View Reviews

Paper ID
3394

Paper Title
Improved guarantees and a multiple-descent curve for the Column Subset Selection Problem and the Nystrom method

Reviewer #2

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
This paper studies a column selection problem for positive definite matrices. Let K be an input matrix. The goal is to select a subset of columns so that the residual error is minimized when we project columns of K onto the subset. The two metrics used are the ratio between the residual error produced by a column selection algorithm (for a subset of size k) and the residual error produced from optimal rank-k approximation.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
The result offers a deeper theoretical understanding of the column selection problem.

3. Please provide an overall evaluation for this submission.
Very good paper, I would like to see it accepted.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
I APOLOGIZE FOR A LATE UPDATE. I REALIZE I MIS-CHOSE THE OVERALL EVALUATION (BORDERLINE ->VERY GOOD).

The paper's central contribution is the design of a randomized algorithm whose performance is characterized by the stable-rank introduced in Bartlett. Roughly speaking, the stable rank $sr_s(A)$ (s is a parameter) measures the mass of the tail compared against the eigenvalue $lambda_{s+1}$. When the rank is high, the eigenvalues of K decay smoothly.

The authors use a technique called determinantal point processes. The probability that a subset S is selected is proportional to the volume of the parallelogram specified by $K_{S,S}$ (i.e., its determinant). Their analysis relates the expected error of the algorithm with the size of the output. Since the size of the output can also be characterized by the eigenvalues of K and a hyper-parameter $alpha$, the structure of the eigenvalues (expressed by the stable rank) can be used to analyze the error of the proposed algorithm.

Their upper bound suggests a multiple-descent phenomenon and it is confirmed by experimental data in Fig 1. The authors also develop a new lower bound result showing the existence of multiple spikes on the lower bound curve (as k changes). Therefore, the phenomenon is not an artifact of their upper bound analysis.

Overall, the paper tackles an interesting problem and provides a new residual error bound that can be significantly better than existing results that rely on the worst-case analysis. The analysis of the algorithm is sound and interesting.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.

Motivation. I thought a primary motivation of the problem is to reduce the running time, compared to using SVD. But it appears that their proposed algorithm could be slow for three reasons: (i) sample through DPP: The original implementation requires running the SVD solver. While they acknowledge that some recent methods could be faster, their algorithm's performance remains unclear (no source code was provided); (ii) the value of alpha: Alpha is defined through opt_k, which seems to require solving SVD; (iii) finding a subset of size exactly k. It appears that they use a standard rejection sampling method to find a subset of size exactly k. Even with carefully tuned concentration, aren't a total number of $sqrt{k}$ trials needed before a size k subset is returned?

Metrics. The spikes for the approximation ratio (over the residual errors) seem like an artifact from the metrics, especially the sharp transition result they present. If we were to use explained variation (r2) as the metrics, probably the spikes will go away or be milder. Therefore, I doubt whether this approximation ratio is a good metric. Are there any concrete examples in which the ratio directly controls the quality of the output of an ML algorithm (e.g., a 10-approximation algorithm has a prediction error that is 5 times higher than a 2-approximation algorithm)?.

Matern kernel: The authors claim that the eigenvalues of a Matern kernel decay in a polynomial manner (around line 72). Most decay rate analyses I know use Chebyshev polynomial/Fourier transform to give the upper bound of the decay. I also think it is rare to have kernels with polynomial decay (e.g., when the kernels are sufficiently smooth, the decay is always super polynomial). I would like the authors to be specific about the reference (a textbook) Rasmussen & Williams. I was not able to find the decay rate mentioned in the paper.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have closely read papers on this topic, and written papers in the broad area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #3

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.

This paper studies the Column Subset Selection Problem (CSSP) and the Nystrom method, which are major tools for constructing small low-rank approximations. The main contribution of this paper is a general theorem on the guarantee of approximation errors (Theorem 1). The new bounds are instance-dependent and be used to derive a number of improved guarantees for special families of matrices.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
The main contribution of this paper is a general theorem on the guarantee of approximation errors of CSSP and the Nystrom method. The new bounds are instance-dependent and can be used to derive a number of improved guarantees for special families of matrices. Tight lower bounds are also provided.

3. Please provide an overall evaluation for this submission.
Very good paper, I would like to see it accepted.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
The new error bounds are instance-dependent and can be used to derive a number of improved guarantees for special families of matrices. More interestingly, these
bounds exhibit a highly non-monotonic behavior of error as a function of k. New lower bound results show that, for certain matrices, the approximation error of the optimal CSSP subset can exhibit any number of peaks as a function of k, which indicates that the upper bound is not an artifact of the analysis, or a property of
the k-DPP distribution.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
This paper studies the Column Subset Selection Problem (CSSP) and the Nystrom method, which are major tools for constructing small low-rank approximations. These problems have been extensively studied and optimal worse-case bounds are known. However, in practice, CSSP algorithms often perform better than such worst-case bounds.

The main contribution of this paper is a general theorem on the guarantee of approximation errors (Theorem 1). The new bounds are instance-dependent and can be used to derive a number of improved guarantees for special families of matrices. More interestingly, these bounds exhibit a highly non-monotonic behavior of error as a function of k. Moreover, new lower bounds results show that, for certain matrices, the approximation error of the optimal CSSP subset can exhibit any number of peaks as a function of k, which indicates that the upper bound is not an artifact of the analysis, or a property of the k-DPP distribution.

I think the theoretical results are very interesting and strong, which fit phenomenons in practice much better than previous worst-case analysis and could be the starting point for other analogous results in related fields.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have closely read papers on this topic, and written papers in the broad area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have

encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #4

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
This paper explores the approximation quality of column subset selection methods, which approximate a matrix by projecting onto a small subset of its columns. These methods are very well studied and optimal algorithms are known, at least in terms of worst case approximation quality. This paper gives nice beyond worst case bounds and matching lower bounds and some interesting insights into when the error from these methods can be large compared to optimal low-rank approximation.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
The paper is well written and gives a nice exploration of CSS approximation factors. The master theorem, which applies specifically to column subset selection with k-DPP sampling (a well known technique) seems to provide a concise 'better than worst case' way of understanding CSS error for this method. Additionally, the lower bounds of Theorem 3 and Cor 1 are interesting, and extend our understanding of worst case CSS approximation factors beyond just focusing on a single rank k. Note that these lower bounds hold in general, and are not just for the k-DPP method.

At the highest qualitative level, the master theorem and the lower bounds suggest that the CSS approximation factor peaks when the rank k is near a steep drop in the matrix spectrum, which is an interesting (if not all that surprising) phenomena.

The master theorem is based on some nice ideas. Lemmas 1 and 2 give a really clean bound on the expected error of a DPP based CSS algorithm in terms of the matrix spectrum, but when the number of selected columns selected by the DPP is random rather than fixed to k. The authors then set the parameters of this DPP so that the number of sampled columns is <= k with high probability. Finally, they argue that running the k-DPP that the master theorem applies to can only give a better bound that this unconstrained DPP, which outputs < k columns with high probability. This last step is intuitively clear, but the proof is not so easy. The authors conjecture that a simpler and tighter proof is possible, by directly comparing and unconstrained DPP with expected subset size k to the k-DPP. This is an interesting conjecture. Overall, these proof techniques seem interesting and valuable, beyond just the master theorem which they are used to show.

The experimental results seem to confirm that the bounds given predict the behavior of CSS approximation, notably the fact that approximate error can increase when the matrix spectrum has steep drops.

3. Please provide an overall evaluation for this submission.
Below the acceptance threshold, I would rather not see it at the conference.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
I have one major reservation about the paper which drives my negative review despite liking the paper overall: I think the connection to double descent is questionable and obscures the message of the paper, making it unpublishable at ICML in its current form.

From what I can tell, the double descent observed here has nothing to do with the double descent phenomena recently popular in understanding generalization of overparameterized ML models. It is literally just another case of a function that has multiple peaks. One particular difference (although there are many) is that the double descent in generalization refers to approximation error actually increasing as the number of parameters increases to the interpolation threshold and then decreasing again. In CSS the error *monotonically decreases* as the rank of the approximation k increases. Only the *relative error* compared to the best rank k approximate may have double descent.

The authors state: "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."

I think this highlights the issue with the connection. Just because two different error functions in two totally different contexts both have multiple peaks doesn't mean they are connected. And it certainly isn't remarkable or surprising that such functions exist in general.

If this paper didn't stress this connection to double descent so much (or really mention it at all) I would be inclined to accept. But in the current state, would not support accepting the paper.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
Some smaller negatives/questions I had about the paper:

1. I would be interested to see more methods (like leverage score or adaptive sampling) compared to, to understand how specific the results are to the k-DPP method. It seems that Cor 1 suggests that the results, at least at a coarse level of predicting error peaks, should apply to general methods, but I wasn't completely sure and this didn't seem to be addressed in the experiments section.

2. The first regime of Remark 1 seems odd. Is it basically the regime where k is so small that no good approximation is given? Say e.g. that A just has s eigenvalues, all equal to 1. the sr_0(A) = s. Then for (1) to hold, we must have k < s -- i.e., k is too small to capture all large eigenvalues of the matrix. The case is even worse when A has s large eigenvalues, but they are not uniform. Then we have sr_0(A) ~< s and so again the bound only holds when k is too small to capture the large eigenvalues of A. Could use more of an intuitive explanation of what is going on here.

3. As compared to the upper bounds, the lower bounds are somewhat less exciting. The proof of Theorem 3 just follows from an existing lower bound for a single value of k. The hard instance for multiple values of k just 'stacks' orthogonal versions of this lower bound instance into one matrix. Nevertheless, these bounds are useful in

understanding the behavior of CSS approximation, even if the proofs are straightforward.

Some more specific comments/questions:

- In intro, after citing the Deshpande result I think would be worth citing the many results which give 1+epsilon bicriteria approximation. These are cited in Section 2, but them missing from intro was odd, since there has been so much focus on them. I think it should be justified a bit more why the focus is on the case of |S| = k. This seems like a somewhat arbitrary restriction. How much do the results depend on this?
- Line 107: 'we use an extended version of this concept' -- This is just a variant, not an extension right?
- The definition of stable rank is a bit odd -it is summing over small, not large eigenvalues. So if a matrix has k large eigenvalues and the remainder all essentially 0, the stable rank in no way corresponds to k, as other common notions of stable rank would. In fact in numerical analysis a notion of stable rank which is just the number of eigenvalues above s is often used, which is in some sense 'opposite' of the notion used here. This, I would suggest renaming this term maybe. It is more a notion of tail error of the singular values than a notion of rank.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have published one or more papers in the narrow area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted