Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Sister Conferences Best Papers Track

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

Michał Dereziński[1†], **Rajiv Khanna**[1], **Michael W. Mahoney**[1,2]
[1]Department of Statistics, University of California at Berkeley
[2]International 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.

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 emerges in the kernel setting under the name *Nyström method* [Williams and Seeger, 2001; Drineas and Mahoney, 2005; Gittens and Mahoney, 2016]. We also discuss this variant, 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:

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

While the best possible rank k approximation has the lowest approximation error, the approximated matrix $\mathbf{B}$ is not input-sparse. As such, it is not *interpretable* since a practitioner is unable to attribute the quality of approximation to specific physical quantities represented by the columns of the matrix $\mathbf{A}$ [Mahoney and Drineas, 2009]. Such *interpretable* dimensionality reduction is desirable in many machine learning applications. Our goal is to find a subset S of size k for which the cost of interpretability is small, as measured by what we call the *approximation factor*: the ratio between $\mathrm{Er}_{\mathbf{A}}(S)$ and OPT_k. Furthermore, a brute force search requires iterating over all $\binom{n}{k}$ subsets, which is 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:

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

While the original algorithm was slow, efficient implementations have been provided since then [Deshpande and Rademacher, 2010; Dereziński, 2019]. The method belongs to the family of cardinality constrained Determinantal Point Processes (DPPs), and will be denoted as $S \sim k\text{-DPP}(\mathbf{A}^\top\mathbf{A})$; for an overview of DPPs, see Section 2 and [Dereziński and Mahoney, 2021]. The approximation factor

*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.

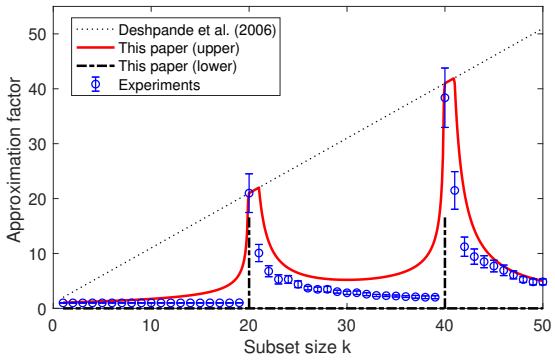

Figure 1: Empirical study of the expected approximation factor $\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]/\mathrm{OPT}_k$ for a k-DPP with different subset sizes $|S| = k$, compared to our theory. We use a data matrix $\mathbf{A}$ whose spectrum exhibits two sharp drops, demonstrating multiple-descent. The lower bounds are based on Theorem 3, whereas, as our upper bound, we plot the minimum over all $\Phi_s(k)$ from Theorem 1. Note that multiple-descent vanishes under smooth spectral decay, resulting in improved guarantees (see Theorem 2).

$k+1$ is optimal in the worst-case, since for any $0 < k < n \leq m$ and $0 < \delta < 1$, an $m \times n$ matrix $\mathbf{A}$ can be constructed for which $\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k} \geq (1-\delta)(k+1)$ 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).

2 Determinantal Point Processes

Since our main results rely on randomized subset selection via determinantal point processes (DPPs), we provide a brief overview of the relevant aspects of this class of distributions. First introduced by [Macchi, 1975], a determinantal point process 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 2. *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 [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 $\mathbf{K} \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, ... \lambda_n$, 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 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 $\mathbf{K}$, 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 [Dereziński, 2019; Dereziński *et al.*, 2019; Calandriello *et al.*, 2020; Anari *et al.*, 2016], 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 [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* $\mathrm{span}\{\mathbf{a}_i : i \in S\}$*. 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 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* $S \sim \mathrm{DPP}(\frac{1}{\alpha}\mathbf{A}^\top\mathbf{A})$*, then*

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

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

$$\frac{\mathbb{E}[\mathrm{Er}_{\mathbf{A}}(S)]}{\mathrm{OPT}_k} \leq (1+2\epsilon)^2\,\Phi_s(k),$$
$$\text{where } \Phi_s(k) = \big(1 + \tfrac{s}{k-s}\big)\sqrt{1 + \tfrac{2(k-s)}{t_s - k}}\,.$$

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. We conjecture that the dependence on ϵ 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* $k \leq \mathrm{sr}_0(\mathbf{A}) - 1$*, then there is 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 size* k *subset* S*,* $\frac{\mathrm{Er}_{\mathbf{A}}(S)}{\mathrm{OPT}_k} \geq 0.9\,k$*.*

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 $\mathrm{sr}_0(\mathbf{A})$. 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* $\lambda_1 \geq \lambda_2 \geq \ldots$ *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)*** $c_1 i^{-p} \leq \lambda_i \leq c_2 i^{-p}$ *for all* i*, with* $p > 1$*, then* $S \sim k\text{-}\mathrm{DPP}(\mathbf{A}^\top\mathbf{A})$ *satisfies*

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

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* $S \sim k\text{-}\mathrm{DPP}(\mathbf{A}^\top\mathbf{A})$ *satisfies*

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

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 sth eigenvalue (Figure 1 for k between 20 and 40). 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.

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* $i \in \{1,\ldots,t\}$*,*

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

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

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

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

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 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 $\mathbf{K}$ whose entry (i,j) is the dot product between the ith and jth vector in the kernel space, $\langle \mathbf{a}_i, \mathbf{a}_j \rangle_{\mathbf{K}}$. 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.

Remark 2. *If $\mathbf{K} = \mathbf{A}^{\top}\mathbf{A}$ and $\|\cdot\|_*$ is the trace norm, then $\left\|\mathbf{K} - \widehat{\mathbf{K}}(S)\right\|_* = \mathrm{Er}_{\mathbf{A}}(S)$ for all $S \subseteq \{1,...,n\}$. 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.,*

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

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 $\mathbf{K}$ for a dataset of size n is defined so that the entry (i,j) is computed using the Gaussian Radial Basis Function (RBF) kernel: $\langle \mathbf{a}_i, \mathbf{a}_j \rangle_{\mathbf{K}} = \exp(-\|\mathbf{a}_i - \mathbf{a}_j\|^2/\sigma^2)$, 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 2), 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, $\|\cdot\|_*$ is the trace norm, and 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 $\mathbf{A}$ 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.

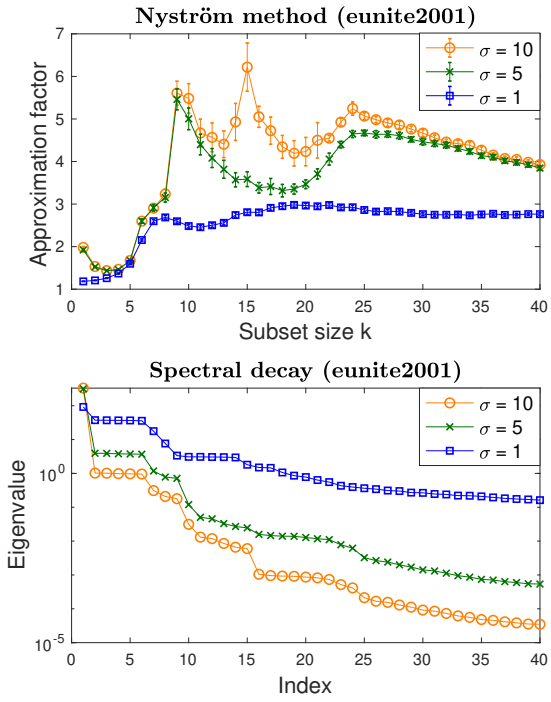

Figure 2: Top plot shows the Nyström approximation factor $\mathbb{E}[\|\mathbf{K} - \widehat{\mathbf{K}}(S)\|_*]/\mathrm{OPT}_k$, where $S \sim k\text{-DPP}(\mathbf{K})$ for the eunite2001 Libsvm dataset (σ is the RBF parameter). Error bars show three times the standard error of the mean over 1000 trials. Bottom plot shows the spectral decay for the top 40 eigenvalues of the kernel $\mathbf{K}$. Note that the peaks in the approximation factor align with the drops in the spectrum.

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 $\mathbf{K}$ in decreasing order, for three different values of the parameter σ.

The dataset *eunite2001* (Figure 2) exhibits a full multiple-descent curve with up to three peaks for large values of σ (see top plot), and the peaks are once again aligned with the spectrum drops (see bottom plot). 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, [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

[Alaoui and Mahoney, 2015] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 775–783, 2015.

[Anari *et al.*, 2016] Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In *29th Annual Conference on Learning Theory*, volume 49, pages 103–115, 2016.

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

[Calandriello *et al.*, 2020] Daniele Calandriello, Michał Dereziński, and Michal Valko. Sampling from a k-dpp without looking at all items. In *Advances in Neural Information Processing Systems*, volume 33, pages 6889–6899, 2020.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Dereziński and Mahoney, 2021] Michał Dereziński and Michael W Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1):34–45, 2021.

[Dereziński *et al.*, 2019] Michał Dereziński, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems*, pages 11542–11554. 2019.

[Dereziński *et al.*, 2020] Michał Dereziński, Feynman Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. In *Advances in Neural Information Processing Systems*, volume 33, pages 5152–5164, 2020.

[Dereziński, 2019] Michał Dereziński. Fast determinantal point processes via distortion-free intermediate sampling. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1029–1049, 2019.

[Deshpande and Rademacher, 2010] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 329–338, 2010.

[Deshpande *et al.*, 2006] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proc. 17th Annual Symposium on Discrete Algorithm*, pages 1117–1126, 2006.

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

[Gautier *et al.*, 2019] Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. DPPy: DPP sampling with python. *Journal of Machine Learning Research*, 20(180):1–7, 2019.

[Gittens and Mahoney, 2016] Alex Gittens and Michael W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17(1):3977–4041, January 2016.

[Gong *et al.*, 2014] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems 27*, pages 2069–2077. 2014.

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

[Hough *et al.*, 2006] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, Bálint Virág, et al. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

[Kim *et al.*, 2016] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, 2016.

[Kulesza and Taskar, 2011] Alex Kulesza and Ben Taskar. k-DPPs: Fixed-Size Determinantal Point Processes. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1193–1200, June 2011.

[Kulesza and Taskar, 2012] Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

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

[Mahoney and Drineas, 2009] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[Mutny *et al.*, 2020] Mojmir Mutny, Michał Dereziński, and Andreas Krause. Convergence analysis of block coordinate algorithms with determinantal sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 3110–3120, 2020.

[Warlop *et al.*, 2019] Romain Warlop, Jérémie Mary, and Mike Gartrell. Tensorized determinantal point processes for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 1605–1615, 2019.

[Williams and Seeger, 2001] Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.