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]

I. General Response and Overview of Revision

I thank the associate editor and the three reviewers for their careful reading of my manuscript and the detailed recommendations for improvements.

Several parts of the paper have been revised accordingly. Particularly, I’ve discussed the potential instabilities of the simple Nyström algorithm and provided regularized algorithms which are stable. A section showing the numerical efficacy of these algorithms has been added. I have reorganized the expository material, been more explicit with my discussion of the previous results on Nyström extensions, and addressed the question of how Nyström extensions formed using standard column selection techniques perform (as compared to the randomized column selection used in the simple Nyström scheme). Additionally, the dependence of the relative error bounds on n and ℓ has been shown to be optimal.

In the following three sections, I address the individual concerns of each reviewer and describe the corresponding amendments to the manuscript.

II. Detailed Response to Reviewer 1

(1) In the first page, I would like to see more details regarding the computation of a truncated SVD in $\mathrm{O}(kn^2)$ time; I feel that this statement is rather simplistic and will not appeal to the readers of SIMAX.

I have now specified that the computation of truncated SVDs requires at least $\Omega(kn^2)$ time since Krylov subspace methods are used, and alluded to the difficulties involved with estimating their precise cost. That seems sufficient to motivate the consideration of methods which are $\mathrm{o}(kn^2)$.

(2) Also, in the first equation in Section 1.4, the author might want to discuss that the diagonal elements of the matrix $\mathbf{P}_{\mathcal{S}}$ are the so-called leverage scores of the matrix $\mathbf{A}$, when $\mathcal{S}$ is a rank-k dominant invariant subspace of $\mathbf{A}$.

I have now noted this observation.

(3) Remark 3 in page 8 might fit better in an conclusion/open problems section.

I have now established that the relative error bound is sharp in n and ℓ.

III. Detailed Response to Reviewer 2

(1) As is stated in the paper, the main results are almost implied elsewhere.

This manuscript’s assumptions and approach differ significantly from that of [CD], which considers CUR approximate decompositions. It would be more accurate to say that these error bounds are consistent with those in [CD], in that we consider two related problems (I note that the two papers were released to ArXiv at around the same time) and both find that the errors scale like $\mathrm{O}(n/\ell)$ where n is the size of the matrix, and ℓ is the number of column samples.

The main contribution of this paper as the elucidation of the connection between Nyström extensions and the column subset selection problem (a connection which does not exist in the case of the general CUR problem), as well as a clear presentation of the use of a matrix exponential probability inequality to bound the required random quantities. The latter may be of interest to SIAM readers as an illustration of how handy these exponential bounds are for getting quantitative estimates of the spectra of random matrices which come up in the analysis of randomized algorithms. The former is important because this connection is potentially algorithmically fruitful and provides a framework for analyzing a variety of other Nyström extension schemes, e.g. sampling by leverage score or sampling from random mixtures of the columns. As another example of the fruitfulness of the connection between the Nyström extensions and the column sampling problem, it allows one to demonstrate that the dependence of the relative spectral norm error bound on n and ℓ is optimal.

In [CD], the authors make the assumption that all n of the eigenvectors of the matrix are flat— which is reasonable for the particular physically motivated problems they are interested in, but is perhaps overly restrictive in data processing— and then find the asymptotic order of the number of measurements required to achieve a certain order of error with nonquantitative guarantees on the failure probability (namely, 'with high probability', but due to elided constants in their explanation, it is unclear what this means). Nor is it hinted that there are situations in which one can expect exact recovery (that the Nyström extension is in fact the original matrix). Further, it is not clear how to extend their results to deal with more intricate sampling schemes like sampling by leverage score.

In this paper, I make what are arguably the minimal reasonable assumptions about the eigenvectors of the matrix: namely that the dominant k-dimensional eigenspace is incoherent. I provide explicit estimates, with small constants, for the number of column samples required to achieve an exact amount of error, with clearly stated quantitative guarantees on the failure probability. I also show that exact recovery occurs with constant probability if the original matrix is exactly low-rank. The analysis is for the sampling scheme where columns are chosen uniformly without replacement, which is of more practical interest than sampling with replacement. Throughout this analysis, it is clear how the positivity of the matrix is being exploited.

I note that for this revision, since it was pointed out that the simple Nyström extension is only conditionally stable, I've used a result from [CD] to analyze the error of two algorithms I propose for computing regularized Nyström extensions (my analysis shows that these algorithms have better error guarantees than the regularized CUR algorithm suggested in [CD], as the positivity allows me to refine the result from [CD] — however, the [CD] algorithm has slightly better performance in practice).

(2) There is no attempt of any kind by the author to supply numerical evidence that supports the results in this paper.

I have addressed this concern by providing numerical studies of the errors of the simple Nyström extension, the regularizations I've proposed, and the regularization proposed in [CD] to select the columns.

IV. Detailed Response to Reviewer 3

(1) On page 8, at the end of the proof of Theorem 2, should it be $\|\mathbf{\Omega}_2\|_2 \leq \|\mathbf{U}_2\|_2\|\mathbf{S}\|_2$?

Yes, thank you. I've corrected the typo.

(2) May want to cite Rudelson and Vershynin's "Sampling from large matrices" which I believe is the first paper with bounds like Proposition 2.

I have noted the connection.

(3) Lowercase "t" is not widely used for transpose

I have changed to a capital "T".

V. Detailed Response to Reviewer 4

(1) *(Regarding Theorem 1)*. The connection to subset selection holds for any generalized Cholesky factor.

I have decided to stick with the matrix square root in the statements, for definiteness, but I have pointed this fact out for the readers.

(2) *(Regarding Theorem 1)*. For the inequality to hold, you need to assume that $\mathbf{\Sigma}_1$ is nonsingular, see the proof of Theorem 9.1 in [HMT11].

Theorem 9.1 in [HMT11] doesn't require $\mathbf{\Sigma}_1$ to be nonsingular (the authors address this point in the proof: they show that if $\mathbf{\Omega}_1$ has full row rank and $\mathbf{\Sigma}_1$ is singular, then you actually get exact recovery).

(3) *(Regarding Theorem 1)*. Considering that you are viewing the problem from the point of view of subset selection, how does your bound compare to the best subset selection, i.e. RRQR bounds?

Thanks for suggesting this. I've compared the Nyström extension you would get by selecting the columns from $\mathbf{A}$ using the RRQR factorization of $\mathbf{A}^{1/2}$ or some other generalized Cholesky factor. The error bound is weaker than that of the simple

Nyström extension when the matrix has an incoherent dominant singular space, but the bound is deterministic and the algorithm is stable.

(4) *(Regarding Theorem 2)*. You need to add the assumption $k \geq \operatorname{rank}(\mathbf{A})$. See the remark above for Theorem 1.

Did you mean $k \leq \operatorname{rank}(\mathbf{A})$ so that $\boldsymbol{\Sigma}_1$ is nonsingular? Either way, this is not a necessary requirement: when $k \leq \operatorname{rank}(\mathbf{A})$, the matrix $\boldsymbol{\Sigma}_1$ is nonsingular. When $k > \operatorname{rank}(\mathbf{A})$, the matrix $\boldsymbol{\Sigma}_1$ is singular, but as mentioned above, you get exact recovery in this situation if ℓ is sufficiently large that $\boldsymbol{\Omega}_1$ has full row rank. I've changed the theorem statement to make this clear.

(5) *(Regarding Theorem 2)*. Your bounds do not seem to reflect the fact that you are sampling without replacement. If $n = \ell$ then, without replacement, $\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 = 0$ but the bound is $\lambda_{k+1}(\mathbf{A})(1 + \varepsilon^{-1})$.

In this case, $\boldsymbol{\Sigma}_{\mathbf{2}} = 0$ and $\boldsymbol{\Omega}_1 = \mathbf{U}_1^T$ has full row rank, so as mentioned above you get exact recovery.

However, yes, the bounds don't directly reflect the fact that the sampling model is without replacement. Ideally I would like to use a matrix analogue of the classical tight bounds for sampling without replacement (e.g. Serfling 1974, *Probability inequalities for the sum in sampling without replacement*), but to my knowledge there are no such matrix exponential inequalities. The natural alternative, concentration bounds (e.g. Cortes et al. 2008, *Stability of transductive regression algorithms*) require an estimate of the expectation of the kth eigenvalue of the sum of random matrices: it is difficult to come up with a nontrivial estimate of this (at least one that gives a better bound that the Chernoff bound I've used here), since λ_k is not concave on $\ell \times \ell$ matrices.

(6) *(Regarding Lemma 1)*. I do not understand why $\mathbb{E}\mathbf{X}_1 = \frac{1}{n}\sum_{j=1}^{n} \mathbf{u}_j\mathbf{u}_j^T$ when you sample without replacement (I see that it is true when you sample with replacement). When you sample uniformly without replacement then $\mathbf{X}_1$ is picked with probability $1/n$. Then there are only $n-1$ matrices left, so uniform sampling picks $\mathbf{X}_2$ with probability $1/(n-1)$, etc.

The $\mathbf{X}_i$ are identically distributed (although not independent), so $\mathbb{E}\mathbf{X}_1 = \ldots = \mathbb{E}\mathbf{X}_\ell$. To calculate this expectation, use the fact that $\mathbf{X}_1$ is drawn uniformly at random from $\mathcal{X} = \{\mathbf{u}_j\mathbf{u}_j^T\}$ (where $\mathbf{u}_j$ is the jth column of $\mathbf{U}_1^T$) without replacement: the probability of choosing $\mathbf{u}_j$ is $1/n$ for all j, so

$$\mathbb{E}\mathbf{X}_1 = \sum_{j=1}^{n} \frac{1}{n}\mathbf{u}_j\mathbf{u}_j^T.$$

(7) *(Regarding Proposition 2)*. These bounds are the same as the ones for independent random psd matrices in Remark 5.3 of Tropp's Found. Comput. Math. paper *User-friendly tail bounds for sums of random matrices.* So I don't see where the sampling without replacement comes in.

As I mentioned above, there aren't any bounds directly available for sampling matrices without replacement. This bound (taken from Tropp's Adv. Adapt. Data Anal. paper *Improved analysis of the subsampled randomized Hadamard transform*) gives exactly the same guarantees as the bound you mention that is stated for sampling with replacement because it connects the two sampling models using ideas due to Hoeffding. It's a different bound only in the sense that the sampling model is different.

(8) *(Regarding the exposition)*. Please specify all your constants... Please explain exactly on what the constants depend.

I've now specified all my constants.

(9) *(Regarding the exposition)*. Please make complete and meaningful connections to existing literature.

This has been addressed.

(10) *(Please state the results in the order in which they are used, and put the proofs in right after the results. Especially with Theorem 2, Lemma 1, and Proposition 2 it is not clear what depends on what.)*.

I've reorganized the statements of the results to improve clarity.

(11) *(Abstract)*. Why are your error bounds relative?

I've now specified that by 'relative' I mean 'relative (in a multiplicative sense) to the optimal low-rank error'.

(12) *(page 1, section 1)*. Please give representative references to document the interest in low rank approximations, and CUR decompositions

I've added material to the introduction demonstrating the ubiquity of low-rank approximations and the usefulness of the CUR decompositions, and Nyström extensions in particular.

(13) *(page 2, section 1.2).* Are the matrices real or complex?

I've now specified that for simplicity all the matrices are real.

(14) *(page 2, section 1.3).* Please explain what you mean by "unique" invariant subspace.
Should $\mathbf{U}_k$ and $\tilde{\mathbf{U}}_k$ have orthonormal columns, instead of just orthogonal columns?
Please define $\mathbf{P}_{\mathbf{U}_k}$.
Please add that the matrix $\mathbf{M}$ is Hermitian, or real symmetric.

I have rewritten this portion to make it clear that I mean there is a gap between λ_k and λ_{k+1} so that the dominant k-dimensional eigenspace is well-defined. I corrected 'orthogonal' to 'orthonormal' and defined the projection matrices I refer to. I've also specified that $\mathbf{M}$ is PSD.

(15) *(page 2).* Exactly which theorem in section VII.3 of [Bha97] are you citing? Please note that the $\sin\Theta$ theorem requires conditions on the spectra of the matrices.

I've now specified which result I used from Bhatia. It is not the $\sin\Theta$, but rather a closely related result that holds for normal matrices.

(16) *(page 3, line 2).* Please explain what C is. This inequality is always true, with $\mathrm{C} = \left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2 / \lambda_j(\mathbf{A})$.

I've removed this statement.

(17) *(page 3, 2 lines above section 1.4).* You can "sanction" the use of the Nyström approximation only if you also show that the floating point computation of the approximation is numerically stable.

I've removed this statement. Instead, I point out that this matrix perturbation result gives us a further reason to be interested in a tighter spectral norm error bound on Nyström extensions.

(18) *(page 3, section 1.4).* Please give references for the original definition of coherence. You cannot assume that SIMAX readers are familiar with this concept. Here you are defining coherence for a subspace, while on page 5 (section 2) you are dening coherence for a matrix. They are the same, and I dont understand why you are making a distinction. Why is the coherence not simply called μ?

I've directed the readers to the matrix completion literature for the original definition of coherence, simplified the notation for coherence, and made the connection between the coherence of a subspace and the coherence of a orthonormal basis for that space explicit.

(19) *(page 3, Corollary 1).* Don't you need to assume that $\lambda_k(\mathbf{A}) > \lambda_{k+1}(\mathbf{A})$?
Why don't you call the coherence μ instead of τ? If you used the same symbol for coherence throughout the whole paper, then the results would be easier to parse.
Please add that $\delta > 0$.
What is the base of the logarithm in the bound for τ?

There does not need to be a spectral gap (because Proposition 1 from [HMT11] does not require one). If there's no gap, $\lambda_{k+1} = \lambda_k$ so you end up with the reasonable result that the error is relative to the optimal rank $(k-1)$ approximation. I've switched to using μ throughout the paper for coherence, added the stipulation on δ, and specified that all logarithms are natural throughout the paper.

(20) *(page 3, line above section 1.5).* What is C in $\mathrm{C}k\log k$? Is it the same C as on the top of the page?

The constant is now explicit.

(21) *(page 3 and section 1.5)*. The purpose of a literature review is to summarize results so that people not familiar with the topic are able to understand them. You cannot assume that SIMAX readers know these probabilistic results. Please state the bounds and their assumptions explicitly.
What do you mean by "additive factor" and "relative Frobenius norm guarantee"? What are "incoherent" singular vectors? What is a "coupling matrix" and a "truncation level"?
In particular the results in [CD] should be presented carefully, because of their strong connection to yours.
In the context of [KMT09b], what do you mean by "constant probability". Please state all assumptions for this bound. What are the Cs in this paragraph? I am getting confused now with all these Cs. How does μ differ from μ_0 on page 2?
What exactly is "the randomized methodology espoused in [HMT11]"?

The statements of the bounds are now given explicitly, with failure probabilities. I have been careful to explain the terms I use. The comparison of my approach with that taken in [CD] has been expanded: essentially, [CD] requires a more restrictive condition on the eigenvectors of the matrix, and the spectral norm error bound given for their regularized CUR decomposition has an additional term (which may be large) that is not present in the error bounds for my algorithms. The difference is due to the fact that their results do not take advantage of the structure implied by the positivity. Of course, the results of [CD] also do not expose the connection between Nyström extensions and the column subset selection problem.

(22) *(page 5, (2))*. What is C? The ℓ in the first bound has the wrong font. Where is the dependence on p and q.? Please explain p and q carefully.

I have explained p and q. I hope it is now clear why there is no dependence on p and q in the given bounds: I substituted the particular values of p and q that correspond to the simple Nyström scheme into the bounds given in [LKL10] in order to see how their results compare to mine for this scheme.

(23) *(page 5, 2 lines below (2))*. Please define "$\succeq$". Is it the same as PSD?

Yes. In the interest of clarity I have removed that notation.

(24) *(page 5, section 2)*. A Hermitian positive-definite matrix has a unique Hermitian positive definite square root. Is this also true for a semi-definite matrix? See the book on matrix functions by Higham.

Yes, since the zero eigenvalues are semisimple.

(25) *(page 5, section 3)*. There has been a lot of work on column subset selection in numerical linear algebra (rank revealing QR factorizations), and the criteria are well defined. Please refer to the work by Eisenstat, Gu, Chandrasekaran, Hong, Pan, etc.

I have added references to the literature on the RRQR approach to column subset selection, in the context of considering the quality of the Nyström approximation you would get from using the RRQR procedure to select the columns.

(26) *(page 6, line 9)*. Please define "randomly uniform permutation matrix".
Please add that $[\mathbf{U}_1\mathbf{U}_2]$ is an orthogonal matrix.

I have clarified the terminology: rather than "randomly uniform permutation matrix", I use the clearer "matrix sampled uniformly at random from the set of all permutation matrices." I've specified that the matrix is orthogonal.

(27) *(page 6, Proposition 1)*. In line 1, ℓ is in the wrong font. It would be easier to just refer to the partitioning in (3) instead of stating everything over again.
As remarked above for Theorem 1, you need to assume that $\mathbf{\Sigma}_1$ is nonsingular.

Thank you for spotting the font changes. I have referred back to the partitioning in (3).

(28) *(page 7, (6) and (7))*. How do they differ from (3) and (4)? I don't see why you need to restate this.

There is no difference, so I removed the redundancy.

(29) *(page 7, Remark 1)*. It would be more helpful if you moved this to page 3, where you first introduce incoherence.

I agree; I moved it.

(30) *(page 8, proof of Theorem 2, line 3)*. Please add "with probability at least $1 - \delta$" for the full row rank of $\mathbf{\Sigma}_1$.

I did so.

(31) *(page 8, Proposition 2)*. Do you mean to say that $\mathcal{X}$ contains matrices of dimension $k \times k$? Shouldn't $\mathcal{X}$ contain at least ℓ matrices?

Yes. I've corrected the statement of this Proposition.