ABSTRACT

We propose a novel technique for dataset summarization by selecting 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 express all dataset 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 most informative samples in a linear order of complexity w.r.t. the data size. We derive a new upper bound for the approximation ratio (AR), the ratio of obtained projection error using selected samples to the best rank-K approximation error. The best previously known AR for self-representative low-rank approximation was presented in ICML 2017, which was further improved by the bound $\sqrt{1+K}$ reported in NeurIPS 2019. Both of these bounds are obtained by brute force search, which is not practical, and these bounds depend solely on $K$, the number of selected samples. In contrast, we describe an adaptive AR that takes into consideration the spectral properties and spikiness measure of the original dataset, $A \in \mathbb{R}^{N \times M}$. In particular, our performance bound is proportional to the condition number $\kappa(A)$. Our derived AR is expressed as $1 + (\kappa(A)^2 - 1)/(N - K)$, which approaches 1 and is optimal in two extreme spectral distribution instances. In the worst case, AR is shown not to exceed $1 + K$ for the proposed method. Our proposed algorithm enjoys linear complexity w.r.t. 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, learning a classifier with representative data, and open-set identification.

INDEX TERMS

Column subset selection, data summarization, graph node selection, low-rank, meta learning, open-set identification, self-rank, spectrum pursuit, SVD.

I. INTRODUCTION

Low-rank approximation is one of the fundamental workhorses of machine learning and optimization. It is a building block for data compression algorithms such as principal component analysis (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. 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 [3]. In spectral-based sampling, the goal is to keep...
the spectrum (spectral decomposition) of representatives as close as possible to the spectrum of the original data. Shannon sampling [4], the most theoretically strong sampling scheme also follows the strategy of keeping the spectrum of the original and sampled function equal to each other. A reconstruction guarantee is then established for signal sampling in an infinite-dimensional space based on spectral properties. In addition to sampling, feature selection, data selection (as in active learning for annotating unlabeled samples), and data summarization (such as video summarization) are amongst other important applications where the representatives must be part of original samples [5].

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 \( A \in \mathbb{R}^{N \times M} \). Let \( S \) be a set of integers between 1 and \( M \), which, when used to index the columns of \( A \), define a collection of samples. We define a matrix of the form \( V = [v_{mk}] \) which projects back from the sampled to the original data. We consider the following optimization problem for finding the optimal samples and back-projection:

\[
\{S^*, V\} = \arg\min_{S,V} \sum_{m=1}^{M} \|a_m - \sum_{k \in S} a_{k,v_{mk}}\|_2^2 + \lambda |S|,
\]

where \( \lambda \) is a parameter that regularizes the sampling rate. We refer to the cardinality of the set \( S^* \) as the self-rank of matrix \( A \) in this paper. Substituting \( a_k \) for the inner summation with an arbitrary vector \( u_k \in \mathbb{S}^{N-1} \) simplifies this problem to the truncated singular value decomposition (SVD). In this case, \( |S| \) is simplified to the conventional definition of rank for matrix \( A \). 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 Eq. 1 results in a self-representative low-rank approximation. This problem is related to the column subset selection problem (CSSP) [6], a problem that is known to be NP-hard [7], [8] and subject of ongoing research [2], [9], [10]. Our research group proposed a novel optimization technique to solve the problem of CSSP which is called iterative projection and matching (IPM) [3]. Recently, the IPM algorithm is extended to a more accurate and non-greedy version which is referred as spectrum pursuit (SP) [5]. Despite superior performance of SP compared to the state-of-the-art (SOTA) in a wide range of applications, it is not provably convergent. Moreover, the performance bound of SP algorithm is not investigated. In the present work, we will show theoretically that why IPM and SP family algorithms are working excellently. In addition to convergence guarantee, another important question will be answered in the present paper which is “how many samples are enough in order to describe a dataset?” The later question is addressed by developing the concept of self-rank. 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 contributions of this paper can be summarized as follows:

- We introduce Spectrum Pursuit with Residual Descent (SP-RD), a developed version of SP that is employed for estimating the self rank and summarizing the dataset by selecting the most informative samples.
- We prove an upper bound for the class of spectrum pursuit algorithms performance, and show that in two special cases, the proposed bound for AR sticks to the tightest bound corresponding to the best rank-K approximation.
- We show that the SP-RD algorithm leads to improved performance compared to the SOTA in a number of applications, including graph node sampling with shortest-path criterion, summarizing data for training a classifier without losing performance, and open-set identification.

II. RELATED WORK AND MOTIVATION

There is a large line of work on data summarization in machine learning which are mainly studied under the context of CSSP and volume sampling (VS). See [1], [2], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] 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 shrinks while preserving its original statistical properties. Based on this observation, [11] studied the reduction in size of a large dataset using random linear projection. On a similar note, [12], [13], [18] constructed a weighted subset of a large dataset, the Bayesian coreset, for a wider class of Bayesian models.

Many other efforts have been devoted to data summarization and coreset construction algorithms. Examples include pivoted QR factorization, non-negative matrix factorization, and matrix subset selection, which are under the umbrella of CSSP [2], [9], [10], [23], [24], [25], [26], 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 [19], sketching algorithms [20] and low-rank approximation problems [21]. Authors in [22] develop a randomized QR with column pivoting (RQRCP), where random sampling (RNDS) is used for column selection. [27], [28] improved the running time of CSSP with different sampling schemes while preserving similar approximation bounds. The CSSP method and its variants are also applied and studied under various different settings. A comprehensive summary of these algorithms and their theoretical guarantees is available in Table 1 in [6]. Most of these algorithms can be roughly categorized into two branches. One branch of algorithms are based on rank-revealing QR (RRQR) decomposition [29] which is related to reduction in E-optimal sense [30]. It has been proved in [6] that RRQR is nearly optimal in terms
of residue norm. On the other hand, sampling based methods [31] 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 [32]. Let us denote the best rank-$K$ approximation of matrix $A$ by $A_K$. The data matrix $A$ is approximated in polynomial-time in [33] and a lower bound on the required number of samples as $O(k \log K + K/\epsilon)$, where $\epsilon$ is the coefficient of the relative error, has been obtained in [33], [34]. In other words, a subset of data is selected such that the projection error of all samples on it is less than $(1 + \epsilon)\|A - A_K\|_F$. For a smaller value of $\epsilon$, we need more selected samples to achieve the desired projection error. A tighter bound for the number of required samples is introduced in [35]. They show that $O(k/\epsilon)$ columns contain a subspace which approximates $A$ with coefficient error of $(1 + \epsilon)$. Here, we introduced our approach based on recent advances on data sampling. Our proposed bound is the first work in the literature that targets rank-$K$ approximation using only $K$ samples while $\epsilon$ 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 [5]. In the present paper we will show theoretically that why SP outperforms SOTA 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 $\kappa(A)$) of the matrix $A$. In this paper, a theoretical upper bound is established which depends on spectral properties of $A$ 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.

III. JOINT SELF-RANK ESTIMATION AND SAMPLING

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

$$S^* = \arg \min_{S, V} \|S\| \text{ s.t. } \sum_{m=1}^{M} \|a_m - \sum_{k \in S} a_{vk} v_{mk}\|^2_2 \leq E_K \|A\|_F^2, \quad (2)$$

where $E_K \in [0, 1]$ is the projection error corresponding to the best rank-$K$ approximation of data normalized to $\|A\|_F$. Parameter $K$ is a user specified target rank. As $K$ increases, more samples are needed in the set $S$ in order to obtain the desired error. The cardinality of set $S^*$ for a specified $K$ is denoted by $S_K(A)$ and it is called the self-rank of dataset $A$ 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 vs. 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 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 5. In other words, there are 15 samples in the dataset which are able to reconstruct all the dataset with an error less than 25%. The normalized error, $E_K$, is a user-specified parameter, which can be set to any value between 0 and 1. The target rank of matrix $A$ corresponding to a desired projection error, $E_K$, is defined as the minimum integer $K$ that satisfies $(\sum_{k=K+1}^N \sigma_k^2)/\|A\|_F^2 \leq E_K$. The $n$th singular value of matrix $A$ is denoted by $\sigma_n$.
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 approximating $A_K$ is the goal of sampling. Our achieved AR is compared with Brute-force [2], Improved CSSP [6], Volume sampling [33], CUR Decomposition [31], and Near-optimal [35].

| Algorithm              | Selected samples | Upper bound                      | Complexity                  |
|------------------------|------------------|----------------------------------|-----------------------------|
| Brute-force             | $K$              | $\sqrt{1 + K} \|A - A_K\|_F$    | brute force $(K^2)$          |
| Improved CSSP          | $K$              | $O \left( \log(K) \right) \|A - A_K\|_F$ | $O \left( \min \{ M^2 N, M^2 N \} \right)$ |
| Volume sampling        | $O(K \log(K)/\varepsilon^2)$ | $O \left( \frac{1 + \varepsilon}{\varepsilon} \|A - A_K\|_F \right)$ | $O \left( \frac{MK^2}{\varepsilon^2} \right) \|A - A_K\|_F$ |
| CUR Decomposition      | $O(2K/\varepsilon)$ | $O \left( \frac{1 + \varepsilon}{\varepsilon} \|A - A_K\|_F \right)$ | $O \left( \frac{NK M}{\varepsilon^2} \right)$ |
| Near-optimal           | $K$              | $O \left( \frac{1 + \varepsilon}{\varepsilon} \|A - A_K\|_F \right)$ | $O \left( \frac{NK M}{\varepsilon^2} \right)$ |
| Ours                   | $K$              | $O \left( \frac{1 + \varepsilon}{\varepsilon} \|A - A_K\|_F \right)$ | $O \left( \frac{NK M}{\varepsilon^2} \right)$ |

The self-rank of matrix $A$ with parameter $K$ is equal to the minimum number of columns in $A$ such that their span approximates $A$ by an error less than the error of best rank-$K$ approximation, i.e., the smallest size for set $S$ that holds

$$
\|A - A_S A_S^\dagger A\|_F \leq \sum_{n=K+1}^{N} \sigma_n^2.
$$

Matrix $A_S$ refers to sampled columns of $A$, and $A_S^\dagger$ is the Moore–Penrose inverse of $A_S$.

IV. SPECTRUM PURSUIT WITH RESIDUAL DESCENT (SP-RD)

Solving Problem (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 for $|S| = K$, and check if the constraint in Problem (2) is satisfied for a desired error threshold.

$$
S^* = \arg\min_S \|S\| \left\{ \sum_{m=1}^{M} \|a_m - \sum_{k \in S} a_k v_{mk} \|^2 \right\} \text{ s.t. } |S| \leq K. \quad (3)
$$

The solution for $S$ does not guarantee that the error in the constraint of Problem (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 $A$ with parameter $K$. If the error is higher than the desired threshold indicated by $E_K$, the cardinality of set $S$ is increased by 1, and Problem (3) is solved with a new value for $K$.

Spectrum Pursuit (SP) Algorithm is proposed in [5] for solving (3). Inspired by SP, a more accurate solver for (3) is proposed in this paper which is called Spectrum Pursuit with Residual Descent (SP-RD).

Projection of all the data onto the subspace spanned by the $K$ columns of $A$, indexed by $S$, i.e., $\pi_S(A)$, can be expressed by a rank-$K$ factorization, $U V^T$. In this factorization, $U \in \mathbb{R}^{N \times K}$, $V^T \in \mathbb{R}^{K \times M}$, and $U$ includes the $K$ columns of $A$, indexed by $S$ which are normalized to have unit length. Therefore, (3) can be restated as

$$
\arg\min_{U,V} \|A - UV^T\|_F \text{ s.t. } u_k \in \mathbb{A}, \quad (4)
$$

where, $\mathbb{A}$ = $\{a_1, a_2, \ldots, a_M\}$, $a_m = a_m/\|a_m\|_2$, 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 (4) involves a combinatorial search and is not easy to tackle, we modify (4) into two consecutive problems. The first sub-problem relaxes the constraint $u_k \in \mathbb{A}$ in (4) to a moderate constraint $\|u\| = 1$, and the second sub-problem reimagines the underlying constraint. These sub-problems are formulated as

$$
(u, v) = \arg\min_{u,v} \|A - U_k V_k^T - u v^T\|_F^2 \text{ s.t. } \|u\| = 1, \quad (5a)
$$

$$
S_k = \arg\max_m \|u^T \tilde{e}_m\|. \quad (5b)
$$

Here, $S_k$ is a singleton containing the index of the $k$th selected data point. Matrix $U_k$ is obtained by removing the $k$th column of $U$. Matrix $V_k$ is defined in a similar manner. Subproblem (5a) is equivalent to finding the first left singular vector (LSV) of $E_k \equiv A - U_k V_k^T$. 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 $A$ and keeps the modified problem close to the recast problem (4). 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 (5b).

The best minimizer for (5a) w.r.t. $u$ is the first LSV of $E_k$. However, restriction to data samples does not imply that the most correlated sample with the first LSV is necessarily the best minimizer for (5a), 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 (5a) 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 $\Omega$ denote the set of $P$ most correlated samples with the first LSV. The modified version of the above problem can be written as follows.

$$
(u, v) = \arg\min_{u,v} \|A - U_k V_k^T - u v^T\|_F^2 \text{ s.t. } \|u\| = 1, \quad (6a)
$$

$$
\Omega = \arg\max_{|\Omega|=P} \sum_{c \in \Omega} |u^T \tilde{e}_c|, \quad (6b)
$$

$$
S_k = \arg\min_{c \in \Omega, v} \|A - U_k V_k^T - u v^T\| \text{ s.t. } u = \tilde{a}_c. \quad (6c)
$$
To make the error monotonically decreasing, as a sufficient condition for convergence, we can include the index of the previously selected sample in $\Omega$ 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 $\|A - A_K\|$ in which $A_K$ is the best rank-$K$ approximation of $A$. Since the cost function is monotonically decreasing, and it is lower bounded, the modified algorithm is convergent. Alg. 1 and Fig. 3 show the steps in SP-RD algorithm. SP-RD is an extension for IPM algorithm which is an incremental algorithm [3].

V. 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 $\|A - A_K\|^2_F$ 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. 4 the cost function for SP algorithm is not necessarily decreasing over iterations. However, our proposed SP-RD algorithm is decreasing, since convergence is guaranteed.

SP-RD simplifies to IPM if set $S$ is initialized by a empty set and $P$ is set to 1. IPM performs only $K$ iterations and $E_T = A$ at the first iteration. As it will be shown in our experiments, SP-RD is a more efficient solver for CSSP compared to SOTA. However, theoretical analysis of IPM is more straightforward. In other words, SP-RD is a more robust implementation of IPM and is privileged with convergence thanks to the descent property (The global solution is not necessarily obtained as for sampling based methods which are of combinatorial nature, the global solution is inevitably reached via brute-force search.)

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 to Alg. 2 as adaptive SP-RD. In Line 1 of the algorithm, $K$ can be initialized using truncated SVD easily. Moreover, $\Pi_S(A)$ is the projection operator on the subspace spanned by columns of $A$ indexed by set $S$. Mathematically, it is equal to $A_S(A_S^t A_S)^{-1} A_S^t A$. It is noteworthy that the computational complexity of the SP-RD algorithm is $O(MN + K^2 N + MNP)$ per iteration, which depends on the computational burden parameter $P$. Note that we only need the first singular vector which can be done using fast methods such as power iteration method. 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 $M$. Fig. 5, compares the run time of our proposed method with other baselines.

The SOTA upper bound for the minimum required number of samples is introduced as $O(\sqrt{K})$ [35], 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 $S$ to ensure that $\|A - \Pi_S(A)\|_F \leq (1 + \epsilon)\|A - A_K\|_F$. In the following theorem, we prove there exists an upper bound for $\epsilon$ when the algorithm SP-RD is used to select one sample. Please note that when $K = 1$, in Line 3, $U_T$ is empty and in Line 5, $E_T = A$. This configuration simplifies the SP-RD algorithm to our previous
Algorithm 1: Spectrum Pursuit With Residual Descent (SP-RD)

**Require:** $A$, $P$, and $K$

**Output:** $\mathcal{S}$

1: **Initialization:**
   - $\mathcal{S} \leftarrow$ a random subset of $\{1, \ldots, M\}$ with $|\mathcal{S}| = K$
   - $[\mathcal{S}_k]_{k=1}^K \leftarrow$ Partition $\mathcal{S}$ into $K$ singletons.
   - $\text{iter} = 0$
   - while the stopping criterion is not met
     2: $k = \text{mod}(\text{iter} + 1)$
     3: $U_\mathcal{S} = \text{normalize column}(A_{\mathcal{S}\setminus\mathcal{S}_k})$
     4: $V_k = A^T U_\mathcal{S} (U_\mathcal{S}^T U_\mathcal{S})^{-1}$
     5: $E_k = A - U_\mathcal{S} V_k^T$
     6: $u_k = \text{first left singular-vector of } E_k$
     7: $\Omega \leftarrow P$ most correlated Col. of $E_k$ with $u_k$
     8: $\Omega = \Omega \cup \mathcal{S}_k$
     9: $\mathcal{S}_k \leftarrow \text{argmin}_{k \in \Omega} \| E_k - a_k a_k^T \|_F$
     10: $\mathcal{S} \leftarrow \cup_{k=1}^K \mathcal{S}_k$
     11: $\text{iter} = \text{iter} + 1$
   - end while

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

**Require:** dataset $A$, $P$, and the desired sampling error $\epsilon \in [0, 1]$

**Output:** set $\mathcal{S}$

1: $K \leftarrow$ smallest integer $K$ such that $\| A - A_K \|_F^2 \leq \epsilon \| A \|_F^2$
   - $E_k = \epsilon$
   - $\alpha = K$
   - $\mathcal{S} \leftarrow$ SP-RD ($A$, $P$, $\alpha$
   - **While** $\| A - \Pi_\mathcal{S}(A) \|_F^2 > E_K \| A \|_F^2$
     2: $\alpha \leftarrow \alpha + 1$
     3: $\mathcal{S} \leftarrow$ SP-RD ($A$, $P$, $\alpha$

work, i.e., IPM. Our theoretical contribution is based on the IPM algorithm. However, the later works, i.e., SP, and SP-RD are more sophisticated implementations of the same approach as IPM algorithm.

From here on, we assume that the matrix of interest is either full-rank or the zero singular values are truncated in the corresponding economic SVD representation so that the condition number is bounded and is not infinitely large as a result of division by small singular values.

**Definition 1:** spikiness measure of a rank $R$ matrix $A$ with singular values $\sigma_1, \sigma_2, \ldots, \sigma_R$ is defined as

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

**Theorem 2:** Let $E_1 = \| A - A_1 \|_F^2$ denote the error achieved by the best first rank-1 approximation of full-rank data matrix $A$. Further, let $\rho(A)$ denote spikiness measure of $A$ defined in [3], and $\kappa(A)$ denote the condition number of matrix $A$. Assume that $N$ is less than $M$ for the given large dataset $A$. Also, let $\{s_1\}$ denote the singleton containing the best sample selected by SP-RD algorithm with $P = 1$. Then, $\| A - \pi_{s_1}(A) \|_F \leq (1 + \epsilon)E_1$, where

$$\epsilon = \min\{\rho(A)^2 (1 - \rho(A)^2), \frac{(\kappa(A)^2 - 1)(1 - \rho(A)^2)}{N - 1}\}. \quad (7)$$

**Proof of Theorem 2:** Given the SVD of a full-rank data matrix $A$ is represented as $A = U \Sigma V^T$, the most correlated column of $A$ with $u_1$, which is denoted by $s_1$ and its normalized version can be written as follows:

$$s_1 = u_1 + \sum_{i>1} \alpha_i u_i$$

where $u_i$ is $i$th column of $U$. The projection error for the first normalized selected sample $s_1$ can be cast as follows:

$$\| A - s_1 s_1^T A \|_F = \| A - u_1 u_1^T A + \sum_{i>1} \alpha_i^2 u_i u_i^T A \|_F^2$$

$$= \sigma_1^2 \left(1 - \frac{1}{\sum_{i>1} \sigma_i^2}\right)^2$$

$$+ \sum_{i>1} \sigma_i^2 \left(1 - \frac{\sigma_i^2}{\sum_{i>1} \sigma_i^2}\right)^2, \quad (9)$$

which is obtained using orthogonality of the cross terms. Now, assume $L = 1 + \sum \sigma_i^2$. In order to compute the desired bound for $\epsilon$ in the format presented in Theorem 2, 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 identical terms from both sides leads to,

$$\sigma_1^2 \left(1 - \frac{1}{L}\right)^2 \leq \sum_{i>1} \sigma_i^2 \left(\frac{2\sigma_i^2}{L} + \epsilon - \frac{\sigma_i^4}{L^2}\right). \quad (10)$$

To establish inequality (10) for some $\epsilon$ values, we relax it to a more tractable inequality. We consider a tighter inequality by decreasing the right hand side and increasing the left hand side of this inequality. $\epsilon$ values working for tighter version generalize to the main inequality as well.

Since $\frac{\sigma_i^2}{L} \leq 1$, we have $\frac{2\sigma_i^2}{L} \geq \frac{\sigma_i^4}{L^2}$. Hence, we remove the $\frac{2\sigma_i^2}{L} - \frac{\sigma_i^4}{L^2}$ from right hand side. Next, we enlarge the left hand side by ignoring the power factor 2 for $1 - \frac{1}{L} < 1$ in inequality (10). Therefore, the resulted inequality to be satisfied for $\epsilon$ will be

$$\sigma_1^2 \left(1 - \frac{1}{L}\right) \leq \sum_{i>1} \sigma_i^2 \left(\epsilon + \frac{\sigma_i^2}{L}\right). \quad (11)$$

We use Eq. (11) to establish two upper bounds for AR. The first bound is achieved by substituting all $\sigma_i$s for $i > 1$ with $\sigma_N$. Hence, (11) reduces to

$$\kappa(A)^2 \left(1 - \frac{1}{L}\right) \leq \sum_{i>1} \left(\frac{\sigma_i^2}{L} + \epsilon\right). \quad (12)$$
Knowing that \( \sum_{i>1} \frac{\sigma_i^2}{L} = 1 - \frac{1}{L} \), the inequality (12) can be written as:

\[
(\kappa(A)^2 - 1)(1 - \frac{1}{L}) \leq (N - 1)\epsilon 
\]

To continue and simplify the bound, we mention the following lemma.

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

\[
\max_k |u^T a_k| \geq \frac{\sigma_1}{\sqrt{M}}. \tag{14}
\]

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

**Proposition 4:** [3] Assume the columns of \( A \) are normalized to lie on the unit sphere. There exists at least one data point, \( a_i \), such that the correlation coefficient between \( a_i \) and the first left singular vector of \( A \) is greater than or equal to \( \rho(A) \).

According to lemma 3 and proposition 4, we observe that \( 1 - \frac{1}{L} \leq 1 - \rho(A)^2 \). By replacing this, we lift the left hand side in order to satisfy the original upper bound, it suffices to find an \( \epsilon \) working for the following inequality:

\[
\frac{(\kappa(A)^2 - 1)(1 - \rho(A)^2)}{N - 1} \leq \epsilon \tag{15}
\]

The other bound is obtained by setting \( a_i \) to 0 in (10). Following the same procedure, we will have

\[
\frac{\rho(A)^2}{1 - \rho(A)^2}(1 - \frac{1}{L})^2 \leq \epsilon \tag{16}
\]

Using the result of lemma 3 for upper bounding \( 1 - \frac{1}{L} \), the inequality of interest is held if the following tighter inequality is satisfied:

\[
\rho(A)^2(1 - \rho(A)^2) \leq \epsilon. \tag{17}
\]

Therefore, the two \( \epsilon \) values which hold the two obtained bounds tightly can work for our interest. We set \( \epsilon = \min\{\rho^2(1 - \rho^2), (\frac{\kappa(A)^2 - 1}{N - 1})\} \). For small condition number, the second bound tends to zero. Also, for large \( \rho \) (close to 1), the overall bound is small. The worst case for the overall upper bound is the maximum of first term which is 0.25. It is worth noting that for rank deficient data, \( N \) can be interchangeably replaced with \( N - N_0 \) for some \( N_0 > 0 \). The minimum of the two is considered as the suggested bound. Therefore, the smallest possible \( \epsilon \) 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 \( \kappa(A) = 1 \), 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, \( \|A - \pi_{S_1}(A)\|_F = \|A - A_1\|_F \). In other words, the upper bound holds with correction factor \( \epsilon = 0 \).

**Remark 2:** \( \rho(A) = 1 \) 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., \( u_1 \). Therefore, \( \epsilon = 0 \), and the upper bound holds tightly. This is trivial as the first sample turns out to be the best rank-1 approximation itself.

Fig. 6 shows the geometrical interpretation of the two extreme cases in Remark 1 and Remark 2. As 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 \( \epsilon = 0 \). 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, \( \epsilon = 0 \) which the tightest upper bound is touched. After observing the optimality of the presented method in two extreme cases for the dataset spectral distribution, we proceed to present a general bound for all dataset distributions in the following theorem. Please note that condition number is the ratio of the largest singular value to the smallest non-zero singular value.

**Theorem 5:** Let \( \|A - A_K\|_F \) denote the minimum error achieved by the best rank-1 approximation of data matrix \( A \). Further, let \( S_K = \{s_1, \ldots, s_K\} \) be the set containing \( K \) samples selected by IPM algorithm which is a simplified version of Alg. 1 with \( P = 1 \). Then,

\[
\|A - \pi_{S_K}(A)\|_F \leq (1 + \epsilon)\|A - A_K\|_F,
\]

for any \( \epsilon \geq \min\{\frac{\kappa(A)^2 - 1}{N - K}, 0.25\} \).

**Proof of Theorem 5:** The subspace of interest \( E_q \) in Alg. 1 Line 5 (also as defined in Eq. (5)) 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 \( s \in Range(E_q) \), then \( s \perp \text{span}\{s_1, \ldots, s_{q-1}\} \). The residual-based progression of Alg. 1 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. 1 permits us to sum over error terms as in Line 9 of Alg. 1 in each iteration. The upper bound for such error terms is established in Theorem 2. Owing to orthogonality property, the Frobenius norm of selected subset representation error
can be written as sum of iteration error terms to yield:
\[
\|E_K\|_F^2 = \sum_{k=1}^{K} \|Z_k\|_F^2, \tag{18}
\]
where \(\|Z_k\|_F^2\) is the minimum error value achieved in Line 9 of Alg. 1. Moreover, \(\|E_K\|_F^2 = \|A - \pi_{\mathcal{B}}(A)\|_F^2\). Let \(E_K\) show the best rank-1 representation error for \(E_K\) in Eq. (4). \(\|Z_k\|\) can be upper bounded as \((1 + \epsilon_k)E_k^\star\) following the same approach as in Theorem 2 where,
\[
\epsilon_k = \min \left\{ \frac{(\kappa(E_k)^2 - 1)(1 - \rho(E_k)^2)}{N - k}, \right. \\
\left. \rho(E_k)^2(1 - \rho(E_k)^2) \right\}. \tag{19}
\]

To watch for the zero singular values appearing in subsequent iterations making the matrix \(E_{\mathcal{F}}\) 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 \(U_k\) and \(E_k\). This can be done by deleting the index of previously selected sample leading to singularity after null-space projection.

Also, the denominator appearing in the lower bound for \(\epsilon_k\) is \(N - k\) (equal to the new reduced dimension which can be derived in a similar manner as observed in Theorem 2 for the \(k\)-th iteration).

**Lemma 6:** Let us define \(k^* = \max_k \kappa(E_k)^\star\). \(k^*\) is upper bounded as \(k^* \leq \kappa(A)\).

**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 [36]. In addition, throughout iterations of Alg. 1, the data passes through projections on null-space of selected samples. Projection does not affect singular values but may turn some into zero. As a result, null-space projections are non-increasing operators on condition number. Therefore, a series of column deletion and null-space projection leads to decreasing (non-increasing) condition number. Thus, \(k^* \leq \kappa(A)\).

The second term is the upper bound for the quadratic term \(\rho^2(1 - \rho^2)\) and therefore 0.25. Inserting \(k^*\) in (19), one can immediately find that the smallest possible coefficient \(\epsilon_k\) holds in:
\[
\epsilon_k = \min \left\{ \frac{k^* - 1}{N - K}, 0.25 \right\} \quad \forall 1 \leq k \leq K \tag{20}
\]
A toy example is illustrated in Fig. 7 which validates the efficacy of the extracted bound for some synthetic data. Using (20), lemma 6, and (18) the proof is completed immediately. Again, we reiterate that for rank deficient matrices \(N\) can be replaced with some \(N - N_0\) without loss of generality.

The proposed SP-RD is the extended version of SP [5] and iterative projection and matching (IPM) algorithm [3]. Theorem 5 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 have no parameter to be fine tuned. Parameter \(P\) in SP-RD does not need fine tuning, and can be set according to our accessible computational power. These favorable properties make the class of SP algorithms problem-independent and also dataset-independent.

**VI. EXPERIMENTAL RESULTS**

In this section, we apply the aforementioned 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. In this section, first a synthetic experiment is designed in order to estimate an oracle self-rank. Then, four new real-world applications are studied.

**Datasets.** i) CMU Multi-PIE Face Dataset. [37] We used a cropped version of this dataset. This dataset contains 249 subjects with 13 poses, 20 illuminations, and 2 expressions. Thus, there are 520 images for each subject. ii) Omniglot [38]. The Omniglot data set is designed for developing more human-like learning algorithms. It contains 1623 different handwritten characters from 50 different alphabets. Each of the 1623 characters was drawn online via Amazon’s Mechanical Turk by 20 different people. Each image is paired with stroke data, a sequences of \([x,y,t]\) coordinates with time \((t)\) in milliseconds. iii) MNIST [39]. The MNIST dataset of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. iv) Cora [40]. The dataset contains 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.
S. Vahidian et al.: Spectrum Pursuit With Residual Descent for Column Subset Selection Problem

**FIGURE 7.** Evaluation of our derived upper bound for approximation ratio. This result is obtained for a 50 × 1000 data with condition number of 1.25.

**FIGURE 8.** The intuition behind the generation of a dataset of a known rank. (a) The base samples that are 20 independent samples are shown. These samples are generated such that they can be approximated by a 15 dimensional subspace. (b) Green samples are the base samples and red samples are linear combination of green samples that are contaminated by a noise from null space of base samples. For this dataset, we know that there exist 20 samples such that their linear combination is able to make a subspace such that it is as accurate as the best 15 dimensional low-rank approximation of all data.

**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 accurately as 20 samples found by SP-RD algorithm.

| $P$ | Estimated self-rank |
|-----|---------------------|
| 1   | 54                  |
| 3   | 43                  |
| 5   | 34                  |
| 7   | 26                  |
| 8   | 20                  |
| 10  | 20                  |

**B. REPRESENTATIVE SELECTION FROM CMU MULTI-PIE DATASET**

Here, CMU Multi-PIE Face Database is considered for representative selection [37]. Fig. 10 compares the performance of different SOTA selection algorithms in terms of CSSP normalized projection error, which is defined as the CSSP cost function in (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 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. 9 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. SP-RD is performed for 200 iterations. Note that SP algorithm is equivalent to the SP-RD algorithm with $P = 1$. 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 9.** 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.

...
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, i.e., 1.

C. 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 backbone 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 [38] and pick the last convolution layer 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 11 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 scenario where data are sampled 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, comparison to K-mediods, SMRS, and dual volume sampling (DVS) [46] algorithms is provided. 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.

D. OPEN-SET IDENTIFICATION BASED ON SELF-RANK

In this experiment, we show how employing adaptive SP-RD gains advantage over random sampling, and fixed rate sampling towards open-set identification. Due to closed-set nature of training procedure for deep neural networks, such models tend to label all samples as belonging to one the classes while a sample may not belong to the observed classes. The capability of rejecting such samples and classifying them as coming from the open set is known as open-set identification [47], [48], [49], [50]. Alg. 3, describes the self-rank based open-set identification procedure. MNIST dataset is used as the closed-set for training. The open set samples are from Omniglot dataset. The ratio of Omniglot to MNIST test samples is $1 : 1$ (10,000 from each) same as the simulation scenario in [49]. A backbone classifier with high-precision (CNN) architecture [51] is trained on MNIST as for step 1 in Alg. 3. The error values resulted from projection (regression) of open and closed-set onto selected representative of each class in step 4 of Alg. 3 differ significantly compared to those of projecting onto the entire dataset (due to effect of out-distribution samples). Results of macro-averaged F1-score [52] for proposed method with different selection methods are listed in Table 3. We observe that the adaptive sampling based on self-rank outperforms random sampling and SP-RD with fixed number of selected samples (without considering the structure of each MNIST training class). In order to carry out a fair comparison, the number of selected samples from each class is set to be the average of the selected samples in the adaptive sampling based on their class-wise self-rank. It has been shown in [5] that the selection-based open-set identification scheme (SOSIS) outperformed SOTA in terms of the projection error for testing samples. In this experiment, we have made SOSIS adaptive in terms of the samples to be selected depending on the self-rank and self-expressiveness of each MNIST classes.
TABLE 3. Performance of sampling-based open-set identification.

| Algorithm       | Random Sampling | Fixed SP-RD | Adaptive SP-RD |
|-----------------|-----------------|-------------|----------------|
|                 | 95.11%          | 95.73%      | 96.33%         |

Algorithm 3: Self-Rank Based Open-Set Identification

Require: $A^X$ (closed-set training data), and $A^Y = \{a_p^Y\}_{p=1}^P$ (test set).

1. Train a classifier on $A^X$ on $H$ classes.
2. Evaluate self-rank $K_h$ for each class $h$ in $A^X$.
3. $S_h \leftarrow$ set of $K_h$ selected samples of class $h$ in $A^X$.
4. $\ell(p) \leftarrow$ label $a_p^Y$ using trained classifier in Step 1 ($\forall p$).
5. $err(p) = \|a_p^Y - \pi_{S(p)}(a_p^Y)\|_2$ ($\forall p$).
6. $c_1$, $c_2 \leftarrow$ perform kmeans on $err$ with 2 centroids.

Output: open-set $\{a_p^Y | err(p) \geq c_1 + c_2\}$.

E. 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 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 $V$. 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 $err = \frac{1}{|S|} \sum_{(v_i, v_j) \in V} \text{dist}(v_i, v_j) - \frac{1}{|S|} \sum_{i \in S, j \in \overline{V}} \text{dist}(v_i, v_j)$ where $|S|$ is the selected samples by the algorithms specified in Table 4. The more selected vertices chosen as these representatives, the less error is obtained. 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 SOTA we consider a pre-defined threshold ($err_{th} = 0.15$) 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 4.

VII. RELATION TO THE CONVENTIONAL SPECTRAL-BASED SAMPLING

Fifty years after Shannon’s work, in 90s researchers revisited Shannon theory for signals with special spectral patterns [4]. A popular investigated spectral structure is union of low-rank subspaces [56] which resulted in compressed sensing, a low-rate sensing technique [57]. 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 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,

$$\{t_k, f\} = \arg\min_{t_k, f} \int |(x(t) - \sum_{k = -\infty}^{+\infty} x(t_k) f(t, t_k) )|^2 dt + \lambda (\text{sampling period})$$

(21)

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 $\{\exp(j\omega t)\}$ is non-zero only for $|\omega| \leq B$, then the cost function of the first term in (21) is 0 using $t_k = k/(2B)$ as the sampled points and $f(t, t_k) = \sin(2B(t - t_k))$ as the mixer function. Thus the achieved sampling period is $1/(2B)$. Problem (21) 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 $A = [a_m] \in \mathbb{R}^{N \times M}$. 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 (21), the joint problem of self-rank estimation and sampling (earlier presented in Problem (21)) can be achieved as follows,

$$\{S^*, V\} = \arg\min_{S, V} \sum_{m=1}^{M} \|a_m - \sum_{k \in S} a_k v_{mk}\|_2^2 + \lambda |S|.$$ (22)

where $\lambda$ is a parameter that regularizes the rate of sampling. Matrix $V$ projects back selected samples to the original dataset similar to function $f$ in Eq. (21). 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 $A$. 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 suggests a minimum number of samples which is a function of spectral properties of the original function. As it is shown in this paper, we also pursue the same approach based on Eigen decomposition.

**VIII. CONCLUDING REMARKS AND DISCUSSION**

In this paper, we study the problem of subset selection. 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. In spite of a large body of work on the subset-selection algorithm and low rank approximation problem, many fundamental questions remain unresolved. Probably one of the most open questions is: “How the selection algorithm can determine the number of needed representative samples” or “how many samples are enough in order to describe a matrix of data (dataset)?” Despite superior performance of prior state-of-the-art methods like IPM [3], and SP [5], in these algorithms the number of selected samples is given; they are not provably convergent; no performance bound has been provided.

In the present work, leveraging SP we have propounded a novel adaptive sampling technique which depends on the spectral properties of the original dataset, $A \in \mathbb{R}^{N \times M}$. In our proposed approach the achieved bound is proportional to the condition number $\kappa(A)$. This is while, the state-of-the-art self-representative low-rank approximation methods are obtained either by brute force search which is not practical or the obtained bounds depend upon $K$, the number of selected data.

We proved performance guarantees by making use of the self-rank concept which showed that the proposed adaptive sampling improves upon continued sampling. We also present experiments on synthetic and real world datasets to demonstrate significant performance superiority to other sampling methods in different learning tasks.

We expect the outcome of our proposed algorithms and theoretical results to have a significant impact on a wide range of applications. This is due to an ever-increasing generation and collection of data in almost all areas, such as computer vision, machine learning, health care systems, surveillance systems, sensor networks, vehicular networks, smart grids, and IoT, to just name a few. In all these applications, it will be crucial to reduce an enormous amount of data to a much smaller subset of representatives. It will be also very important to efficiently learn critical features of data. Our proposed approach has direct impact on active learning, few-shot learning, self-supervised learning and data compression which are not explored here.

The introduced concept of self-rank is a useful tool which can be employed in other areas of data science in order to establish tighter theoretical guarantees. Specifically, spectral properties such as self-rank and condition number can be investigated in more details to achieve better performance bounds and constructive algorithms based on theoretical supports. The underlying model of our work is a union of linear subspaces. In the future work, more sophisticated models can be considered especially, latent spaces of neural networks which is a union of non-linear manifolds.

Some potential concerns/limitations of our work is that even though the assumptions made in this paper are realistic and compatible with the prior arts, the value of $N$ could be much less than the number of data and also $K$ could be much smaller than $N$.

**ACKNOWLEDGMENT**

The views, findings, opinions, and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF. (Saeed Vahidian, Mohsen Joneidi, and Ashkan Esmaeili contributed equally to this work.)

**REFERENCES**

[1] F. Chierichetti, S. Gollapudi, R. Kumar, S. Lattanzi, R. Panigrahy, and D. P. Woodruff, “Algorithms for lp low-rank approximation,” in *Proc. 34th Int. Conf. Mach. Learn. (PMLR)*, vol. 70, 2017, pp. 806–814.

[2] C. Dan, H. Wang, H. Zhang, Y. Zhou, and P. K. Ravikumar, “Optimal analysis of subset-selection based lp low-rank approximation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2537–2548.

[3] A. Zaeemzadkh, M. Joneidi, N. Rahnavard, and M. Shah, “Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5414–5423.

[4] M. Unser, “Sampling-50 years after Shannon,” *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, Apr. 2000.

[5] M. Joneidi, S. Vahidian, A. Esmaeili, W. Wang, N. Rahnavard, B. Lin, and M. Shah, “Select to better learn: Fast and accurate deep learning using data selection from nonlinear manifolds,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7819–7829.

[6] C. Boutsidis, M. W. Mahoney, and P. Drineas, “An improved approximation algorithm for the column subset selection problem,” in *Proc. 20th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2009, pp. 968–977.

[7] Y. Shitov, “Column subset selection is NP-complete,” 2017, arXiv:1701.02764.

[8] A. Cevir, “Column subset selection problem is UG-hard,” *J. Comput. Syst. Sci.*, vol. 80, no. 4, pp. 849–859, Jun. 2014.

[9] Z. Song, D. Woodruff, and P. Zhong, “Towards a zero-one law for column subset selection,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6120–6131.

[10] Z. Song, D. Woodruff, and P. Zhong, “Average case column subset selection for entrywise $l_1$-norm loss,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10111–10121.

[11] R. Bardenet and O.-A. Maillard, “A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets,” *Tech. Rep. Ilrhal-01248841f*, 2015.

[12] A. B. Dieng, D. Tran, R. Ranganath, J. W. Paisley, and D. M. Blei, “Variational inference via $l^1$-norm minimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 2729–2738.

[13] T. Campbell and T. Broderick, “Bayesian coreset construction via greedy iterative geometric ascent,” in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 698–706.

[14] M. Langberg and L. J. Schulman, “Universal epsilon-approximators for integrals,” in *Proc. 21st Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, Austin, TX, USA, Jan. 2010, pp. 598–607.

[15] S. Dasgupta, “Two faces of active learning,” *Theor. Comput. Sci.*, vol. 412, no. 19, pp. 1767–1781, Apr. 2011.

[16] S. Dasgupta, “Coarse sample complexity bounds for active learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 235–242.
C. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Tsvetko, J. Tang, and H. Liu, “Feature selection: A data perspective,” ACM Comput. Surv., vol. 50, no. 6, pp. 1–45, Jan. 2018.

J. H. Huggins, T. Campbell, and T. Broderick, “Coresets for scalable Bayesian logistic regression,” in Proc. Adv. Neural Inf. Process. Syst., Barcelona, Spain, 2016, pp. 4081–4089.

K. L. Clarkson and D. P. Woodruff, “Low rank approximation and regression in input sparsity time,” in Proc. 45th Annu. ACM Symp. Theory Comput. (STOC), 2013, pp. 81–90.

J. W. Demmel, L. Grigori, M. Gu, and H. Xiang, “Communication avoiding rank-revealing QR factorization with column pivoting,” SIAM J. Matrix Anal. Appl., vol. 36, pp. 55–89, Jan. 2015.

M. Gu, “Subspace iteration randomization and singular value problems,” SIAM J. Sci. Comput., vol. 37, no. 3, pp. A1139–A1173, Jan. 2015.

J. A. Duerchs and M. Gu, “True BLAS-3 performance QRCP using random sampling,” 2015, arXiv:1509.06820.

S. Paul, M. Magdon-Ismail, and P. Drineas, “Column selection via adaptive sampling,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 406–414.

Z. Ji, G. Meng, D. Huang, X. Yue, and B. Wang, “NNMFBS: A NMF-based feature selection method in identifying pivotal clinical symptoms of hepatocellular carcinoma,” Comput. Math. Methods Med., vol. 2015, Oct. 2015, Art. no. 846992.

X. Xie, X. Gu, Y. Li, and Z. Li, “K-size partial reduct: Positive region optimization for attribute reduction,” Knowl.-Based Syst., vol. 228, Sep. 2021, Art. no. 107253.

E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, “Feature selection for classification using principal component analysis and information gain,” Exp. Syst. Appl., vol. 174, Jul. 2021, Art. no. 114765.

A. Deshpande, I. L. Hadamacher, S. S. Vempala, and G. Wang, “Matrix approximation and projective clustering via volume sampling,” Theory Comput., vol. 2, no. 12, pp. 225–247, 2006.

A. Deshpande and S. S. Vempala, “Adaptive sampling and fast low-rank matrix approximation,” in Proc. 9th Int. Workshop Approximation Algorithms Combinat. Optim., Approximation, Randomization, Combinat. Optim. Algorithms Techn. (Lecture Notes in Computer Science), vol. 4110, J. Diaz, K. Jansen, J. D. P. Rolim, and U. Zwick, Eds, 2006, pp. 292–303.

M. Gu and S. C. Eisenstat, “Efficient algorithms for computing a strong rank revealing OR factorization,” SIAM J. Sci. Comput., vol. 17, no. 4, pp. 848–869, Jul. 1996.

M. Joneidi, A. Zaeemhaddeh, B. Shahrashi, G.-J. Qi, and N. Raahnabadv, “E-optimal sensor selection for comprehensive sensing-based purposes,” IEEE Trans. Big Data, vol. 6, no. 1, pp. 51–65, Mar. 2020.

M. M. P. Drineas and S. Muthukrishnan, “Relative-error CUR matrix,” SIAM J. Matrix Anal. Appl., vol. 3, no. 2, pp. 844–881, 2008.

A. N. S. Srinadh Bhojanapalli and B. N. S. guarantee generative classifiers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 753–757.

K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis. The Netherlands: Springer, Oct. 2016, pp. 630–645.

Z. C. Lipton, C. Elkan, and B. N. S. guarantee generative classifiers,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases. France: Springer, Sep. 2014, pp. 225–239.

E. Elhamifar, G. Sapiro, and S. S. Sastry, “Dissimilarity based sparse subset selection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 11, pp. 2182–2197, Nov. 2016.

A. K. O. Bachem and M. Lucic, “Practical coreset constructions for machine learning,” M.S. thesis, Dept. Comput. Sci., ETH Zurich, Zürich, Switzerland, 2017.

L. Bo, X. Ren, and D. Fox, “Hierarchical matching pursuit for image classification: Architecture and fast algorithms,” in Proc. Adv. Neural Inf. Process. Syst., Granada, Spain, Dec. 2011, pp. 2115–2123.

Y. C. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” IEEE Trans. Inf. Theory, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, “Introduction to compressed sensing,” Preprint, vol. 93, no. 1, p. 2, 2011.
ASHKAN ESMAEILI received the B.S. degree in electrical and computer engineering from the Sharif University of Technology, Tehran, Iran, in 2014, the M.Sc. degree in electrical and computer engineering from Stanford University, Stanford, CA, USA, in 2016, and the Ph.D. degree in electrical and computer engineering from the Sharif University of Technology. He is currently a Postdoctoral Researcher at the University of Central Florida, Orlando, FL, USA. His current research interests include machine learning, column subset selection, and meta learning.

SIAVASH KHODADADEH received the B.S. degree in computer science from the Sharif University of Technology, Tehran, Iran, in 2013, the M.Sc. degree in computer science from the University of Tehran, Tehran, in 2016, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA, in 2021. His current research interests include meta learning and computer vision.

SHARAREH ZEHTABIAN received the B.S. degree in computer science from the Sharif University of Technology, Tehran, Iran, in 2012, the M.Sc. degree in computer science from the University of Tehran, Tehran, in 2016, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA in 2021. Her current research interests include machine learning and computer vision.

BILL LIN (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA. He is currently a Professor with the Department of Electrical and Computer Engineering, UCSD. He is also an Adjunct Professor with the Department of Computer Science and Engineering, UCSD. Prior to joining UCSD, he was the Head of the System Control and Communications Group, IMEC. He also works a variety of computer engineering problems. He has coauthored over 200 journals and conference publications, including two best paper awards, three best paper nominations, and two distinguished paper citations. He has served on panels and given invited presentations at several major conferences, and he has served on over 50 program committees. He also holds five awarded patents.