Improved Guarantees and a Multiple-descent Curve for Column Subset Selection and the Nyström Method (Extended Abstract)*

Michał Dereziński†, Rajiv Khanna†, Michael W. Mahoney*†

1Department of Statistics, University of California at Berkeley
2International Computer Science Institute, Berkeley
{mderezin, rajivak}@berkeley.edu, mmahoney@stat.berkeley.edu

Abstract

The Column Subset Selection Problem (CSSP) and the Nyström method are among the leading tools for constructing interpretable low-rank approximations of large datasets by selecting a small but representative set of features or instances. A fundamental question in this area is: what is the cost of this interpretability, i.e., how well can a data subset of size $k$ compete with the best rank $k$ approximation? We develop techniques which exploit spectral properties of the data matrix to obtain improved approximation guarantees which go beyond the standard worst-case analysis. Our approach leads to significantly better bounds for datasets with known rates of singular value decay, e.g., polynomial or exponential decay. Our analysis also reveals an intriguing phenomenon: the cost of interpretability as a function of $k$ may exhibit multiple peaks and valleys, which we call a multiple-descent curve. A lower bound we establish shows that this behavior is not an artifact of our analysis, but rather it is an inherent property of the CSSP and Nyström tasks. Finally, using the example of a radial basis function (RBF) kernel, we show that both our improved bounds and the multiple-descent curve can be observed on real datasets simply by varying the RBF parameter.

1 Introduction

We consider the task of selecting a small but representative sample of column vectors from a large matrix. Known as the Column Subset Selection Problem (CSSP), this is a well-studied combinatorial optimization task with many applications in machine learning. In a commonly studied variant of this task, we aim to minimize the squared error of projecting all columns of the matrix onto the subspace spanned by the chosen column subset.

*This is an abridged version invited to IJCAI 2021 of a longer paper with the same title that appeared in NeurIPS 2020 and received a Best Paper Award.
†Corresponding author.
where we use \([n]\) to denote the set \([1, \ldots, n]\). The relative probability of a subset being drawn is governed by a positive semi-definite (p.s.d.) matrix \(K \in \mathbb{R}^{n \times n}\), as stated in the definition below, where we use \(K_{S,S}\) to denote the \(|S| \times |S|\) submatrix of \(K\) with rows and columns indexed by \(S\).

**Definition 2.** For an \(n \times n\) p.s.d. matrix \(K\), define \(S \sim \text{DPP}(K)\) as a distribution over all subsets \(S \subseteq [n]\) so that

\[
\Pr(S) = \frac{\det(K_{S,S})}{\det(I + K)}.
\]

A restriction to subsets of size \(k\) is denoted as \(k\)-DPP(\(K\)).

DPPs can be used to introduce diversity in the selected set or to model the preference for selecting dissimilar items, where the similarity is stated by the kernel matrix \(K\). DPPs are commonly used in many machine learning applications where these properties are desired, e.g., recommender systems [Warlop et al., 2019], model interpretation [Kim et al., 2016], text and video summarization [Gong et al., 2014], and others [Kulesza and Taskar, 2012]. For a recent survey, see [Dereziński and Mahoney, 2021].

Given a p.s.d. matrix \(K \in \mathbb{R}^{n \times n}\) with eigenvalues \(\lambda_1, \ldots, \lambda_n\), the size of the set \(S \sim \text{DPP}(K)\) is distributed as a Poisson binomial random variable, namely, the number of successes in \(n\) Bernoulli random trials where the probability of success in the \(i\)th trial is given by \(\frac{\lambda_i}{\lambda_1 + 1}\). This leads to a simple expression for the expected subset size:

\[
\mathbb{E}[|S|] = \sum_i \frac{\lambda_i}{\lambda_1 + 1} = \text{tr}(K(I + K)^{-1}). \tag{2}
\]

Note that if \(S \sim \text{DPP}(1_a K)\), where \(a > 0\), then \(\Pr(S)\) is proportional to \(a^{-|S|} \det(K_{S,S})\), so rescaling the kernel by a scalar only affects the distribution of the subset sizes, giving us a way to set the expected size to a desired value (larger \(a\) means smaller expected size). Nevertheless, it is still often preferable to restrict the size of \(S\) to a fixed \(k\), obtaining a \(k\)-DPP(\(K\)) [Kulesza and Taskar, 2011].

Both DPPs and \(k\)-DPPs can be sampled efficiently, with some of the first algorithms provided by [Hough et al., 2006], [Deshpande and Rademacher, 2010], [Kulesza and Taskar, 2011] and others. These approaches rely on an eigendecomposition of the kernel \(K\), at the cost of \(O(n^3)\). When \(K = A^T A\), as in the CSSP, and the dimensions satisfy \(m < n\), then this can be improved to \(O(nm^2)\). More recently, algorithms that avoid computing the eigendecomposition have been proposed [Dereziński, 2019; Dereziński et al., 2019; Calandriello et al., 2020; Anari et al., 2016], resulting in running times of \(\tilde{O}(n)\) when given matrix \(K\) and \(\tilde{O}(nm)\) for matrix \(A\), assuming small desired subset size. See [Gautier et al., 2019] for an efficient Python implementation of DPP sampling.

The key property of DPPs that enables our analysis is a formula for the expected value of the random matrix that is the orthogonal projection onto the subspace spanned by vectors selected by DPP(\(A^T A\)). In the special case when \(A\) is a square full rank matrix, the following result can be derived as a corollary of Theorem 1 by [Mutny et al., 2020], and a variant for DPPs over continuous domains can be found as Lemma 8 of [Dereziński et al., 2020].
Lemma 1. For any $\mathbf{A}$ and $S \subseteq [n]$, let $\mathbf{P}_S$ be the projection onto the span$\{\mathbf{a}_i : i \in S\}$. If $\mathbf{S} \sim \text{DPP}(\mathbf{A}^\top \mathbf{A})$, then

$$\mathbb{E}[\mathbf{P}_S] = \mathbf{A}(\mathbf{I} + \mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top.$$ 

Lemma 1 implies a simple closed form expression for the expected error in the CSSP presented next. Here, we use a rescaling parameter $\alpha > 0$ for controlling the distribution of the subset sizes. Note that it is crucial that we are using a DPP with random subset size, because the corresponding expression for the expected error of the fixed size $k$-DPP is combinatorial, and therefore much harder to work with.

Lemma 2. For any $\alpha > 0$, if $\mathbf{S} \sim \text{DPP}(\frac{1}{\alpha} \mathbf{A}^\top \mathbf{A})$, then

$$\mathbb{E}[\mathbf{E}_\alpha(S)] = \text{tr}(\mathbf{A}\mathbf{A}^\top (\mathbf{I} + \frac{1}{\alpha} \mathbf{A}\mathbf{A}^\top)^{-1}) = \mathbb{E}[||S||] \cdot \alpha.$$ 

3 Main Results

Our upper bounds rely on the notion of effective dimensionality called stable rank [Alaoui and Mahoney, 2015]. Here, we use an extended version of this concept, as defined by [Bartlett et al., 2019].

Definition 3 (Stable rank). Let $\lambda_1 \geq \lambda_2 \geq \ldots$ denote the eigenvalues of the matrix $\mathbf{A}^\top \mathbf{A}$. For $0 \leq s < \text{rank}(\mathbf{A})$, we define the stable rank of order $s$ as $\text{sr}_s(\mathbf{A}) = \lambda_s^{-1} \sum_{i > s} \lambda_i$.

In the following result, we define a family of functions $\Phi_s(k)$ which bound the approximation factor $\mathbb{E}[\mathbf{E}_\alpha(S)]/\text{OPT}_k$ in the range of $k$ between $s$ and $s + \text{sr}_s(\mathbf{A})$. We call this the Master Theorem because we use it to derive a number of more specific upper bounds.

Theorem 1 (Master Theorem). Given $0 \leq s < \text{rank}(\mathbf{A})$, let $t_s = s + \text{sr}_s(\mathbf{A})$, and suppose that $s + \frac{7}{16} \ln \frac{2}{k} \leq k \leq t_s - 1$, where $0 < \epsilon \leq \frac{1}{2}$. If $\mathbf{S} \sim k$-DPP$(\mathbf{A}^\top \mathbf{A})$, then

$$\frac{\mathbb{E}[\mathbf{E}_\alpha(S)]}{\text{OPT}_k} \leq (1 + 2\epsilon)^2 \Phi_s(k),$$

where $\Phi_s(k) = (1 + \frac{s}{\sqrt{k}}) \sqrt{1 + \frac{2(k-s)}{t_s-k}}$.

Note that we separated out the dependence on $\epsilon$ from the function $\Phi_s(k)$, because the term $(1 + 2\epsilon)^2$ is an artifact of a concentration of measure analysis that is unlikely to be of practical significance. We conjecture that the dependence on $\epsilon$ can be eliminated from the statement entirely.

We next examine the consequences of the Master Theorem, starting with a sharp transition that occurs as $k$ approaches the stable rank of $\mathbf{A}$.

Remark 1 (Sharp transition). For any $k$ it is true that:

1. For all $\mathbf{A}$, if $k \leq \text{sr}_0(\mathbf{A}) - 1$, then there is a subset $S$ of size $k$ such that $\frac{\mathbb{E}[\mathbf{E}_\alpha(S)]}{\text{OPT}_k} = O(\sqrt{k})$.

2. There is $\mathbf{A}$ such that $\text{sr}_0(\mathbf{A}) - 1 < k < \text{sr}_0(\mathbf{A})$ and for every size $k$ subset $S$, $\frac{\mathbb{E}[\mathbf{E}_\alpha(S)]}{\text{OPT}_k} \geq 0.9k$.

Part 1 of Remark 1 follows from the Master Theorem by setting $s = 0$, whereas part 2 follows from the lower bound of [Guruswami and Sinop, 2012]. Observe how the worst-case approximation factor jumps from $O(\sqrt{k})$ to $\Omega(k)$, as $k$ approaches $\text{sr}_0(\mathbf{A})$. An example of this sharp transition is shown in Figure 1, where the stable rank of $\mathbf{A}$ is around 20.

While certain matrices directly exhibit the sharp transition from Remark 1, many do not. In particular, for matrices with a known rate of spectral decay, the Master Theorem can be used to provide improved guarantees on the CSSP approximation factor over all subset sizes.

To illustrate this, we give novel bounds for the two most commonly studied decay rates: polynomial and exponential.

Theorem 2 (Examples without sharp transition). Let $\lambda_1 \geq \lambda_2 \geq \ldots$ be the eigenvalues of $\mathbf{A}^\top \mathbf{A}$. There is an absolute constant $c$ such that for any $0 < c_1 \leq c_2$, with $\gamma = c_2/c_1$, if:

1. (polynomial spectral decay) $c_1 1^{-p} \leq \lambda_i \leq c_2 1^{-p}$ for all $i$, with $p > 1$, then $\mathbf{S} \sim k$-DPP$(\mathbf{A}^\top \mathbf{A})$ satisfies

$$\frac{\mathbb{E}[\mathbf{E}_\alpha(S)]}{\text{OPT}_k} \leq c\gamma^p.$$ 

2. (exponential spectral decay) $c_1 (1-\delta)^i \leq \lambda_i \leq c_2 (1-\delta)^i$ for all $i$, with $\delta \in (0, 1)$, then $\mathbf{S} \sim k$-DPP$(\mathbf{A}^\top \mathbf{A})$ satisfies

$$\frac{\mathbb{E}[\mathbf{E}_\alpha(S)]}{\text{OPT}_k} \leq c(1+\delta k).$$

Note that for polynomial decay, unlike in (1), the approximation factor is constant, i.e., it does not depend on $k$. For exponential decay, our bound provides an improvement over (1) when $\delta = o(1)$. To illustrate how these types of bounds can be obtained from the Master Theorem, consider the function $\Phi_s(k)$ for some $s > 0$. The first term in the function, $(1 + \frac{s}{\sqrt{k}})$, decreases with $k$, whereas the second term (the square root) increases, albeit at a slower rate. This creates a U-shaped curve which, if sufficiently wide, has a valley where the approximation factor can get arbitrarily close to 1. This will occur when $\text{sr}_s(\mathbf{A})$ is large, i.e., when the spectrum of $\mathbf{A}^\top \mathbf{A}$ has a relatively flat region after the $s$th eigenvalue (Figure 1 for $k$ between 20 and 40). Note that a peak value of some function $\Phi_s(k)$ may coincide with a valley of some $\Phi_{s+1}$, so only taking a minimum over all functions reveals the true approximation landscape predicted by the Master Theorem.

The peaks and valleys of the CSSP approximation factor suggested by Theorem 1 are in fact an inherent property of the problem, rather than an artifact of our analysis or the result of using a particular algorithm. We prove this by constructing a family of matrices $\mathbf{A}$ for which the best possible approximation factor is large, i.e., close to the worst-case upper bound of [Deshpande et al., 2006], not just for one size $k$, but for a sequence of increasing sizes.

Theorem 3 (Lower bound). For any $\delta \in (0, 1)$ and $0 = k_0 < k_1 < \ldots < k_t < n \leq m$, there is a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that for any subset $S$ of size $k_i$, where $i \in \{1, \ldots, t\}$,

$$\frac{\mathbb{E}[\mathbf{E}_\alpha(S)]}{\text{OPT}_k_i} \geq (1-\delta)(k_i - k_{i-1}).$$

Combining the Master Theorem with the lower bound of Theorem 3 we can easily provide an example matrix for which the optimal solution to the CSSP problem exhibits multiple peaks and valleys. We refer to this phenomenon as the multiple-descent curve.
**Corollary 1** (Multiple-descent curve). For \( t \in \mathbb{N} \) and \( \delta \in (0, 1) \), there is a sequence \( 0 < k_1 < k_2 < \cdots < k_t \), and \( A \in \mathbb{R}^{m \times n} \) such that for any \( i \in \{1, \ldots, t\} \):

\[
\min_{S:|S|=k_i} \frac{\text{Er}_A(S)}{\text{OPT}_k_i} \leq 1 + \delta, \quad \text{and} \quad \min_{S:|S|=k_i} \frac{\text{Er}_A(S)}{\text{OPT}_k_i} \geq (1 - \delta)(k_i^n + 1).
\]

**The Nyström method.** We briefly discuss how our results translate to guarantees for the Nyström method, a variant of the CSSP in the kernel setting which has gained considerable interest in the machine learning literature [Drineas and Mahoney, 2005; Gittens and Mahoney, 2016]. In this context, rather than being given the column vectors explicitly, we consider the \( n \times n \) matrix \( K \) whose entry \((i, j)\) is the dot product between the \( i \)-th and \( j \)-th vector in the kernel space, \((\mathbf{a}_i, \mathbf{a}_j)_K\). A Nyström approximation of \( K \) based on subset \( S \) is defined as \( \tilde{K}(S) = \mathbf{B}^\top \mathbf{C}^\top \), where \( \mathbf{B} \) is the \(|S| \times |S|\) submatrix of \( K \) indexed by \( S \), whereas \( \mathbf{C} \) is the \( n \times |S| \) submatrix with columns indexed by \( S \).

**Remark 2.** If \( K = A^\top A \) and \( \| \cdot \|_s \) is the trace norm, then \( \| K - \tilde{K}(S) \|_s = \text{Er}_A(S) \) for all \( S \subseteq \{1, \ldots, n\} \). Moreover, the trace norm error of the best rank \( k \) approximation of \( K \) is equal to the squared Frobenius norm error of the best rank \( k \) approximation of \( A \), i.e.,

\[
\min_{R: \text{rank}(R) = k} \| K - R \|_s = \text{OPT}_{k}.
\]

**4 Empirical Evaluation**

In this section, we provide an empirical evaluation designed to demonstrate how our improved guarantees for the CSSP and Nyström method, as well as the multiple-descent phenomenon, can be easily observed on real datasets. We use a standard experimental setup for data subset selection using the Nyström method [Gittens and Mahoney, 2016], where an \( n \times n \) kernel matrix \( K \) for a dataset of size \( n \) is defined so that the entry \((i, j)\) is computed using the Gaussian Radial Basis Function (RBF) kernel: \((\mathbf{a}_i, \mathbf{a}_j)_K = \exp(-\| \mathbf{a}_i - \mathbf{a}_j \|^2/\sigma^2)\), where \( \sigma \) is a free parameter. We are particularly interested in the effect of varying \( \sigma \). Nyström subset selection is performed using \( S \sim k\text{-DPP}(K) \) (Definition 2), and we plot the expected approximation factor \( \mathbb{E}[\| K - \tilde{K}(S) \|_s]/\text{OPT}_k \) (averaged over 1000 runs), where \( \tilde{K}(S) \) is the Nyström approximation of \( K \) based on the subset \( S \), \( \| \cdot \|_s \) is the trace norm, and \( \text{OPT}_k \) is the trace norm error of the best rank \( k \) approximation. This task is equivalent to the CSSP task defined on the matrix \( A \) such that \( K = A^\top A \).

The aim of our empirical evaluation is to verify the following two claims motivated by our theory (and to illustrate that doing so is as easy as varying the RBF parameter \( \sigma \)):

1. When the spectral decay is sufficiently slow/smooth, the approximation factor for CSSP/Nyström is much better than suggested by previous worst-case bounds.
2. A drop in spectrum around the \( k \)-th eigenvalue results in a peak in the approximation factor near subset size \( k \). Several drops result in the multiple-descent curve.

In Figure 2 (top), we plot the approximation factor against the subset size \( k \) (in the range of 1 to 40) for a benchmark regression dataset *eunite2001* from the Libsvm repository [Chang and Lin, 2011]. In Figure 2 (bottom), we also show the top 40 eigenvalues of the RBF kernel \( K \) in decreasing order, for three different values of the parameter \( \sigma \).

The dataset *eunite2001* (Figure 2) exhibits a full multiple-descent curve with up to three peaks for large values of \( \sigma \) (see top plot), and the peaks are once again aligned with the spectrum drops (see bottom plot). Decreasing \( \sigma \) gradually eliminates the peaks, resulting in a uniformly small approximation factor. Thus, both of our theoretical claims can easily be verified on this dataset simply by adjusting the RBF parameter.

While the right choice of the parameter \( \sigma \) ultimately depends on the downstream machine learning task, it has been observed that varying \( \sigma \) has a pronounced effect on the spectral properties of the kernel matrix, [Gittens and Mahoney, 2016]. The main takeaway from our results here is that, depending on the structure of the problem, we may end up in the regime where the Nyström approximation factor exhibits a multiple-descent curve (e.g., due to a hierarchical nature of the data) or in the regime where it is relatively flat.
References


