Response to Reviewers on "The spectral norm error of a simple CUR decomposition for positive semidefinite matrices"

Alex Gittens

[Manuscript ID 085880, Submitted 13 Dec 2011, Reviewed 29 Feb 2012, Revised 16 Dec 2012, Reviewed 13 Feb 2013, Revised 18 Mar 2013]

I. General Response and Overview of Revision

I thank the associate editor and the referees for their careful reading of my manuscript and the recommendations for improvements. In the following sections, I address the individual concerns of each reviewer and describe the corresponding amendments to the manuscript.

II. Detailed Response to Referee 2

(1) The author uses the term "Nyström extension" throughout the article. We thought the title should use "a simple Nyström extension" instead. The term CUR can be listed as one of the keywords for the paper.

Changed accordingly.

(2) A CUR decomposition doesn't always assume that $\mathbf{U}$ is the "generalized inverse of the overlap of $\mathbf{C}$ and $\mathbf{R}$." It could also be obtained by projecting $\mathbf{A}$ onto $\mathrm{range}(\mathbf{C})$ and $\mathrm{range}(\mathbf{R}^T)$.

The explanation of CUR approximants has been updated.

(3) *(Section 1.3)*. We can quote Golub and van Loan's book for the distance between subspaces.

A reference has been added.

(4) *(Section 1.4)*. The notion of coherence dates back to much earlier work such as Donoho and Elad's "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization" in 2002.

In the setting of the paper mentioned, the dictionary comprises unit length vectors and the coherence that the authors use measures the maximal inner product of the dictionary elements. To the best of my knowledge, Candes and Recht were the first to introduce the particular notion of coherence that I use here.

(5) *(bottom of page 3)*. Is the orthonormal matrix $n \times k$ instead of $k \times n$?

Yes, thank you. This has been corrected.

(6) *(bottom of page 3)*. At the same place, we can state that it is Lemma 2.2 of [CR09].

The reference has been modified accordingly.

(7) *(Algorithm 1)*. Should we return the coupling matrix $\mathbf{W}_\rho^{-1}$ instead of $\mathbf{W}_\rho$? It seems more natural that we return the coupling matrix instead of its inverse. The same applies to other algorithms in this paper.

I chose to return the inverse because the matrices $\mathbf{C}$ and $\mathbf{W}_\rho^{-1}$ implicitly represent the Nyström extension without need for any further manipulations.

(8) *(beginning of page 5)*. Typo: "that that" should be "that is"

Thank you. Corrected.

(9) *(Algorithm 2)*. Perhaps we can dispense with checking whether $\lambda_{\text{min}}(\mathbf{W}) \geq \rho$. The bounds obtained would be similar.

The bound is the same, but one can imagine the empirical errors would not be, depending on the actual value of ρ. For example, if ρ is large, then one would not want to use the regularization $\mathbf{W}_\rho = \mathbf{W} + \rho\mathbf{I}$ unnecessarily.

(10) *(Theorem 1)*. The bound using Markov's inequality is not very strong but we appreciate the idea of dealing with the case where the singular values decay much faster. This new addition to the paper is however similar to what is considered in [CD].

The use of Markov's inequality is standard when dealing with randomized matrix multiplies (in this case, the product $\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}$). I was motivated to look at this way of incorporating spectral decay by the somewhat conflicting facts that that the dependence on $\sqrt{n/\ell}$ of my original bound is optimal in the worst case, but the error is much smaller in practice on real datasets. I agree this is similar to the bound given in [CD]; I have mentioned this in the relevant literature section.

(11) *(page 6)*. Regarding [DKM06], the no. of columns c scales with $1/\delta$, not $\log \delta^{-1}$.

The bounds I cite come from equations (22) and (23) of Theorem 4 in [DKM06], in which c scales with $\log \delta^{-1}$ and r scales with $1/\delta$.

(12) *(page 6)*. Quote Theorem 3 of [DM05].

I have added a reference to this result to the relevant discussion.

(13) We feel that too much detail is given in the section on previous work. Also, how does the previous work relate to your work? You can say the algorithms are slow or the error bounds are not as good as yours.

Originally this section was shorter; I added detail in response to previous referees' requests that this section be made more expository than standard because the average SIMAX reader is perhaps not familiar with this body of work.

The two major differences between previous work and this work is that this establishes an equivalence between Nyström extensions and the column subset selection problem, and is the first to provide truly relative error spectral norm bounds (concurrent with [CD].)

I have tried to emphasize the contributions of this work at the conclusion of the relevant literature section.

(14) It is mentioned in the footnote of [CD] that the notion of coherence can be extended in at least two ways. Clearly, they chose a simpler definition for the sake of exposition. The same applies to [KMT09a]. There is no need to criticize their definition of coherence.

I have removed this criticism.

(15) The constants in [CD] are about as small because they are based on similar probabilistic results, namely Rudelson, Vershynin's work or Tropp's improved analysis. The probability bounds are stated in a different form. In a lot of works, the failure probability is given as $1/n$ or $1/n^2$ just to convey that a $\log n$ factor in the running time suffices to reduce the failure probability to $1/\text{poly}(n)$. It is a matter of taste and the criticism is too harsh.

I have removed this criticism.

(16) *(Theorem 2)*. We can emphasize that (5) contains an equality. Also, instead of writing $\mathbf{\Sigma}_1$ being singular, we can write $\mathbf{A}$ having rank $< k$, which might be clearer. We also prefer that the remark that "$\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}$ if $\text{rank}(\mathbf{A}) < k$" to be made outside the theorem.

The succeeding remark has been modified to emphasize the equality, and the sufficient condition for exact recovery is now stated in terms of the rank of $\mathbf{A}$. The location of the exact recovery condition was not changed, since the claim is most naturally proven along with the rest of the theorem.

(17) *(the end of Section 3.1)*. We do not necessarily need to compute the square root of $\mathbf{A}$. We can also project $\mathbf{A}$ onto $\text{range}(\mathbf{C})$ on both sides which can be much faster if $\mathbf{A}$ has a sparse representation or a closed form.

As far as I can tell, this scheme still requires forming $\mathbf{A}^{1/2}$ since to form $\mathbf{C}$ you first have to obtain the sampling matrix $\mathbf{S}$, which requires knowledge of the RRQR factorization of $\mathbf{A}^{1/2}$.

(18) *(Lemma 1)*. You can emphasize that Lemma 1 is new because it involves column sampling without replacement. That said, Lemma 1 is an easy consequence of [Tro11] and the reader can be referred to some other proof, e.g., Rudelson, Vershynin's work, on how to apply [Tro11]'s result to prove Lemma 1.

True, Lemma 1 is not technically difficult. But given the ease and shortness of the argument, I feel it is good style to provide it in full.

(19) Regarding the example in [BDMI11], which supplies the lower bound for column selection, is the basis coherent? The $\sqrt{n/\ell}$ factor could come from the growing coherence. In other words, it is not clear that this example implies that the relative error bounds in Theorem 3 are optimal.

My claim of optimality is limited to this assertion: that Theorem 3 is a worst-case bound with optimal dependence on n and ℓ. The [BDMT11] matrix demonstrates this claim. I expect that indeed the $\sqrt{n/\ell}$ factor only shows up when the dominant eigenspaces are quite coherent, and a bound which replaced the $\sqrt{n/\ell}$ with something that depended explicitly on the coherence of the matrix (or some other summary characteristic) would probably be optimal in a more completely satisfying sense.

Currently, the coherence enters in Theorem 3 only in saying how large the number of samples needs to be before the bound applies. I do not claim that this dependence is optimal (although this number of samples is required for $\mathbf{\Omega}_1$ to have full row rank so Proposition 1 applies, there may be some other way of establishing the bound without requiring $\mathbf{\Omega}_1$ to have full row rank).

(20) *(Proposition 3)*. We need to mention the assumption that $\mathbf{W}$ is nonsingular.

This assumption has been added.

(21) *(Theorem 1)*. The perturbation argument in the proof of Theorem 1 is similar to what is done in [CD]. Even though it is not very hard, some acknowledgment would be welcome.

I have corrected this omission (their work did motivate this argument).

(22) *(Equation (8))*. Should the last term be 4ρ, not 2ρ?

Corrected.

(23) We can describe briefly how (8) leads to the final bound in Theorem 1. Remind the reader that e_ρ is not just the RHS of Equation (7) but contains some ρ factors.

Done.

(24) On page 16, can we elaborate on the statement “the no. of samples needed to obtain a small relative error is much less sensitive to the coherence”? What do we mean by “less sensitive”? Do you mean that the simple Nystrom extension works well for dense matrices over a wider range of coherence? Do you mean that the simple Nystrom extension fails abruptly for a sparse $\mathbf{U}$ when the coherence exceeds a threshold? Will a plot with varying coherence instead of varying ℓ be more informative here?

I mean that the Nystrom extensions give lower errors for a fixed coherence and number of column samples when the eigenvectors are dense than when the eigenvectors are sparse. This has been clarified. It may also be the case that there is some coherence value that depends on the sparseness of $\mathbf{U}_1$ above which the Nystrom extensions have significantly higher error— this is an interesting suggestion—, but I do not pursue this investigation.

(25) In Figure 4, what happens when ρ is very small, like 10^{-15}? From numerical experiments that we conducted ourselves, the error of Algorithm 2 seems to blow up for very small ρ as well. However, we don’t see anything wrong with the proof. For your convenience, we included the code on the last page. Please comment on the output of the attached code.

For too small ρ, $\mathbf{W}_\rho$ can be ill-conditioned. You can see this in the output of the code you provided: if you overplot the error of the simple Nyström extension, you see that as ρ decreases, first the regularized approximants have errors that are orders of magnitude smaller than the simple extension because the regularization counteracts the ill-conditioning, but as ρ continues to decrease, the regularization becomes ineffective and the errors of the regularized approximants begin increasing until they are the same as the error of the simple extension.

I have modified the relevant experiment and Figure 4 to demonstrate this phenomenon.

(26) Please do more than 20 runs for the numerical experiments.

I have increased the number of runs to 60.

III. Detailed Response to Referee 3

(1) SIAM has a policy of not publishing color figures unless they are necessary, and in this paper I do not think color is necessary.

The figures have been changed to gray scale.

IV. Detailed Response to Referee 4

(1) The condition numbers in Section 5 (and section 1.2, 3.1 and 6) need to be of the form $\lambda_1(\mathbf{W})/\lambda_k(\mathbf{W})$ for a $k \times k$ spd matrix $\mathbf{W}$. See the perturbation bounds for linear system solution in [GV96]. It is not sufficient to merely bound the smallest eigenvalue from below.

I left implicit the fact that $\lambda_1(\mathbf{W}) \leq \lambda_1(\mathbf{A})$ could be used to bound the condition number of $\mathbf{W}$. I have updated the manuscript to make this clear. I have also noted that one could estimate $\lambda_1(\mathbf{W})$ using other matrix concentration results if one were so inclined.

(2) In Figures 2 and 3 the relative errors are all greater or equal to one. Therefore none of the results have any accuracy at all. In general, a relative error $\leq 5 * 10^{-p}$ implies that the largest elements in magnitude have at least p correct digits.

There is no concept of accuracy here because there is no claim that anything more specific than a low-rank approximation is being produced. The goal of Nyström extensions is to obtain low-rank approximations whose errors are comparable to those of the optimal SVD approximants, so the relative error being plotted is the reconstruction error of the Nyström extension over the reconstruction error of an optimal rank-k approximant. I have clarified this in the paper.

Originally the relative errors in Figures 2 and 3 were massive because I chose an unfavorable regime. In this revision, the parameters have been modified to illustrate a regime that is more representative of a typical application of the Nyström extension (insofar as this is possible with synthetic data).

(3) Please give explicit expressions for what is plotted on the vertical axes of Figures 2 and 3; and cite equation numbers for the bounds you are plotting, or else restate the bounds.

I have done so. (No bounds were being plotted, only empirical relative errors)

(4) *(Page 1, section 1, 1. paragraph).* Please give primary-source references for the operation count of computing truncated SVDs of dense matrices by Lanczos methods. Why would you use a Krylov space method for a dense matrix?

When k is small, Krylov space methods are an efficient approach (e.g. this is how Matlab's `svds` computes truncated SVDs of dense matrices) in practice. When k is on the order of n, it may be cheaper in practice to switch over to using a truncated SVD computed from a RRQR factorization, so I mentioned the cost of computing such SVDs also.

As far as I am aware, there are no meaningful operation counts for computing SVDs of sparse or dense matrices with Krylov methods; convergence depends in a complicated way on the specifics of the matrix and the parameters of the method. I am only aware of operation counts for the individual iterations of some methods. However, to motivate looking at the simple Nyström extension as an attractive route to low-rank approximants, it seems sufficient to note that Krylov methods all cost at least kn^2 operations, the cost of k dense matrix-vector multiplies.

(5) *(Page 5, Algorithm 2).* Why is it not $k \geq \operatorname{rank}(\mathbf{A})$, as in Algorithm 1?

No restrictions are necessary. I have updated the statements accordingly.

(6) *(Page 9, section 3, 2. paragraph).* Minimizing the residual is only one of several possible criteria for subset selection, see [IC94]. The order of authors in this reference is wrong.

I have corrected the citation order and updated the manuscript to reflect this observation.

(7) *(Page 9, (3)).* Please add that $\mathbf{\Sigma}_1$ contains the k largest eigenvalues of A.

Done.

(8) *(Page 10, line 1).* In the range inequalities, the font for ℓ is not correct.

Corrected.

(9) *(Page 13, line 4)*. You cannot say that Theorem 3 ”promises exact recovery” because the bounds are only probabilistic.

The claim has been appropriately qualified.

(10) *(Page 14, proof of Proposition 4)*. Since you are writing $(\mathbf{SAS})^{-1}$, you have to add the assumption that $\mathbf{W}$ is nonsingular.

This assumption has been added.