View Reviews

Paper ID
8208

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

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper studies the important column subset selection problem (as well as extensions to the Nyström method), where the aim is to find a subset of k columns such that the projection cost of all points to the span of selected points is minimized. Classic randomized algorithms for the problem could only guarantee a worst-case approximation factor of O(k+1), and a lower bound construction was known where the approximation is \Omega(k). It is also known that the algorithms often perform much better than this worst case.

The paper thus inspects the behavior of upper and lower bounds parameterized by k in more detail by providing a kind of "beyond worst case analysis", that still considers the worst case but for each value of k (or regions thereof) individually rather than only for the worst case choice of k.

The main contributions are:
1. A sequence of upper bounds (for sampling the columns via a k-Determinantal Point Process (k-DPP)), depending on a stable rank parameter, that coincide with the Theta(k) bound at values of k where the spectrum has significant jumps. But the new bounds have significantly lower "valleys" in the regions between those jumps. E.g. for all values of k lower than the stable rank (of order 0) a simple consequence of their "master theorem" is a O(sqrt(k)) upper bound. They call this behavior with multiple peaks and valleys "multiple descent curve".

2. An application to matrices where the spectrum does't possess sudden jumps but a "regular" decay. For polynomial decay (of degree p) the bound becomes O(p). For an exponential decay, (1-delta)^i the bound becomes (1+delta k) which is better than the worst case whenever delta is o(1).

3. A worst case construction where the spectrum has a sequence of jumps, so that the lower bound becomes large at those jumps but the error remains upper bounded by a constant between those peaks. This is done by putting the known single jump construction into orthogonal subspaces so they have independent impact. This construction shows that the "multiple descent curve" structure of the upper and lower bounds is inherent to the problem, rather than an artifact of the k-DPP sampling method or of the analysis. (additionally the behavior is also observed for a greedy selection algorithm in the supplement)

4. It is explained how the above results extend to the Nyström approximation with trace norm error. This also yields further upper bounds in terms of defining parameters of the RBF, and Matern kernels.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS

community.
The paper makes significant progress in the study of the CSSP and Nyström approximation problems. Notably under the original constraint of subsets with at most k elements. (no oversampling allowed)

It provides a natural form of beyond worst case analysis with highly non-trivial upper and lower bounds dependent on the subset size.

It provides experimental assessments of the theoretical claims that even further support the intuition.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
no considerable weaknesses

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
all claims seem correct; the empirical methodology satisfies highest scientific standards clearly stating the research question, explaining how it is assessed and answering the questions by drawing correct conclusions from the experiment.

5. Clarity: Is the paper well written?
The paper is very well written. It is one of few papers that have the perfect balance between formal technical and intuitive explanations in the main writeup, and additional much more detailed technical results in the supplement.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
There is an extensive discussion on related work (partly in the supplement). It is clearly stated what was known before and what are the new contributions in the present work.

I was missing only one thing: there should be some references on the area of "beyond worst case analysis". Maybe you could add a survey on this topic and one or two examples of beyond worst case analysis among recent ICML or NeurIPS papers.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
some minor comments to improve the paper:

- add references for "beyond worst-case analysis"

- 89: is sr(A)-1 < k < sr(A) correct? if sr(A) is an integer, there is no value of k to satisfy this. maybe the upper bound should be rank(A) instead of sr(A) ?

- 92: O(sqrt(k)) to O(k) -> "O(sqrt(k)) to \Omega(k)"

-111: "k between 20 and 50" -> "k between 20 and 40"

- 169: explain parameters \nu and \ell.

- 171: note that Matern turns to RBF with \ell = \sigma the limit for nu -> \infty. Also note that this limit is approached very quickly so that in practice \nu is almost always bounded by a small constant 7/2 (Rasmussen & Williams, 2006), which supports even more that O(1+nu) is indeed a very good approximation factor!

- "sufficiently wide valley". There is a quantification of "sufficiently wide" in the lemmas. Maybe it would be nice to

have it also somewhere in the text, especially since the expression can become quite a large "constant" depending on eps.

- Figure 2: (top) -> (left), and (bottom) -> (right)

- Figure 2: is the min Phi_s(k) line taken over s\in{10, 15, 25} or over all values of s\in [0,rank(A)] ?

- In the proof of thm 3 I am missing one additional argument concluding wrt the "key challenge" described above lemma 4. What is the consequence of the simplices being orthogonal to one another?

9. Please provide an "overall score" for this submission.
9: Top 15% of accepted NeurIPS papers. An excellent submission; a strong accept.

10. Please provide a "confidence score" for your assessment of this submission.
5: You are absolutely certain about your assessment. You are very familiar with the related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This paper studies Column Subset Selection (CSS) problem. They proposed a new algorithm of which approximation ratio depends on the stable rank of the input matrix. They also provided the lower bound. Their upper and lower bounds show that for some matrices the CSS approximation factor can exhibit peaks and valleys as a function of the subset size k。

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The algorithm is novel. It is surprising that running Determinantal point processes (DPP) on a proper scaled matrix can give a good CSS with approximation factor depending on the stable rank. It is very clean and easy to implement. The construction of the hard instance is interesting. The combination of the upper bound and the lower bound reveals an interesting phenomenon that the optimal approximation ratio may be a multiple-descent curve.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
For the upper bound side, the main issue is that the analysis breaks for small k. It is not clear the behavior of the algorithm when k is a small constant.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
I did not find any explicit issue in the proofs or experiments.

5. Clarity: Is the paper well written?
The paper is clean and well written in general.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
In Appendix, authors gave detailed comparison with previous work.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
- line 47: should mention the definition of DPP is in appendix.
- I felt that the authors missed some related work. For example, "Algorithms for ℓ_p Low-Rank Approximation" (Chierichetti et. al., ICML'2017) gives a iterative column selection methods. An another algorithm "Optimal CUR Matrix Decompositions" (Boutsidis & Woodruff, STOC'2014) uses leverage score and adaptive sampling methods. It would be good that authors can discuss whether these algorithms can achieve the similar guarantees as the proposed algorithm or not.

9. Please provide an "overall score" for this submission.
7: A good submission; accept.

10. Please provide a "confidence score" for your assessment of this submission.
2: You are willing to defend your assessment, but it is quite likely that you did not understand central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The paper considers Column Subset Selection Problem (CSSP) (a variant also known as the Nystro¨m method) which selects a subset S of k columns from an m by n matrix, so that the difference bewteen the matrix and its projection on the span of S is small (in terms of Frobenius norm). It focuses on the ratio between this error and that of the best rank k approximation. It exploits the spectral properties of the matrix and gets better bounds than known worst-case bounds for matrices with good singular value decay. It also shows that the approximation ratio as a function of k has multiple peaks and valleys, and shows that this is inherent by providing a lower bound. Finally, it empirically verifies the phenomenon using the RBF kernel on real datasets.

pros:
+ The paper explains why the method works well in practice, much better than predicted by the worst-case analysis. It shows that the method makes good use of the spectral decay of the data, which is presented in many practical datasets.
+ It further observes the interesting multiple descent phenomenon of the approximation ratio for some matrices (i.e., with multiple significant drops in the spectral decay). Importantly, it shows that this is not an artifact of their analysis but rather an inherent property of the task on such matrices, by showing a lower bound that, together with the upper bound, justifies the phenomenon.

cons:
- The connection of multiple-descent phenomenon with the spectral phase transition has been observed in the previous works, as discussed in the paragraph of Line 128. While differences and similarities are mentioned, one commonality not discussed is that a bit overparameterization can help the learning significantly. Maybe the authors

can comment more on whether this help of overparameterization is due to the same/similar reasons as in previous works.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
+ The paper explains why the method works well in practice, much better than predicted by the worst-case analysis. It shows that the method makes good use of the spectral decay of the data, which is presented in many practical datasets.
+ It further observes the interesting multiple descent phenomenon of the approximation ratio for some matrices (i.e., with multiple significant drops in the spectral decay). Importantly, it shows that this is not an artifact of their analysis but rather an inherent property of the task on such matrices, by showing a lower bound that, together with the upper bound, justifies the phenomenon.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
- The connection of multiple-descent phenomenon with the spectral phase transition has been observed in the previous works, as discussed in the paragraph of Line 128. While differences and similarities are mentioned, one commonality not discussed is that a bit overparameterization can help the learning significantly. Maybe the authors can comment more on whether this help of overparameterization is due to the same/similar reasons as in previous works.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
yes

5. Clarity: Is the paper well written?
Yes

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes. I will apprecitate more discussion about the effect of overparameterization in the task.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
8: Top 50% of accepted NeurIPS papers. A very good submission; a clear accept.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #4

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This is a high-quality theoretical paper that provides unique insights on the origins of the multiple descent curve for

feature selection. The master theorem as function of the spectral decay and the lower bound helps us to understand why we can observe multiple peaks and descents of the approximation factor.

The supplementary material is also very useful providing excellent context and additional experiments.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
High quality rigorous paper that provides upper and lower bounds that helps us understand the qualitative and quantitative of the expected approximation factor.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
Although the paper is mostly theoretical I wish the authors could have shown the analysis on other data sets with different types of spectral decays. Having such an empirical analysis can educate us in what to expect in different situations.

If we compare Fig. 3 and the supplementary material for k-DPP and the greedy method only the enuite2001 spectral decay appears to have a differing approximation factor. Why?

The smaller eigenvalues are typically quite noisy and just by sorting then they may appear as having an exponential spectral decay. Yet, they are likely reflecting the sampling of an exponential distribution rather than the actual decay. Have the authors thought how to discard/correct for it?

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
I did not find obvious errors in the manuscript.

5. Clarity: Is the paper well written?
very clearly written.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
properly discussed.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
I enjoyed learning about the relationship of the spectral decay and the approximation factor. Thank you.

9. Please provide an "overall score" for this submission.
9: Top 15% of accepted NeurIPS papers. An excellent submission; a strong accept.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes