima & Brbic (2019), K-medoids Vijaya et al. (2004) and volume sampling Deshpande & Rademacher (2010b).

the performance of different state-of-the-art selection algorithms in terms of CSSP normalized projection error, which is defined as the CSSP cost function in equation 3 for a given selection method normalized by $\|A - A_K\|$. The normalized error is averaged for all 249 subjects. Parameter P is set as 10 for SP-RD algorithm and in order to select K samples, $5K$ iterations are performed. As it can be seen, the obtained approximation ratios are much better than the tightest theoretical bound in the literature which is $\sqrt{1+K}$. In practice, for multi-pie face dataset, the achieved AR using different algorithms is better than 1.42 for $K \leq 15$. Moreover, the proposed SP-RD reaches an AR around 1.15 which is the closest to the best possible AR which is 1.

ADAPTIVE SAMPLING FROM MNIST DATASET FOR CLASSIFICATION

To evaluate the effectiveness of our algorithm, we apply our algorithm on MNIST dataset classification task. In order to apply adaptive sampling, first MNIST data points are transferred to a feature space. A simple back-bone architecture is utilized for both feature selection and classification tasks to prevent architecture-specific bias. We train our feature net for classification task on Omniglot dataset Lake et al. (2011) and pick the last convolution layer as the features. We transform every image in MNIST training set into a vector of length 256 with the trained feature net. Our classifier also has the same architecture as our feature net and the only difference is that the last layer consists of only 10 neurons which is the number of classes in MNIST.

Figure 5 shows the result of these experiments. In adaptive sampling scenario using SP-RD algorithm, we consider the complexity of each class in MNIST training data determined by the introduced self-rank for each class. We compute the self-rank of each class using Alg. 2 alongside with optimal samples. Therefore, the number of selected samples are class dependent. This forms a non-uniform sampling which samples data for learning from each class in an adaptive fashion. We compare this to random selection, and the selection carried out by SP-RD with a fixed number of selected samples from each class. Moreover, K-medoids, SMRS, and dual volume sampling (DVS) Li et al. (2017) algorithms are compared. Splitting the sampling budget adaptively according to the introduced self-rank results in improvement of classification performance. This experiment is designed based on a generic architecture with a typical dataset. However, other selection algorithms do not provide a gain over random selection.

ADAPTIVE GRAPH SUMMARIZATION

A graph network is characterized by structural features such as degree distribution, average shortest-path length (ASPL) and the clustering coefficient. Calculating the ASPL of a large graph is memory space and computation intensive. Hence, we propose an alternative efficient method for computing the ASPL with a reasonable error. Instead of the shortest path between all pairs of vertices we compute

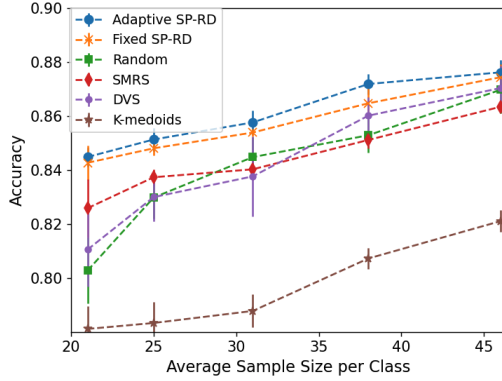


Figure 5: Comparison between the accuracy of MNIST classification task when a portion of each class is sampled. In this scenario, K-medoids performs even worse than random selection and the performance of SMRS and dual volume sampling are close to that of random selection. The vertical lines show a 95% confidence interval.

Table 3: Number of samples needed to achieve the threshold of 0.15 on the error of APSL on the Cora dataset.

Algorithms	Number of samples
SP-RD (P=10)	105
FFS You et al. (2018)	138
DS3 Elhamifar et al. (2016)	145
VSDeshpande & Rademacher (2010b)	147
IS Olivier Bachem (2017)	250
SP Joneidi et al. (2020)	140
FWCampbell & Broderick (2018)	384
MP Bo et al. (2011)	475
GIGA Campbell & Broderick (2018)	490

the ASPL between all the vertices and some selected vertices. Given a graph G we first select a subset of the vertices and then we exploit the measure of ASPL to evaluate the accuracy. Let G be the graph of the real world citation graph datasets, *Cora* with the set of vertices V ⁴. Further, let $\text{dist}(v_1, v_2)$ denote the shortest distance between v_1 and v_2 ($v_1, v_2 \in V$). Then, the error is defined as

$$\text{err} = \left| \frac{1}{|V|} \sum_{(i,j) \in V} \text{dist}(v_i, v_j) - \frac{1}{|S|} \sum_{i \in S, j \in V} \text{dist}(v_i, v_j) \right| \text{ where } |S| \text{ is the selected samples by the algorithms}$$

specified in the table. The more representative these selected vertices are the less error we get. More importantly, the less selected samples are adopted by each algorithm the faster the ASPL will be obtained. For a fair comparison between our algorithm and the state-of-the-art we consider a pre-defined threshold ($\text{err}_{th} = 0.15$) on the error of the approximation based on the topology of the graph and we observe each algorithm to see how many data each algorithm needs to satisfy that precision of error on the shortest path. The results of these experiments are shown in Table 3.

CONCLUSION

In this paper, we study the problem of subset selection, which has many applications in machine learning. The general goal is to select a subset from a large set of data such that their linear combination is able to target rank- K approximation. We have propounded a novel adaptive sampling technique from a given dataset for machine learning tasks. This sampling method is based on considering the spectral properties of the given data matrix. Our algorithm delivers a tight theoretical bound on the approximation ratio. The proposed algorithm enjoys a linear computational complexity and at the same time it provides an outstanding practical performance and a theoretical approximation guarantee. These favorable properties make our proposed algorithm a new paradigm in the literature of data sampling. The elongated proof is included in our supplementary material. We also present experiments on synthetic and real world datasets to demonstrate significant performance superiority to other sampling methods in different learning tasks.

⁴The dataset contain sparse bag-of-words feature vectors for each document and a list of citation links between documents. We treat the citation links as (undirected) edges and construct a binary, symmetric adjacency matrix A with 2,708 nodes and 5,429 edges.

Asymptotic Optimality of Self-Representative Low-Rank Approximation and Its Applications

Supplementary Material

In this supplementary document, we will show a constructive intuition about our proposed sampling. Then, the details of spectrum pursuit with residual descent (SP-RD) algorithm are presented. Finally, the proofs of the mentioned theoretical results are presented.

RELATION TO THE CONVENTIONAL SPECTRAL-BASED SAMPLING

Fifty years after Shannon work, in 90s researchers revisited Shannon theory for signals with special spectral patterns Unser (2000). A popular investigated spectral structure is union of low-rank subspaces Eldar & Mishali (2009) which resulted in compressed sensing, a low-rate sensing technique Davenport et al. (2011). In compressed sensing we measure a combination (sense) of samples and it is not appropriate for selecting a small number of actual samples. On the other hand, in the recent 20 years information systems are extensively developed based on recent advances in machine learning and big datasets. The present paper links the conventional spectral sampling approaches to sampling from now-days big datasets represented in a finite space for a machine learning task.

To get a constructive intuition about the Shannon sampling theorem, we cast the optimal Shannon sampling in terms of an optimization problem as follows,

$$\{t_k, f\} = \underset{t_k, f}{\operatorname{argmin}} \int_{-\infty}^{+\infty} \left(x(t) - \sum_{k=-\infty}^{+\infty} x(t_k) f(t, t_k) \right)^2 dt + \lambda \text{ (sampling period)} \quad (5)$$

The sampling period is the average of $|t_k - t_{k-1}|$ for all k . This problem aims to find sampling epochs t_k such that function f is able to recover the original signal accurately. Shannon theorem solves this problem using an assumption on the bandwidth of signal $x(t)$. Assume that the decomposition of signal $x(t)$ in terms of bases $\{\exp(jwt)\}$ is non-zero only for $|w| \leq B$, then the cost function of the first term in equation 5 is 0 using $t_k = k/(2B)$ as the sampled points and $f(t, t_k) = \operatorname{sinc}(2B(t - t_k))$ as the mixer function. Thus the achieved sampling period is $1/(2B)$. Problem equation 5 holds for signals in an infinite-dimension space.

However, practical data measurements result in datasets which are composed of a large number of vectors in a finite-dimensional space. Assume M data samples in an N -dimensional space are organized in matrix $\mathbf{A} = [\mathbf{a}_m] \in \mathbb{R}^{N \times M}$. Variable t in the continuous signal representation is substituted with integer variable m for $m = 1$ up to M . The goal is to sample few columns such that their mixture reconstructs all columns. Inspired by equation 5, Problem (1) in the main paper is achieved as follows,

$$\{\mathbb{S}^*, \mathbf{V}\} = \underset{\mathbb{S}, \mathbf{V}}{\operatorname{argmin}} \sum_{m=1}^M \|\mathbf{a}_m - \sum_{k \in \mathbb{S}} \mathbf{a}_k v_{mk}\|_2^2 + \lambda |\mathbb{S}|. \quad (6)$$

where λ is a parameter that regularizes the rate of sampling. Matrix \mathbf{V} projects back selected samples to the original dataset similar to function f in Eq. equation 5. Similar to the spectral conditions for a perfect signal reconstruction, we establish an upper bound for the performance of data sampling according to the spectral properties of matrix \mathbf{A} . In the continuous signal reconstruction we need $2B$ samples for each $\Delta t = 1$. However, Self-rank indicates the minimum required samples for a dataset in a finite dimensional space.

Shannon sampling theorem requires a minimum number of samples which is a function of spectral properties of the original function. As it is shown in the main paper, we also pursuit the same fashion.

SPECTRUM PURSUIT WITH RESIDUAL DESCENT (SP-RD) ALGORITHM

Projection of all the data onto the subspace spanned by the K columns of \mathbf{A} , indexed by \mathbb{T} , i.e., $\pi_{\mathbb{T}}(\mathbf{A})$, can be expressed by a rank- K factorization, $\mathbf{U}\mathbf{V}^T$. In this factorization, $\mathbf{U} \in \mathbb{R}^{N \times K}$, $\mathbf{V}^T \in \mathbb{R}^{K \times M}$, and \mathbf{U} includes the K columns of \mathbf{A} , indexed by \mathbb{T} which are normalized to have unit

Algorithm 3: Spectrum Pursuit with Residual Descent

Require: A, P and K **Output:** \mathbb{S} **1: Initialization:** $\mathbb{S} \leftarrow$ A random subset of $\{1, \dots, M\}$ with $|\mathbb{S}| = K$ $\{\mathbb{S}_k\}_{k=1}^K \leftarrow$ Partition \mathbb{S} into K singletons.

iter=0

while the stopping criterion is not met

2: $k = \text{mod}(\text{iter}, K) + 1$ 3: $U_{\bar{k}} = \text{normalize column}(A_{\mathbb{S} \setminus \mathbb{S}_k})$ 4: $V_{\bar{k}} = A^T U_{\bar{k}} (U_{\bar{k}}^T U_{\bar{k}})^{-1}$ 5: $E_{\bar{k}} = A - U_{\bar{k}} V_{\bar{k}}^T$ 6: $u_k = \text{first left singular-vector of } E_{\bar{k}}$ 7: $\Omega \leftarrow P$ most correlated columns of $E_{\bar{k}}$ with u_k 8: $\Omega = \Omega \cup \mathbb{S}_k$ 9: $\mathbb{S}_k \leftarrow \text{argmin}_{k \in \Omega} \|E_{\bar{k}} - a_k a_k^T E_{\bar{k}}\|_F$ 10: $\mathbb{S} \leftarrow \bigcup_{k'=1}^K \mathbb{S}_{k'}$

11: iter=iter+1

end while

length. Therefore, optimization problem (6) and the problem in Eq. (2) of the main paper can be restated as

$$\underset{U, V}{\text{argmin}} \|A - UV^T\|_F^2 \text{ s.t. } u_k \in \mathbb{A}, \quad (7)$$

where, $\mathbb{A} = \{\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_M\}$, $\tilde{a}_m = a_m / \|a_m\|_2$, and u_k is the k^{th} column of U . It should be noted that U is restricted to be a collection of K normalized columns of A , while there is no constraint on V . Since Problem (7) involves a combinatorial search and is not easy to tackle, we modify (7) into two consecutive problems. The first sub-problem relaxes the constraint $u_k \in \mathbb{A}$ in (7) to a moderate constraint $\|u\| = 1$, and the second sub-problem reimposes the underlying constraint. These sub-problems are formulated as

$$(u, v) = \underset{u, v}{\text{argmin}} \|A - \underbrace{U_{\bar{k}} V_{\bar{k}}^T}_{E_{\bar{k}}} - uv^T\|_F^2 \text{ s.t. } \|u\| = 1, \quad (8a)$$

$$\mathbb{S}_k = \underset{m}{\text{argmax}} |u^T \tilde{e}_m|. \quad (8b)$$

Here \mathbb{S}_k is a singleton containing the index of the k^{th} selected data point. Matrix $U_{\bar{k}}$ is obtained by removing the k^{th} column of U . Matrix $V_{\bar{k}}$ is defined in a similar manner. Subproblem (8a) is equivalent to finding the first left singular vector (LSV) of $E_{\bar{k}} \triangleq A - U_{\bar{k}} V_{\bar{k}}^T$. The constraint $\|u\| = 1$ keeps u on the unit sphere to remove scale ambiguity between u and v . Moreover, the unit sphere is a superset for \mathbb{A} and keeps the modified problem close to the recast problem (7). After solving for u (which is not necessarily one of our data points), we find the data point that matches u the most (makes the smallest angle with u) in (8b).

The best minimizer for equation 8a w.r.t. u is the first LSV of $E_{\bar{k}}$. However, restriction to data samples does not imply that the most correlated sample with the first LSV is necessarily the best minimizer for equation 8a, although in most cases the most correlated sample with the first LSV is the best minimizer. In order to find the best sample that minimizes equation 8a at each iteration, we propose to collect a few samples that are correlated with the first LSV and compute residual error for these samples. Let Ω denote the set of P most correlated samples with the first LSV. The modified version of the above problem can be written as follows.

$$(u, v) = \underset{u, v}{\text{argmin}} \|A - U_{\bar{k}} V_{\bar{k}}^T - uv^T\|_F^2 \text{ s.t. } \|u\| = 1, \quad (9a)$$

$$\Omega = \underset{|\Omega|=P}{\text{argmax}} \sum_{c \in \Omega} |u^T \tilde{e}_c|, \quad (9b)$$

$$\mathbb{S}_k = \underset{c \in \Omega, v}{\text{argmin}} \|A - U_{\bar{k}} V_{\bar{k}}^T - uv^T\| \text{ s.t. } u = \tilde{a}_c. \quad (9c)$$

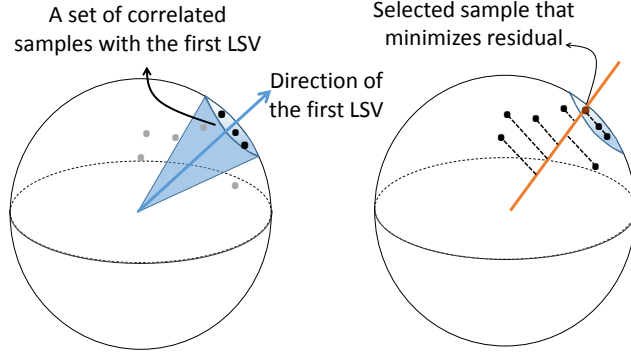


Figure 6: (left) a small subset of samples which are correlated with the first left SV are grouped (Eq. (5a) and (5b) and Line 6,7 & 8 of SPRD algorithm). (right) the sample which is the best minimizer for eq. equation 9c is selected (Eq. (5c) and Line 9 of SP-RD algorithm).

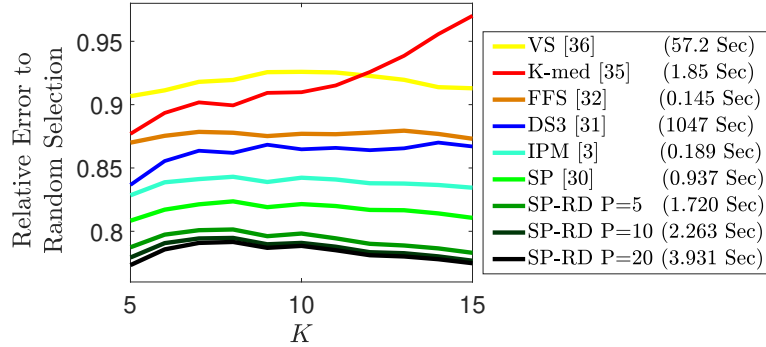


Figure 7: The projection error of selection normalized by the projection error of random selection. Regarding the reported time $K = 10$ is considered. The experiment setup is identical to that of Fig. 4 in the main paper.

To make the error monotonically decreasing, as a sufficient condition for convergence, we can include the index of the previously selected sample in for Ω for updating the k^{th} selected sample. This step is performed in Line 8 of SP-RD algorithm. In order to select K samples, the lower bound for selection cost function is $\|\mathbf{A} - \mathbf{A}_K\|$ in which \mathbf{A}_K is the best rank- K approximation of \mathbf{A} . Since the cost function is monotonically decreasing, and it is lower bounded, the modified algorithm is convergent. Alg. 3 and Fig. 6 show the steps in SP-RD algorithm. Alg. 1 is also mentioned in the main paper and it is repeated here for a more clear presentation.

Approximation ratio is a constant greater than 1 since it is normalized by the best (lowest) possible projection error. In Fig. 7 errors are normalized by projection error of random selection. Since random selection is a rough selection method, the ratio is usually (not necessarily) smaller than 1. As it can be seen increasing P improves the accuracy of selection. However, the gain after $P = 10$ is marginal.

CONVERGENCE ANALYSIS

At each iteration of SP-RD, a new sample is selected only if the resulted residual error decreases (Alg. 1 (SP-RD), line 9). This way the error is non-increasing. The error is also lower bounded by $\|\mathbf{A} - \mathbf{A}_K\|_F^2$ since it is the least error that K vectors can achieve as the projection error. These two conditions guarantee that the algorithm converges and quality of the selected subset always improves or remains the same over iterations. As you can see in Fig. 8 the cost function for SP algorithm is not necessarily decreasing over iterations. However, our proposed SP-RD algorithm is decreasing, since convergence is guaranteed.

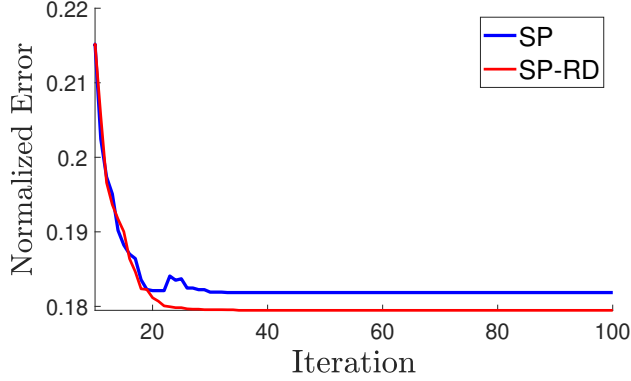


Figure 8: Selecting 10 representatives from the first 10 classes of Multi-pie dataset. Each class has 520 samples and the error trajectory is depicted in order to show converge behavior of SP-RD algorithm and its comparison with the plain SP algorithm. Both are initialized using IPM algorithm.

Proofs of Propositions and Theorems

Lemma 3. *Zaeemzadeh et al. (2019)* Let \mathbf{A} denote an $N \times M$ matrix. Let σ_1 and \mathbf{u} denote the first singular value and the corresponding left singular vector of \mathbf{A} , respectively. Then, there exists at least one column in \mathbf{A} such that the absolute value of its inner product with \mathbf{u} is greater than or equal to $\frac{\sigma_1}{\sqrt{M}}$. Hence,

$$\max_k |\mathbf{u}^T \mathbf{a}_k| \geq \frac{\sigma_1}{\sqrt{M}}. \quad (10)$$

The following proposition states a lower bound on the maximum of the absolute value of the correlation between columns of a matrix and \mathbf{u} , when data are normalized on the unit sphere.

Proposition 4. *Zaeemzadeh et al. (2019)* Assume the columns of \mathbf{A} are normalized to lie on the unit sphere. There exists at least one data point, \mathbf{a}_i , such that the correlation coefficient between \mathbf{a}_i and the first left singular vector of \mathbf{A} is greater than or equal to $\frac{\sigma_1}{\|\mathbf{A}\|_F}$.

Definition 1. Rank-oneness measure of a rank R matrix \mathbf{A} with singular values $\sigma_1, \sigma_2, \dots, \sigma_R$ is defined as

$$\rho(\mathbf{A}) = \sqrt{\frac{\sigma_1^2}{\sum_{r=1}^R \sigma_r^2}} = \frac{\sigma_1}{\|\mathbf{A}\|_F}.$$

Proof of Theorem 1: Given the SVD of a full-rank data matrix \mathbf{A} is represented as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. The most correlated column of \mathbf{A} with \mathbf{u}_1 , which is denoted by \mathbf{s}_1 and its normalized version can be written as follows:

$$\mathbf{s}_1 = \frac{\mathbf{u}_1 + \sum_{i>1} \alpha_i \mathbf{u}_i}{\sqrt{1 + \sum_{i>1} \alpha_i^2}} \quad (11)$$

where \mathbf{u}_i is i th column of \mathbf{U} . The projection error for the first normalized selected sample \mathbf{s}_1 can be cast as follows:

$$\begin{aligned} & \|\mathbf{A} - \mathbf{s}_1 \mathbf{s}_1^T \mathbf{A}\|_F^2 \\ &= \left\| \mathbf{A} - \frac{\mathbf{u}_1 \mathbf{u}_1^T \mathbf{A} + \sum_i \alpha_i^2 \mathbf{u}_i \mathbf{u}_i^T \mathbf{A}}{1 + \sum_i \alpha_i^2} \right\|_F^2 \\ &= \sigma_1^2 \left(1 - \frac{1}{1 + \sum_i \alpha_i^2}\right)^2 + \sum_{i>1} \sigma_i^2 \left(1 - \frac{\alpha_i^2}{1 + \sum_{i>1} \alpha_i^2}\right)^2. \end{aligned} \quad (12)$$

Now, assume $L = 1 + \sum_i \alpha_i^2$. In order to compute the desired bound for ϵ in the format presented in Theorem 1, it only suffices to show that the last term in the above equation is upper bounded by $E_1 = \sum_{i>1} \sigma_i^2 (1 + \epsilon)$. Rewriting and cancelling common terms from both sides leads to,

$$\sigma_1^2 \left(1 - \frac{1}{L}\right)^2 \leq \sum_{i>1} \sigma_i^2 \left(\frac{2\alpha_i^2}{L} + \epsilon - \frac{\alpha_i^4}{L^2}\right) \quad (13)$$

To establish inequality equation 13, it is enough to relax it to a more tractable looser inequality. To this end, we shrink the right hand side by substituting all σ_i s for $i > 1$ with σ_N , and since $\frac{\alpha_i^2}{L} \leq 1$, we have $\frac{\alpha_i^2}{L} \geq \frac{\alpha_i^4}{L^2}$. Hence, we remove the $\frac{\alpha_i^2}{L} - \frac{\alpha_i^4}{L^2} \geq 0$ from right hand side. Next, we enlarge the left hand side by ignoring the power factor 2 on the left hand side of inequality equation 13 as $1 - \frac{1}{L} < 1$. Therefore, the inequality to be satisfied which guarantees the inequality equation 13 will be held can be written as follows:

$$\kappa(\mathbf{A})^2 \left(1 - \frac{1}{L}\right) \leq \sum_{i>1} \left(\frac{\alpha_i^2}{L} + \varepsilon\right) \quad (14)$$

Knowing that $\sum_{i>1} \frac{\alpha_i^2}{L} = 1 - \frac{1}{L}$, the inequality equation can be reduced to:

$$(\kappa(\mathbf{A})^2 - 1) \left(1 - \frac{1}{L}\right) \leq (N - 1)\varepsilon \quad (15)$$

According to lemma 3, we observe that $1 - \frac{1}{L} \leq 1 - \rho(\mathbf{A})^2$. Again this means we lift the left hand side and in order to satisfy the original upper bound, it suffices that the following inequality to be held:

$$\frac{(\kappa(\mathbf{A})^2 - 1)(1 - \rho(\mathbf{A})^2)}{N - 1} \leq \varepsilon \quad (16)$$

First, we need the following lemma.

Lemma 5. *Let us define $\kappa^* = \max_k \kappa(\mathbf{E}_{\bar{k}})$. κ^* is upper bounded as $\kappa^* \leq \kappa(\mathbf{A})$.*

Proof: It is known that removing columns of a data matrix, the largest singular value decreases and the smallest singular value increases and as a result, the condition number of a given matrix reduces Queiró (1987). In addition, throughout iterations of Alg. 3, the data passes through projections on null-space of selected samples. Projection does not affect singular values but may make some zero. As a result, null-space projections are non-increasing operators on condition number. A series of column deletion and null-space projection therefore, lead to decreasing (non-increasing) condition number. Thus, $\kappa^* \leq \kappa(\mathbf{A})$. ■

Proof of Theorem 2: The subspace of interest $\mathbf{E}_{\bar{k}}$ in Alg. 3 Line 5 (also as defined in Eq. equation 8) is obtained by projection on the null-space of the previously selected samples. Therefore, due to null-space projection property at each iteration, every vector in the resulted sub-spaces after projection is orthogonal to the previously selected samples. In other words, if $\mathbf{s} \in \text{Range}(\mathbf{E}_{\bar{k}})$, then $\mathbf{s} \perp \text{span}\{\mathbf{s}_1, \dots, \mathbf{s}_{k-1}\}$. The residual-based progression of Alg. 3 selects a novel sample to capture the highest information in projected version of dataset to the null-space of previously selected samples. Thus, representation errors in subsequent iterations fall in orthogonal sub-spaces and this fruitful linear algebraic property of Alg. 3 permits us to sum over error terms as in Line 9 of Alg. 3 in each iteration. The upper bound for such error terms is established in Theorem 1. Owing to orthogonality property, the Frobenius norm of selected subset representation error can be written as sum of iteration error terms to yield:

$$\|\tilde{\mathbf{E}}_K\|_F^2 = \sum_{k=1}^K \|\mathbf{Z}_k\|_F^2 \quad (17)$$

where $\|\mathbf{Z}_k\|_F^2$ is the minimum error value achieved in Line 9 of Alg. 3. Moreover, $\|\tilde{\mathbf{E}}_K\|_F^2 = \|\mathbf{A} - \pi_{\mathbb{S}}(\mathbf{A})\|_F^2$. Let $E_{\bar{k}}$ show the best rank-1 representation error for $\mathbf{E}_{\bar{k}}$ in Eq. (4). $\|\mathbf{Z}_k\|$ can be upper bounded as $(1 + \varepsilon_k)E_{\bar{k}}$ following the same approach as in Theorem 1 where,

$$\varepsilon_k = \frac{(\kappa(\mathbf{E}_{\bar{k}})^2 - 1)(1 - \rho(\mathbf{E}_{\bar{k}})^2)}{N - k}. \quad (18)$$

To watch for the zero singular values appearing in subsequent iterations making the matrix $\mathbf{E}_{\bar{k}}$ singular, we truncate the nonzero singular values so that the condition number remains finite and meaningful. This requires a reduction in the number of columns in matrices used to select sample indices $\mathbf{U}_{\bar{k}}$ and $\mathbf{E}_{\bar{k}}$. This can be done by deleting the index of previously selected sample leading to singularity after null-space projection.

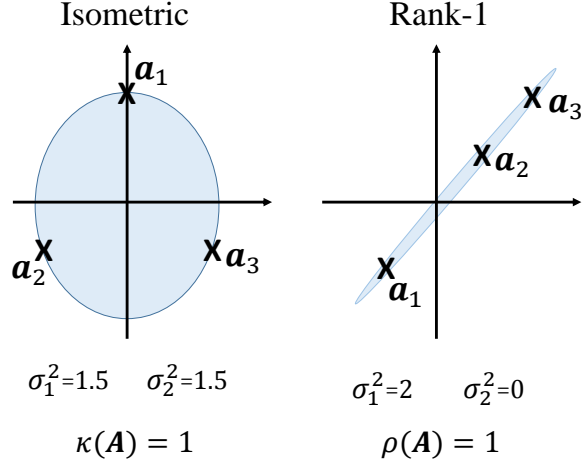


Figure 9: Assume three points in a two dimensional space represented by a 2×3 matrix $\mathbf{A} = [\mathbf{a}_1; \mathbf{a}_2; \mathbf{a}_3]$. Two extreme cases of eigenvectors' orientation are shown where the self-representative approximation error tightly sticks to the best rank-1 approximation error according to Theorem 1 ($\varepsilon = 0$). In both cases, there is no preference on samples to be selected and any random selection provides the same projection error.

Also, the denominator appearing in the lower bound for ε_k is $N - k$ (equal to the new reduced dimension which can be derived in a similar manner as observed in Theorem 1 for the k -th iteration).

Inserting κ^* in equation 18, one can immediately find that the smallest possible coefficient ε_K holds in:

$$\varepsilon_k \leq \frac{\kappa^{*2} - 1}{N - K}, \quad \forall 1 \leq k \leq K \quad (19)$$

Using equation 19 and equation 17 the proof is completed immediately.

Two asymptotic cases are discussed in the main paper in Remark 1 and Remark 2. Fig. 9 shows the geometrical interpretation of these two special conditions. As it can be seen, for an isometric dataset, there is no preference for samples to be selected. In this scenario, projection error of selecting 1 sample is equal to projection error of the best rank-1 approximation. Thus, the tightest bound will be reached and $\varepsilon = 0$. Similarly, when the dataset is rank-1, there will be the same story since span of any sample will be equal to the span of best rank-1 approximation. Thus, $\varepsilon = 0$ which the tightest upper bound is touched.

REFERENCES

- Rémi Bardenet and Odalric-Ambrym Maillard. A note on replacing uniform subsampling by random projections in mcmc for linear regression of tall datasets. In *2015. fflhal-01248841f*, 2015.
- Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pp. 2115–2123, 2011.
- Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pp. 968–977. SIAM, 2009.
- Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.
- Trevor Campbell and Tamara Broderick. Bayesian coresets construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, 2018.

- Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P. Woodruff. Algorithms for p low-rank approximation. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 806–814. JMLR.org, 2017.
- Michael Mahoney Christos Boutsidis and Petros Drinea. An improved approximation algorithm for the column subset selection problem. In *In Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009.
- Ali Çivril. Column subset selection problem is ug-hard. *Journal of Computer and System Sciences*, 80(4):849–859, 2014.
- K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *in Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, pp. 81–90, 2013.
- Chen Dan, Hong Wang, Hongyang Zhang, Yuchen Zhou, and Pradeep K Ravikumar. Optimal analysis of subset-selection based L_p low-rank approximation. In *Advances in Neural Information Processing Systems*, pp. 2537–2548, 2019.
- Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems*, pp. 235–242, 2006.
- Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.
- Mark A Davenport, Marco F Duarte, Yonina C Eldar, and Gitta Kutyniok. Introduction to compressed sensing. *preprint*, 93(1):2, 2011.
- Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pp. 329–338, 2010a.
- Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 329–338. IEEE, 2010b.
- Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John W. Paisley, and David M. Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 2729–2738, 2017.
- Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.
- J. A. Duersch and M. Gu. True blas-3 performance qrcp using random samplin. In *arXiv preprint arXiv:1509.06820*, 2015.
- Yonina C Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.
- Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.
- Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1600–1607. IEEE, 2012. ISBN 9781467312264. doi: 10.1109/CVPR.2012.6247852.
- Ehsan Elhamifar, Guillermo Sapiro, and S. Shankar Sastry. Dissimilarity based sparse subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2182–2197, 2016. ISSN 01628828. doi: 10.1109/TPAMI.2015.2511748.
- Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 5 2010. ISSN 0262-8856. doi: 10.1016/J.IMAVIS.2009.08.002. URL <https://www.sciencedirect.com/science/article/pii/S0262885609001711>.

- M. Gu. Subspace iteration randomization and singular value problems. In *SIAM Journal on Matrix Analysis and Applications*, volume 37, pp. A1139–A117, 2015.
- Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank revealing qr factorization. In *SIAM Journal on Scientific Computing*, 1996.
- Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4080–4088, 2016.
- M. Gu J. W. Demmel, L. Grigori and H. Xiang. Communication avoiding rank revealing qr factorization with column pivoting. In *SIAM Journal on Matrix Analysis and Applications*, volume 36, pp. 55–89, 2015.
- Mohsen Joneidi, Saeed Vahidian, Ashkan Esmaeili, Weijia Wang, Nazanin Rahnavard, Bill Lin, and Mubarak Shah. Select to better learn: Fast and accurate deep learning using data selection from nonlinear manifolds. *Computer Vision and Pattern Recognition, CVPR 2020. IEEE Conference on*, 2020.
- Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.
- Michael Langberg and Leonard J. Schulman. Universal epsilon-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pp. 598–607, 2010.
- Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems*, pp. 5038–5047, 2017.
- Shin Matsushima and Maria Brbic. Selective sampling-based scalable sparse subspace clustering. In *Advances in Neural Information Processing Systems*, 2019.
- Andreas Krause Olivier Bachem, Mario Lucic. Practical coreset constructions for machine learning. *Thesis at Department of Computer Science, ETH Zurich*, 2017.
- M.W. Mahoney P. Drineas and S. Muthukrishnan. Polynomial time algorithm for column-row based relative-error low-rank matrix approximation. In *Report, DIMACS*, 2006.
- Saurabh Paul, Malik Magdon-Ismail, and Petros Drineas. Column selection via adaptive sampling. In *Advances in neural information processing systems*, pp. 406–414, 2015.
- Michael W Mahoney Petros Drineas and S Muthukrishnan. Relative-error cur matrix. In *SIAM Journal on Matrix Analysis and Applications*, volume 3, pp. 844–881, 2008.
- João Filipe Queiró. On the interlacing property for singular values and eigenvalues. *Linear Algebra and Its Applications*, 97:23–28, 1987.
- Yaroslav Shitov. Column subset selection is np-complete. *arXiv preprint arXiv:1701.02764*, 2017.
- Zhao Song, David Woodruff, and Peilin Zhong. Average case column subset selection for entrywise ℓ_1 -norm loss. In *Advances in Neural Information Processing Systems*, pp. 10111–10121, 2019a.
- Zhao Song, David Woodruff, and Peilin Zhong. Towards a zero-one law for column subset selection. In *Advances in Neural Information Processing Systems*, pp. 6120–6131, 2019b.
- and Nathan Srebro Srinadh Bhojanapalli, Behnam Neyshabur. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 43, 2016.
- Michael Unser. Sampling-50 years after shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.
- P A Vijaya, M Narasimha Murty, and D K Subramanian. Leaders–Subleaders: An efficient hierarchical clustering algorithm for large data sets. *Pattern Recognition Letters*, 25(4):505–513, 2004.

Chong You, Chi Li, Daniel P Robinson, and René Vidal. Scalable exemplar-based subspace clustering on class-imbalanced data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 67–83, 2018.

Alireza Zaeemzadeh, Mohsen Joneidi, Nazanin Rahnavard, and Mubarak Shah. Iterative Projection and Matching: Finding Structure-Preserving Representatives and Its Application to Computer Vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5414–5423, 2019. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zaeemzadeh_Iterative_Projection_and_Matching_Finding_Structure-Preserving_Representatives_and_Its_Application_CVPR_2019_paper.html.