Improved guarantees and a multiple-descent curve for the Column Subset Selection Problem and the Nyström method

Anonymous Authors[1]

Abstract

The Column Subset Selection Problem (CSSP) and the Nyström method are among the leading tools for constructing small low-rank approximations of large datasets in machine learning and scientific computing. A fundamental question in this area is: how well can a data subset of size k compete with the best rank k approximation? To address this question, we develop improved approximation guarantees which go beyond the standard worst-case analysis by relying on the spectral properties of the data. 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 approximation factor as a function of k may exhibit multiple peaks and valleys, which we call a multiple-descent curve. Moreover, our new lower bounds show that this behavior is not an artifact of our analysis, but rather an inherent property of the CSSP and Nyström tasks. Finally, we verify our theoretical analysis on real datasets.

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 (e.g., feature selection, see Guyon & Elisseeff, 2003; Boutsidis et al., 2008), scientific computing (e.g., Chan & Hansen, 1992; Drineas et al., 2008a) and signal processing (e.g., Balzano et al., 2010). 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.

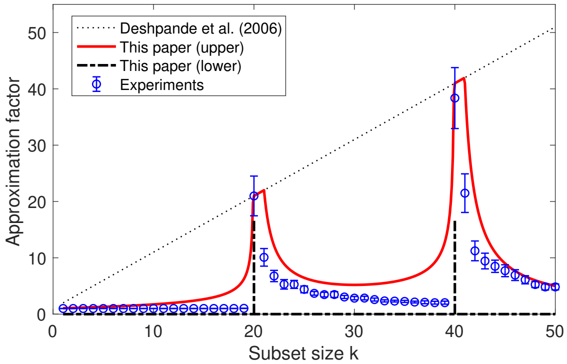

Figure 1. An empirical study showing the expected approximation factor $\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]/\mathrm{OPT}_k$ for $S \sim k\text{-DPP}(\mathbf{A}^\top\mathbf{A})$, with different subset sizes $|S| = k$, compared to our theory. We use a data matrix constructed as in Theorem 3 to demonstrate the multiple-descent phenomenon. As our upper bound, we plot the minimum over all functions $\Phi_s(k)$ from Theorem 1.

Definition 1 (CSSP). *Given an $m \times n$ matrix $\mathbf{A}$, pick a set $S \subseteq \{1, ..., n\}$ of k column indices, to minimize*

$$\mathrm{Er}_{\mathbf{A}}(S) := \|\mathbf{A} - \mathbf{P}_S\mathbf{A}\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{P}_S$ is the projection onto $\mathrm{span}\{\mathbf{a}_i : i \in S\}$ and $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$.

Another variant of the CSSP, of particular interest in machine learning, emerges in the kernel setting under the name *Nyström method* (Williams & Seeger, 2001; Drineas & Mahoney, 2005; Gittens & Mahoney, 2016). We discuss this variant in Section 1.2, showing how our analysis applies in this context. Both the CSSP and the Nyström method are ways of constructing accurate low-rank approximations by using submatrices of the target matrix. Therefore, it is natural to ask how close we can get to the best possible rank k approximation error:

$$\mathrm{OPT}_k := \min_{\mathbf{B}:\ \mathrm{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant \min_{S:\ |S|=k} \mathrm{Er}_{\mathbf{A}}(S).$$

Our goal is to find a subset S of size k for which the ratio between $\mathrm{Er}_{\mathbf{A}}(S)$ and OPT_k is small. Furthermore, a brute force search requires iterating over all $\binom{n}{k}$ subsets, which is

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

prohibitively expensive, so we would like to find our subset more efficiently.

In terms of worst-case analysis, Deshpande et al. (2006) gave a randomized method which returns a set S of size k such that:

$$\frac{\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leqslant k+1. \tag{1}$$

While the original algorithm was slow, efficient implementations have been provided since then (e.g., see Deshpande & Rademacher, 2010). The method belongs to the family of cardinality constrained determinantal point processes (see Definition 3), and will be denoted as $S \sim k\text{-DPP}(\mathbf{A}^\top\mathbf{A})$. The approximation factor $k+1$ is optimal in the worst-case, since for any $0<k<n\leqslant m$ and $0<\delta<1$, an $m\times n$ matrix $\mathbf{A}$ can be constructed for which $\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k} \geqslant (1-\delta)(k+1)$ for all subsets S of size k. Yet it is known that, in practice, CSSP algorithms perform better than worst-case, so the question we consider is: how can we go beyond the usual worst-case analysis to accurately reflect what is possible in the CSSP?

Contributions. We provide improved guarantees for the CSSP approximation factor, which go beyond the worst-case analysis and which lead to surprising conclusions.

1. New upper bounds: We develop a family of upper bounds on the CSSP approximation factor (Theorem 1), which we call the Master Theorem as they can be used to derive a number of new guarantees. In particular, we show that when the data matrix $\mathbf{A}$ exhibits a known spectral decay, then (1) can often be drastically improved (Theorem 2).

2. New lower bound: Even though the worst-case upper bound in (1) can often be loose, there are cases when it cannot be improved. We give a new lower bound construction (Theorem 3) showing that there are matrices $\mathbf{A}$ for which multiple different subset sizes exhibit worst-case behavior.

3. Multiple-descent curve: Our upper and lower bounds reveal that for some matrices the CSSP approximation factor can exhibit peaks and valleys as a function of the subset size k (see Figure 1). We show that this phenomenon is an inherent property of the CSSP (Corollary 1), which leads us to a new connection with the recently discovered double descent curve (Belkin et al., 2019a; Dereziński et al., 2019b).

1.1. Main results

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

Definition 2 (Stable rank)**.** *Let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$ denote the eigenvalues of the matrix $\mathbf{A}^\top\mathbf{A}$. For $0\leqslant s<\mathrm{rank}(\mathbf{A})$, we define the stable rank of order s as $\mathrm{sr}_s(\mathbf{A}) = \lambda_{s+1}^{-1}\sum_{i>s}\lambda_i$.*

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

Theorem 1 (Master Theorem)**.** *Given $0\leqslant s<\mathrm{rank}(\mathbf{A})$, let $t_s = s+\mathrm{sr}_s(\mathbf{A})$, and suppose that $s+\frac{7}{\epsilon^4}\ln^2\frac{1}{\epsilon}\leqslant k\leqslant t_s-1$, where $0<\epsilon\leqslant\frac{1}{2}$. If $S\sim k\text{-DPP}(\mathbf{A}^\top\mathbf{A})$, then*

$$\frac{\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leqslant (1+2\epsilon)^2\,\Phi_s(k),$$

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

Note that we separated out the dependence on ϵ 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. In fact, we believe that the dependence on ϵ can be eliminated from the statement entirely (see Conjecture 1).

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 $k\leqslant\mathrm{sr}_0(\mathbf{A})-1$, then there exists a subset S of size k such that $\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k} = O(\sqrt{k})$.*

2. *There is $\mathbf{A}$ such that $\mathrm{sr}_0(\mathbf{A})-1<k<\mathrm{sr}_0(\mathbf{A})$ and for every subset S of size k we have $\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k}\geqslant 0.9\,k$.*

Part 1 of the remark follows from the Master Theorem by setting $s=0$, whereas part 2 follows from the lower bound of Guruswami & Sinop (2012). Observe how the worst-case approximation factor jumps from $O(\sqrt{k})$ to $O(k)$, as k approaches $\mathrm{sr}_0(\mathbf{A})$. 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 $\lambda_1\geqslant\lambda_2\geqslant\ldots$ be the eigenvalues of $\mathbf{A}^\top\mathbf{A}$. There is an absolute constant c such that for any $0<c_1\leqslant c_2$, with $\gamma=c_2/c_1$, if:*

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

$$\frac{\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leqslant c\gamma p.$$

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

$$\frac{\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leqslant c\gamma(1+\delta k).$$

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, $1 + \frac{s}{k-s}$, 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 $\mathrm{sr}_s(\mathbf{A})$ 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 50). Note that a peak value of some function Φ_{s_1} may coincide with a valley of some Φ_{s_2}, so only taking a minimum over all functions reveals the true approximation landscape predicted by the Master Theorem. To prove Theorem 2, we show that the stable ranks $\mathrm{sr}_s(\mathbf{A})$ are sufficiently large so that any k lies in the valley of some function $\Phi_s(k)$ (see Section 4).

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* $0 = k_0 < k_1 < \ldots < k_t < n \leqslant m$, *there is a matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *such that for any subset* S *of size* k_i, *where* $i \in \{1,\ldots,t\}$,

$$\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_{k_i}} \geqslant (1-\delta)(k_i - k_{i-1}).$$

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* $0 < k_1^l < k_1^u < k_2^l < k_2^u < \ldots < k_t^l < k_t^u$ *and* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *such that for any* $i \in \{1,\ldots,t\}$*:*

$$\min_{S:|S|=k_i^l} \frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_{k_i^l}} \leqslant 1+\delta \quad \textit{and}$$

$$\min_{S:|S|=k_i^u} \frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_{k_i^u}} \geqslant (1-\delta)(k_i^u + 1).$$

The multiple-descent phenomenon that emerges from our analysis bears similarity to the double descent curve described by Belkin et al. (2019a). This curve illustrates the sharp transition between the generalization error of under- and over-parameterized machine learning models as we change the ratio between the number of parameters and the number of samples. We further discuss this connection in Section 2.

1.2. The Nyström method

We briefly discuss how our results translate to guarantees for the Nyström mehod, a variant of the CSSP in the kernel setting which has gained considerable interest in the machine learning literature (Drineas & Mahoney, 2005; Gittens & Mahoney, 2016). In this context, rather than being given the column vectors explicitly, we consider the $n \times n$ matrix $\mathbf{K}$ whose i,j-th entry is the dot product between the ith and jth vector in the kernel space, $\langle \mathbf{a}_i, \mathbf{a}_j\rangle_{\mathrm{K}}$. A Nyström approximation of $\mathbf{K}$ based on subset S is defined as $\widehat{\mathbf{K}}(S) = \mathbf{C}\mathbf{B}^\dagger\mathbf{C}^\top$, where $\mathbf{B}$ is the $|S| \times |S|$ submatrix of $\mathbf{K}$ indexed by S, whereas $\mathbf{C}$ is the $n \times |S|$ submatrix with columns indexed by S. The Nyström method has numerous applications in machine learning, including for kernel machines (Williams & Seeger, 2001), Gaussian Process regression (Burt et al., 2019) and Independent Component Analysis (Bach & Jordan, 2003).

Remark 2. *If* $\mathbf{K} = \mathbf{A}^\top\mathbf{A}$ *and* $\|\cdot\|_*$ *is the trace norm, then*

$$\big\|\mathbf{K} - \widehat{\mathbf{K}}(S)\big\|_* = \mathrm{Er}_{\mathbf{A}}(S) \quad \textit{for all} \quad S \subseteq \{1,\ldots,n\}.$$

Moreover, the trace norm error of the best rank k approximation of $\mathbf{K}$, *is equal to the squared Frobenius norm error of the best rank k approximation of* $\mathbf{A}$, *i.e.,*

$$\min_{\widehat{\mathbf{K}}:\,\mathrm{rank}(\widehat{\mathbf{K}})=k} \|\mathbf{K} - \widehat{\mathbf{K}}\|_* = \mathrm{OPT}_k.$$

This connection was used by Belabbas & Wolfe (2009) to adapt the $k+1$ approximation factor bound of Deshpande et al. (2006) to the Nyström method. Similarly, all of our results for the CSSP, including the multiple-descent curve that we have observed, can be translated into analogous statements for the trace norm approximation error in the Nyström method. Of particular interest are the improved bounds for kernel matrices with known eigenvalue decay rates. Such matrices arise naturally in machine learning when using standard kernel functions such as the squared exponential kernel and the Matérn kernel (Burt et al., 2019).

Squared exponential kernel: If $\mathbf{K}$ is the SE kernel with length-scale ℓ and the data is drawn from $\mathcal{N}(0, \sigma^2)$, then for large enough n (Santa et al., 1997), $\lambda_i \asymp \lambda_1(\frac{b}{a+b+c})^i$, where $a = 1/(4\sigma^2)$, $b = 1/(2\ell^2)$ and $c = \sqrt{a^2 + 2ab}$, so Theorem 2 yields a Nyström approximation factor bound of $O(1 + \frac{a+c}{a+b+c}k)$, which is better than $k + 1$ when $\ell^2 \ll \sigma^2$.

Matérn kernel: If $\mathbf{K}$ is the Matérn kernel with parameters ν and ℓ and the data is distributed according to a uniform measure in one dimension, then $\lambda_i \asymp \lambda_1 i^{-2\nu-1}$ (Rasmussen & Williams, 2006), so Theorem 2 yields a Nyström approximation factor of $O(1 + \nu)$ for any subset size k.

In Section 6, we also empirically demonstrate our improved guarantees and the multiple-descent curve for the Nyström method with the Radial Basis Function (RBF) kernel.

2. Related work

The Column Subset Selection Problem is one of the most classical tasks in matrix approximation (Boutsidis et al., 2008). The original version of the problem compares the projection error of a subset of size k to the best rank k approximation error. The techniques used for finding good subsets have included many randomized methods (Deshpande et al., 2006; Boutsidis et al., 2008; Belhadji et al., 2018), as well as deterministic methods (Gu & Eisenstat, 1996). Later on, most works have relaxed the problem formulation by allowing the number of selected columns $|S|$ to exceed the rank k. These approaches include deterministic sparsification based algorithms (Boutsidis et al., 2011), greedy selection (e.g., Altschuler et al., 2016) and randomized methods (e.g., Drineas et al., 2008b; Guruswami & Sinop, 2012; Paul et al., 2015). Note that we study the *original* version of the CSSP (i.e., without the relaxation), where the number of columns $|S|$ must be equal to the rank k.

The Nyström method has been given significant attention independently of the CSSP. The guarantees most comparable to our setting are due to Belabbas & Wolfe (2009), who show the approximation factor $k + 1$ for the trace norm error. Many recent works allow the subset size $|S|$ to exceed the target rank k, which enables the use of i.i.d. sampling techniques such as leverage scores (Gittens & Mahoney, 2016) and ridge leverage scores (Alaoui & Mahoney, 2015; Musco & Musco, 2017). In addition to the trace norm error, these works consider other types of guarantees, e.g., based on spectral and Frobenius norms, which are not as readily comparable to the CSSP error bounds.

The double descent curve was introduced by Belkin et al. (2019a) to explain the remarkable success of machine learning models which generalize well despite having more parameters than training data. This research has been primarily motivated by the success of deep neural networks, but double descent has also been observed in linear regression (Belkin et al., 2019b; Bartlett et al., 2019; Dereziński et al., 2019b) and other learning models. Double descent occurs when we plot the generalization error as a function of the number of parameters used in the learning model. The definition of generalization error relies on assuming a probabilistic generative model of the data. Importantly, our setting is different in that it is a *deterministic* combinatorial optimization problem. In particular, Corollary 1 shows that our multiple-descent curve can occur as a purely deterministic property of the optimal CSSP solution.

Determinantal point processes have been shown to provide near-optimal guarantees not only for the CSSP but also other tasks in numerical linear algebra, such as least squares regression (e.g., Avron & Boutsidis, 2013; Dereziński & Warmuth, 2018; Dereziński et al., 2019). They are also used in recommender systems, stochastic optimization and other tasks in machine learning (for a review, see Kulesza & Taskar, 2012). Efficient algorithms for sampling from these distributions have been proposed both in the CSSP setting (i.e., given matrix $\mathbf{A}$; see, e.g., Deshpande & Rademacher, 2010; Dereziński, 2019) and in the Nyström setting (i.e., given kernel $\mathbf{K}$; see, e.g., Anari et al., 2016; Dereziński et al., 2019). The term "cardinality constrained DPP" (also known as a "k-DPP" or "volume sampling") was introduced by Kulesza & Taskar (2011) to differentiate from standard DPPs which have random cardinality. Our proofs rely in part on converting DPP bounds to k-DPP bounds via a refinement of the concentration of measure argument used by Dereziński et al. (2019a).

3. Determinantal point processes

A Determinantal Point Process (DPP, introduced by Macchi, 1975) is a probability distribution over subsets $S \subseteq [n]$, where we use $[n]$ to denote the set $\{1, ..., n\}$. The relative probability of a subset being drawn is governed by a positive semidefinite (p.s.d.) matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, as stated in the definition below, where we use $\mathbf{K}_{S,S}$ to denote the $|S| \times |S|$ submatrix of $\mathbf{K}$ with rows and columns indexed by S.

Definition 3. *For an $n \times n$ p.s.d. matrix $\mathbf{K}$, define $S \sim \mathrm{DPP}(\mathbf{K})$ as a distribution over all subsets $S \subseteq [n]$ so that*

$$\Pr(S) = \frac{\det(\mathbf{K}_{S,S})}{\det(\mathbf{I} + \mathbf{K})}.$$

A restriction to subsets of size k is denoted as k-$\mathrm{DPP}(\mathbf{K})$.

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 $\mathbf{K}$. DPPs are commonly used in many machine learning applications where these properties are desired, e.g., recommender systems (Warlop et al., 2019), model interpretation (Been Kim & Koyejo, 2016), text and video summarization (Gong et al., 2014), and others (Kulesza & Taskar, 2012).

Given a p.s.d. matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, \dots \lambda_n$, the size of the set $S \sim \mathrm{DPP}(\mathbf{K})$ 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 ith trial is given by $\frac{\lambda_i}{\lambda_i+1}$. This leads to a simple expression for the expected subset size:

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

Note that if $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{K})$, where $\alpha > 0$, then $\Pr(S)$ is proportional to $\alpha^{-|S|}\det(\mathbf{K}_{S,S})$, 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 α means smaller expected size). Nevertheless, it is still often preferrable to restrict the size of S to a fixed k, obtaining a k-$\mathrm{DPP}(\mathbf{K})$ (Kulesza & Taskar, 2011).

Both DPPs and k-DPPs can be sampled efficiently, with some of the first algorithms provided by Hough et al. (2006), Deshpande & Rademacher (2010), Kulesza & Taskar (2011) and others. These approaches rely on an eigendecomposition of the kernel $\mathbf{K}$, at the cost of $O(n^3)$. When $\mathbf{K} = \mathbf{A}^\top\mathbf{A}$, as in the CSSP, and the dimensions satisfy $m \ll n$, then this can be improved to $O(nm^2)$. More recently, algorithms that avoid computing the eigendecomposition have been proposed (Anari et al., 2016; Dereziński et al., 2019; Dereziński, 2019), resulting in running times of $\widetilde{O}(n)$ when given matrix $\mathbf{K}$ and $\widetilde{O}(nm)$ for matrix $\mathbf{A}$, 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 $\mathrm{DPP}(\mathbf{A}^\top\mathbf{A})$. In the special case when $\mathbf{A}$ is a square full rank matrix, the following result can be derived as a corollary of Theorem 1 by Mutný et al. (2019), and a variant for DPPs over continuous domains can be found as Lemma 8 of Dereziński et al. (2019b). For completeness, we also provide a proof in Appendix A.

Lemma 1. *For any $\mathbf{A}$ and $S \subseteq [n]$, let $\mathbf{P}_S$ be the projection onto the* $\mathrm{span}\{\mathbf{a}_i : i \in S\}$. *If $S \sim \mathrm{DPP}(\mathbf{A}^\top\mathbf{A})$, 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. 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 $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{A}^\top\mathbf{A})$, then*

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

Proof. Using Lemma 1, the expected loss is given by:

$$\begin{aligned}\mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S)\big] &= \mathbb{E}\big[\|(\mathbf{I}-\mathbf{P}_S)\mathbf{A}\|_F^2\big] = \mathrm{tr}(\mathbf{A}\mathbf{A}^\top\mathbb{E}[\mathbf{I}-\mathbf{P}_S])\\ &= \mathrm{tr}\big(\mathbf{A}\mathbf{A}^\top(\mathbf{I}-\tfrac{1}{\alpha}\mathbf{A}(\mathbf{I}+\tfrac{1}{\alpha}\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top)\big)\\ &\overset{(*)}{=} \mathrm{tr}\big(\mathbf{A}\mathbf{A}^\top(\mathbf{I}+\tfrac{1}{\alpha}\mathbf{A}\mathbf{A}^\top)^{-1}\big),\end{aligned}$$

where $(*)$ follows from the matrix identity $(\mathbf{I}+\mathbf{A}\mathbf{A}^\top)^{-1} = \mathbf{I}-\mathbf{A}(\mathbf{I}+\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$. □

The challenge in using the above formula for the expected error is that it applies to a DPP with a randomized subset size rather than a fixed size k-DPP. Therefore, the random subset S with some positive probability has cardinality much greater than k. Our strategy in addressing this is to choose the rescaling parameter α so that the subset size is bounded by k with sufficiently high probability, via a concentration of measure argument, as discussed in the following section.

4. Upper bounds

In this section, we derive the upper bound given in Theorem 1 by using the expectation formula for the squared projection error of a DPP (Lemma 2). We then show how this result can be used to obtain improved guarantees for matrices with known eigenvalue decays, i.e., Theorem 2.

Recall that both the expected error formula and the expected subset size of $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{A}^\top\mathbf{A})$ depend on the rescaling parameter α, and our analysis relies on a careful selection of this parameter. To illustrate this, consider setting it to $\alpha = \mathrm{OPT}_k = \sum_{i=k+1}^n \lambda_i$, where λ_i are the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ in decreasing order. Now, (2) implies that:

$$\mathbb{E}[|S|] = \sum_{i=1}^n \frac{\lambda_i}{\alpha+\lambda_i} \leqslant \sum_{i=1}^k \frac{\lambda_i}{\alpha+\lambda_i} + 1 \leqslant k+1.$$

Together with Lemma 2, this recovers the upper bound of Deshpande et al. (2006) since $\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)] = \mathbb{E}[|S|]\cdot\alpha \leqslant (k+1)\cdot\mathrm{OPT}_k$, except that the subset size is randomized with expectation bounded by $k+1$, instead of a fixed subset size equal k. However, a more refined choice of the parameter α allows us to significantly improve on the above error bound in certain regimes, as shown below.

Lemma 3. *For any $\mathbf{A}$, $0 \leqslant \epsilon < 1$ and $s < k < t_s$, where $t_s = s + \mathrm{sr}_s(\mathbf{A})$, suppose that $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{A}^\top\mathbf{A})$ for $\alpha = \frac{\gamma_s(k)\mathrm{OPT}_k}{(1-\epsilon)(k-s)}$ and $\gamma_s(k) = \sqrt{1+\frac{2(k-s)}{t_s-k}}$. Then:*

$$\frac{\mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S)\big]}{\mathrm{OPT}_k} \leqslant \frac{\Phi_s(k)}{1-\epsilon} \quad \textit{and} \quad \mathbb{E}[|S|] \leqslant k - \epsilon\,\frac{k-s}{\gamma_s(k)},$$

where $\Phi_s(k) = \big(1+\frac{s}{k-s}\big)\gamma_s(k)$.

Note that, setting $\epsilon = 0$, the above lemma implies that we can achieve approximation factor $\Phi_s(k)$ with a DPP whose expected size is bounded by k. We introduce ϵ so that we can convert the bound from DPP to the fixed size k-DPP via a concentration argument. Intuitively, our strategy is to show that the randomized subset size of a DPP is sufficiently concentrated around its expectation that with high probability it will be bounded by k, and for this we need the expectation to be strictly below k. A careful application of the Chernoff bound for a Poisson binomial random variable yields the following concentration bound.

Lemma 4. *Let S be sampled as in Lemma 3 with $\epsilon \leqslant \frac{1}{2}$. If $s + \frac{7}{\epsilon^4}\ln^2\frac{1}{\epsilon} \leqslant k \leqslant t_s - 1$, then $\Pr(|S| > k) \leqslant \epsilon$.*

Finally, any expected bound for random size DPPs can be converted to an expected bound for a fixed size k-DPP via the following result.

Lemma 5. *For any $\mathbf{A} \in \mathbb{R}^{m\times n}$, $k \in [n]$ and $\alpha > 0$, if $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{A}^\top\mathbf{A})$ and $S' \sim k\text{-}\mathrm{DPP}(\mathbf{A}^\top\mathbf{A})$, then*

$$\mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S')\big] \leqslant \mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S) \mid |S| \leqslant k\big].$$

The above inequality may seem intuitively obvious since adding more columns to a set S to complete it to size k always reduces the error. However, a priori, it could happen that going from subsets of size $k-1$ to subsets of size k results in a redistribution of probabilities to the subsets with larger error. To show that this will not happen, our proof relies on classic but non-trivial combinatorial bounds called Newton's inequalities. Putting together Lemmas 3, 4 and 5, we obtain our Master Theorem.

Proof of Theorem 1 Let $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{A}^\top\mathbf{A})$ be sampled as in Lemma 3, and let $S' \sim k\text{-}\mathrm{DPP}(\mathbf{A}^\top\mathbf{A})$. We have:

$$\begin{aligned}\mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S')\big] &\overset{(a)}{\leqslant} \mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S) \mid |S| \leqslant k\big] \\ &\leqslant \frac{\mathbb{E}\big[\mathrm{Er}_{\mathbf{A}}(S)\big]}{\Pr(|S| \leqslant k)} \overset{(b)}{\leqslant} \frac{\Phi_s(k)}{(1-\epsilon)^2}\cdot \mathrm{OPT}_k,\end{aligned}$$

where (a) follows from Lemma 5 and (b) follows from Lemmas 3 and 4. Since $0 < \epsilon \leqslant \frac{1}{2}$, we have $\frac{1}{(1-\epsilon)^2} \leqslant (1+2\epsilon)^2$, which completes the proof. ■

We now demonstrate how Theorem 1 can be used as the Master Theorem to derive new bounds on the CSSP approximation factor under additional assumptions on the singular value decay of matrix $\mathbf{A}$. Rather than a single upper bound, Theorem 1 provides a family of upper bounds Φ_s, each with a range of applicable values k. Since each $\Phi_s(k)$ forms a U-shaped curve, its smallest point falls near the middle of that range. In Figure 2 we visualize these bounds as a sliding window that sweeps accross the axis representing possible subset sizes. The width of the window varies: when it starts at s then its width is the stable rank $\mathrm{sr}_s(\mathbf{A})$. The wider the window, the lower is the valley of the corresponding U-curve. Thus, when bounding the approximation factor for a given k, we should choose the widest window such that k falls near the bottom of its U-curve. Showing a guarantee that holds for all k requires lower-bounding the stable ranks $\mathrm{sr}_s(\mathbf{A})$ for each s. This is straightforward for both polynomial and exponential decay. Specifically, using the notation from Theorem 2, in Appendix C we prove that:

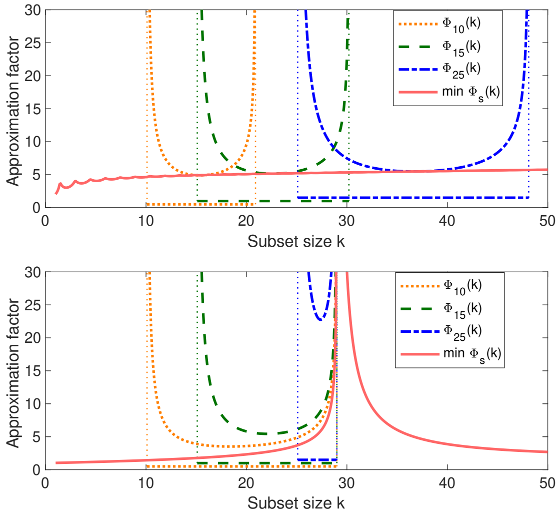

Figure 2. Illustration of the upper bound functions $\Phi_s(k)$ for different values of s, with a 200×200 matrix $\mathbf{A}$ such that the ith eigenvalue of $\mathbf{A}^\top\mathbf{A}$ is set to: (top) $1/i$; (bottom) 1 for $i < 30$ and 0.01 for $i \geqslant 30$. For each function, we marked the window of applicable k's with a horizontal line. For polynomial spectral decay (top), the stable rank $\mathrm{sr}_s(\mathbf{A})$ (i.e., the width of the window starting at s) increases, while for the sharp spectrum drop (bottom) the stable rank shrinks as the window approaches the drop, causing a peak in the upper bound.

$$\mathrm{sr}_s(\mathbf{A}) = \begin{cases}\Omega(s/p), & \text{for polynomial rate } \lambda_i \asymp 1/i^p,\\ \Omega(1/\delta), & \text{for exponential rate } \lambda_i \asymp (1-\delta)^i.\end{cases}$$

As an example, Figure 2 (top) shows that the stable rank $\mathrm{sr}_s(\mathbf{A})$, i.e., the width of the window starting at s, grows linearly with s for eigenvalues decaying polynomially with $p = 1$. As a result, the bottom of each U-shaped curve remains at roughly the same level, making the CSSP approximation factor independent of k, as in Theorem 2. In contrast, Figure 2 (bottom) provides the same plot for a different matrix $\mathbf{A}$ with a sharp drop in the spectrum. The U-shaped curves cannot slide smoothely accross that drop because of the shrinking stable ranks, which results in a peak similar to the ones observed in Figure 1.

5. Lower bound

As discussed in the previous section, our upper bounds for the CSSP approximation factor exhibit a peak (a high point, with the bound decreasing on either side) around a subset size k when there is a sharp drop in the spectrum of $\mathbf{A}$

around the kth singular value. It is natural to ask whether this peak is an artifact of our analysis, or a property of the k-DPP distribution, or whether even optimal CSSP subsets exhibit this phenomenon. In this section, we extend a lower bound construction of Deshpande et al. (2006) and use it to show that for certain matrices the approximation factor of the optimal CSSP subset, i.e., $\min_{|S|=k}\mathrm{Er}_{\mathbf{A}}(S)/\mathrm{OPT}_k$, can exhibit not just one but any number of peaks as a function of k, showing that the multiple-descent curve from Figure 1 describes an inherent phenomenon in the CSSP.

The lower bound construction of Deshpande et al. (2006) relies on arranging the column vectors of a $(k+1)\times(k+1)$ matrix $\mathbf{A}$ into a centered symmetric k-dimensional simplex. This way, the $k+1$ columns are spanning a k dimensional subspace which contains the k leading singular vectors of $\mathbf{A}$. They then proceed to shift the columns slightly in the direction orthogonal to that subspace so that the $(k+1)$st singular value of $\mathbf{A}$ becomes non-zero. This results in an instance of the CSSP with a sharp drop in the spectrum. Due to the symmetry in this construction, all subsets of size k have an identical squared projection error. It is easy to show that this error satisfies $\mathrm{Er}_{\mathbf{A}}(S)\geqslant(1-\delta)(k+1)\mathrm{OPT}_k$, where δ is a parameter which depends on the condition number of matrix $\mathbf{A}$ and it can be driven arbitrarily close to 0. Another variant of this construction was also provided by Guruswami & Sinop (2012). The key limitation of both of these constructions is that they only provide a lower bound for a single subset size k in a given matrix, whereas our goal is to show that the CSSP can exhibit the multiple-descent curve, which requires lower bounds for multiple different values of k holding with respect to the same matrix $\mathbf{A}$.

Our strategy for constructing the lower bound matrix is to concatenate together multiple sets of columns, each of which represents a simplex spanning some subspace of $\mathbb{R}^m$. The key challenge that we face in this approach is that, unlike in the construction of Deshpande et al. (2006), different subsets of the same size will have different projection errors. Nevertheless, we are able to lower bound these errors.

Lemma 6. *Fix $\delta\in(0,1)$ and consider unit vectors $\mathbf{a}_{i,j}\in\mathbb{R}^m$ in general position, where $i\in[t]$, $j\in[l_i]$, such that $\sum_j\mathbf{a}_{i,j}=0$ for each i, and for any i,j,i',j', if $i\neq i'$ then $\mathbf{a}_{i,j}$ is orthogonal to $\mathbf{a}_{i',j'}$. Also, let unit vectors $\{\mathbf{v}_i\}_{i\in[t]}$ be orthogonal to each other and to all $\mathbf{a}_{i,j}$. There are positive scalars α_i,β_i for $i\in[t]$ such that matrix $\mathbf{A}$ with columns $\alpha_i\mathbf{a}_{i,j}+\beta_i\mathbf{v}_i$ over all i and j satisfies:*

$$\min_{|S|=k_i}\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_{k_i}}\geqslant(1-\delta)l_i,\quad \textit{for } k_i=l_1+...+l_i-1.$$

Proof of Theorem 3 We let $l_1=k_1+1$ and then for $i>1$ we set $l_i=k_i-k_{i-1}$. We then construct the vectors $\mathbf{a}_{i,j}$ that satisfy Lemma 6 by letting each set $\{\mathbf{a}_{i,j}\}_j$ be the corners of a centered (l_i-1)-dimensional regular simplex. We ensure that each simplex is orthogonal to every other simplex by placing them in orthogonal subspaces. ■

We also use Lemma 6 in Appendix E to construct a matrix which exhibits the multiple-descent curve (Corollary 1).

6. 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 & Mahoney, 2016), where an $n\times n$ kernel matrix $\mathbf{K}$ for a dataset of size n is defined so that the i,j-th entry is computed using the Radial Basis Function (RBF) kernel: $\langle\mathbf{a}_i,\mathbf{a}_j\rangle_{\mathrm{K}}=\exp(-\|\mathbf{a}_i-\mathbf{a}_j\|^2/\sigma^2)$, where σ is a free parameter. We are particularly interested in the effect of varying σ. Nyström subset selection is performed using $S\sim k\text{-DPP}(\mathbf{K})$ (Definition 3), and we plot the expected approximation factor $\mathbb{E}[\|\mathbf{K}-\widehat{\mathbf{K}}(S)\|_*]/\mathrm{OPT}_k$ (averaged over 1000 runs), where $\widehat{\mathbf{K}}(S)$ is the Nyström approximation of $\mathbf{K}$ based on the subset S (see Section 1.2), $\|\cdot\|_*$ is the trace norm, and OPT_k is the trace norm error of the best rank k approximation. Additional experiments, using greedy selection instead of a k-DPP, are in Appendix F. As discussed in Section 1.2, this task is equivalent to the CSSP task defined on the matrix $\mathbf{A}$ such that $\mathbf{K}=\mathbf{A}^\top\mathbf{A}$.

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 σ):

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 kth eigenvalue results in a peak in the approximation factor near subset size k. Several drops result in the multiple-descent curve.

In Figure 3 (top), we plot the approximation factor against the subset size k (in the range of 1 to 40) for an artificial toy dataset and for two benchmark regression datasets from the Libsvm repository (*bodyfat* and *eunite2001*, see Chang & Lin, 2011). The toy dataset is constructed by scaling the eigenvalues of a random 50×50 Gaussian matrix so that the spectrum is flat with a single drop at the 21-st eigenvalue. For each dataset, in Figure 3 (bottom), we also show the top 40 eigenvalues of the kernel $\mathbf{K}$ in decreasing order. For the toy dataset, to maintain full control over the spectrum we use the linear kernel $\langle\mathbf{a}_i,\mathbf{a}_j\rangle_{\mathrm{K}}=\mathbf{a}_i^\top\mathbf{a}_j$, and we show results for three different values of the condition number κ of kernel $\mathbf{K}$. For the benchmark datasets, we show results on the RBF kernel with three different values of the parameter σ.

Examining the toy dataset (Figure 3, left), it is apparent that a larger drop in spectrum leads to a sharper peak in

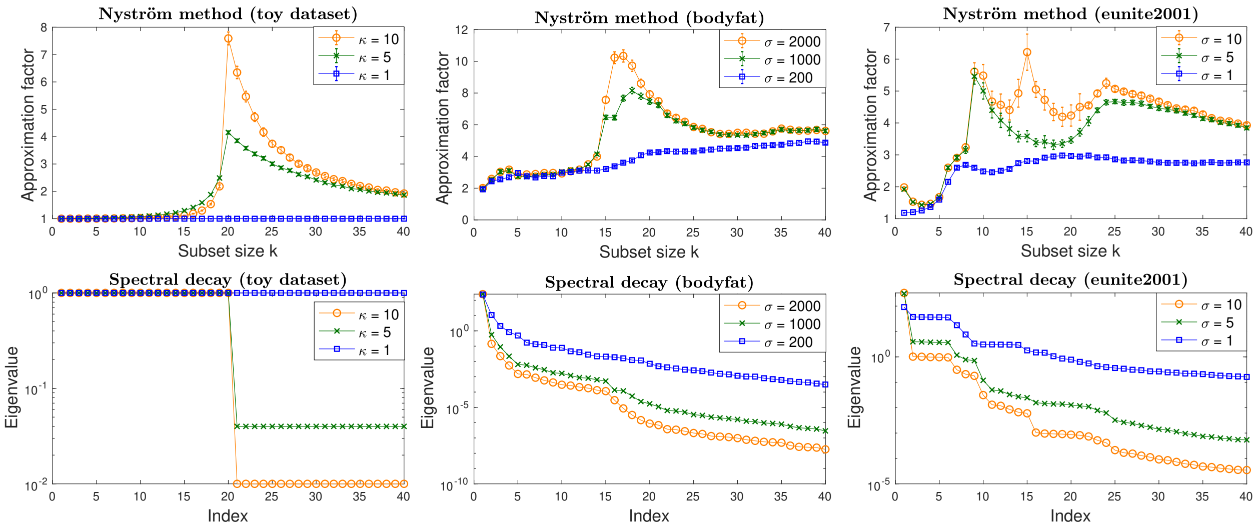

Figure 3. Top three plots show the Nyström approximation factor $\mathbb{E}[\|\mathbf{K}-\widehat{\mathbf{K}}(S)\|_*]/\mathrm{OPT}_k$, where $S \sim k\text{-DPP}(\mathbf{K})$ (experiments using greedy selection instead of a k-DPP are in Appendix F), for a toy dataset (κ is the condition number) and two Libsvm datasets (σ is the RBF parameter). Error bars show three times the standard error of the mean over 1000 trials. Bottom three plots show the spectral decay for the top 40 eigenvalues of each kernel $\mathbf{K}$. Note that the peaks in the approximation factor align with the drops in the spectrum.

the approximation factor as a function of the subset size k, whereas a flat spectrum results in the approximation factor being close to 1. A similar trend is observed for dataset *bodyfat* (Figure 3, center), where large parameter σ results in a peak that is aligned with a spectrum drop, while decreasing σ makes the spectrum flatter and the factor closer to 1. Finally, dataset *eunite2001* (Figure 3, right) exhibits a full multiple-descent curve with up to three peaks for large values of σ, and the peaks are once again aligned with the spectrum drops. Decreasing σ 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 σ ultimately depends on the downstream machine learning task, it has been observed that varying σ has a pronounced effect on the spectral properties of the kernel matrix, (see, e.g., Gittens & Mahoney, 2016; Lawlor et al., 2016; Wang et al., 2019). 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.

7. Conclusions and open problems

We derived new guarantees for the Column Subset Selection Problem (CSSP) and the Nyström method, going beyond worst-case analysis by exploiting the structural properties of a dataset, e.g., when the spectrum exhibits a known rate of decay. Our upper and lower bounds for the CSSP/Nyström approximation factor reveal an intruiguing phenomenon we call the multiple-descent curve: the approximation factor can exhibit a highly non-monotonic behavior as a function of k, with multiple peaks and valleys. These observations suggest a connection to the double descent curve exhibited by the generalization error of many machine learning models (see Section 2). This new connection is remarkable, since, unlike generalization error, the CSSP approximation factor is a deterministic objective in a combinatorial optimization problem without any underlying statistical model.

Our analysis technique relies on converting an error bound from random-size DPPs to fixed-size k-DPPs, which results in an additional constant factor of $(1+2\epsilon)^2$ in Theorem 1. We put forward a conjecture which would eliminate this factor from Theorem 1 and is of independent interest to the study of elementary symmetric polynomials, a classical topic in combinatorics (Hardy et al., 1952).

Conjecture 1. *The following function is convex with respect to $k \in [n]$ for any positive sequence $\lambda_1, ..., \lambda_n$:*

$$f(k) = (k+1)\frac{\sum_{S:|S|=k+1}\prod_{i\in S}\lambda_i}{\sum_{S:|S|=k}\prod_{i\in S}\lambda_i}.$$

Deshpande et al. (2006) showed that if $S \sim k\text{-DPP}(\mathbf{A}^\top\mathbf{A})$ and λ_i are the eigenvalues of $\mathbf{A}^\top\mathbf{A}$, then $\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)] = f(k)$. If $f(k)$ is convex then Jensen's inequality implies:

$$\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)] \leqslant \mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S')] \quad \text{for } S' \sim \mathrm{DPP}(\tfrac{1}{\alpha_k}\mathbf{A}^\top\mathbf{A}),$$

where α_k is chosen so that $\mathbb{E}[|S'|] = k$. This would allow us to use the bound from Lemma 3 directly on a k-DPP without relying on the concentration argument of Lemma 4, thereby improving the bounds in Theorems 1 and 2.

References

Alaoui, A. E. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 775–783, Montreal, Canada, December 2015.

Altschuler, J., Bhaskara, A., Fu, G., Mirrokni, V., Rostamizadeh, A., and Zadimoghaddam, M. Greedy column subset selection: New bounds and distributed algorithms. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2539–2548, New York, New York, USA, 20–22 Jun 2016. PMLR.

Anari, N., Gharan, S. O., and Rezaei, A. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In Feldman, V., Rakhlin, A., and Shamir, O. (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 103–115, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Avron, H. and Boutsidis, C. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, March 2003. ISSN 1532-4435.

Balzano, L., Recht, B., and Nowak, R. High-dimensional matched subspace detection when data are missing. In *2010 IEEE International Symposium on Information Theory*, pp. 1638–1642. IEEE, 2010.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. Technical Report Preprint: arXiv:1906.11300, 2019.

Been Kim, R. K. and Koyejo, S. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, 2016.

Belabbas, M.-A. and Wolfe, P. J. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0810600105.

Belhadji, A., Bardenet, R., and Chainais, P. A determinantal point process for column subset selection. *arXiv e-prints*, art. arXiv:1812.09771, Dec 2018.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. USA*, 116: 15849–15854, 2019a.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b.

Boutsidis, C., Mahoney, M., and Drineas, P. An improved approximation algorithm for the column subset selection problem. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 12 2008. doi: 10.1145/1496770.1496875.

Boutsidis, C., Drineas, P., and Magdon-Ismail, M. Near optimal column-based matrix reconstruction. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pp. 305–314, Oct 2011. doi: 10.1109/FOCS.2011.21.

Burt, D., Rasmussen, C. E., and Van Der Wilk, M. Rates of convergence for sparse variational Gaussian process regression. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 862–871, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Chan, T. F. and Hansen, P. C. Some applications of the rank revealing qr factorization. *SIAM Journal on Scientific and Statistical Computing*, 13(3):727–741, 1992.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Chung, F. and Lu, L. *Complex Graphs and Networks (Cbms Regional Conference Series in Mathematics)*. American Mathematical Society, Boston, MA, USA, 2006. ISBN 0821836579.

Dereziński, M. Fast determinantal point processes via distortion-free intermediate sampling. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1029–1049, Phoenix, USA, 25–28 Jun 2019.

Dereziński, M. and Warmuth, M. K. Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39, 2018.

Dereziński, M., Calandriello, D., and Valko, M. Exact sampling of determinantal point processes with sublinear time preprocessing. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11542–11554. Curran Associates, Inc., 2019.

Dereziński, M., Clarkson, K. L., Mahoney, M. W., and Warmuth, M. K. Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1050–1069, Phoenix, USA, 25–28 Jun 2019.

Dereziński, M., Liang, F., and Mahoney, M. W. Bayesian experimental design using regularized determinantal point processes. *arXiv e-prints*, art. arXiv:1906.04133, Jun 2019a.

Dereziński, M., Liang, F., and Mahoney, M. W. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv e-prints*, art. arXiv:1912.04533, Dec 2019b.

Deshpande, A. and Rademacher, L. Efficient volume sampling for row/column subset selection. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 329–338, Las Vegas, USA, October 2010.

Deshpande, A., Rademacher, L., Vempala, S., and Wang, G. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1117–1126, Miami, FL, USA, January 2006.

Drineas, P. and Mahoney, M. W. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008a.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 30(2):844–881, September 2008b.

Gautier, G., Polito, G., Bardenet, R., and Valko, M. DPPy: DPP Sampling with Python. *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS), in press*, 2019. URL http://arxiv.org/abs/1809.07258. Code at http://github.com/guilgautier/DPPy/ Documentation at http://dppy.readthedocs.io/.

Gittens, A. and Mahoney, M. W. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17(1):3977–4041, January 2016. ISSN 1532-4435.

Gong, B., Chao, W.-L., Grauman, K., and Sha, F. Diverse sequential subset selection for supervised video summarization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2069–2077. Curran Associates, Inc., 2014.

Gu, M. and Eisenstat, S. C. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

Guruswami, V. and Sinop, A. K. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1207–1214, Kyoto, Japan, January 2012.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null): 1157–1182, March 2003. ISSN 1532-4435.

Hardy, G., Littlewood, J., and Pólya, G. *Inequalities*. Cambridge University Press, 2nd edition, 1952.

Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

Kulesza, A. and Taskar, B. k-DPPs: Fixed-Size Determinantal Point Processes. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1193–1200, Bellevue, WA, USA, June 2011.

Kulesza, A. and Taskar, B. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

Lawlor, D., Budavári, T., and Mahoney, M. W. Mapping the similarities of spectra: Global and locally-biased approaches to SDSS galaxy data. *Astrophysical Journal*, 833(1), 12 2016.

Macchi, O. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975. ISSN 00018678.

Musco, C. and Musco, C. Recursive sampling for the nystrom method. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3833–3845. Curran Associates, Inc., 2017.

Mutný, M., Dereziński, M., and Krause, A. Convergence analysis of the randomized newton method with determinantal sampling. *arXiv e-prints*, art. arXiv:1910.11561, Oct 2019.

Paul, S., Magdon-Ismail, M., and Drineas, P. Column selection via adaptive sampling. In *Proceedings of the*

28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, pp. 406–414, Cambridge, MA, USA, 2015. MIT Press.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Santa, H. Z., Zhu, H., Williams, C. K. I., Rohwer, R., and Morciniec, M. Gaussian regression and optimal finite dimensional linear models. In *Neural Networks and Machine Learning*, pp. 167–184. Springer-Verlag, 1997.

Wang, R., Li, Y., Mahoney, M. W., and Darve, E. Block basis factorization for scalable kernel evaluation. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1497–1526, 2019.

Warlop, R., Mary, J., and Gartrell, M. Tensorized determinantal point processes for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 1605–1615, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6201-6.

Williams, C. K. I. and Seeger, M. Using the Nyström method to speed up kernel machines. In Leen, T. K., Dietterich, T. G., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, pp. 682–688. MIT Press, 2001.