Eigenvalues of a matrix in the streaming model

Alexandr Andoni[*] Huy L. Nguyễn[†]

Abstract

We study the question of estimating the eigenvalues of a matrix in the streaming model, addressing a question posed in [Mut05]. We show that the eigenvalue "heavy hitters" of a matrix can be computed in a single pass. In particular, we show that the ϕ-heavy hitters (in the ℓ_1 or ℓ_2 norms) can be estimated in space proportional to $1/\phi^2$. Such a dependence on ϕ is optimal.

We also show how the same techniques may give an estimate of the residual error tail of a rank-k approximation of the matrix (in the Frobenius norm), in space proportional to k^2.

All our algorithms are linear and hence can support arbitrary updates to the matrix in the stream. In fact, what we show can be seen as a form of a bi-linear dimensionality reduction: if we multiply an input matrix with projection matrices on both sides, the resulting matrix preserves the top eigenvalues and the residual Frobenius norm.

1 Introduction

The past few decades have seen a surge of research on sampling and sketching matrices for the purpose of producing low-complexity approximations to the large matrices. Most notably, there has been an influential line of work on computing a low-rank approximation of a given matrix, starting with the work of [PTRV98, FKV04]. In the extreme case of rank $k = 1$, one recovers an approximation to the eigenvector corresponding to the largest eigenvalue (or large eigenvalues). Classically, one can compute such low-rank approximations via singular-value decomposition (SVD). Since the matrices of interest and the application areas are typically in the realm of "massive datasets"[1], it is imperative to develop algorithms that are very efficient for processing such large matrices.

Massive data computation typically have a space and/or communication bottleneck, and thus it is crucial to consider these algorithms in the *streaming* framework, where the data (matrix) is stored externally (or in a distributed fashion), and the algorithm passes through the data sequentially, potentially a few times. The main consideration here is the working/RAM space, which we would like to make as small as possible. For example, for rank approximation algorithms, the space complexity is usually proportional to $n \cdot k$ for a $n \times n$ matrix and a target rank k. Besides squashing the space, at least part of the research has been also driven by the need to reduce the number of passes from constant (or logarithmic) to one pass only, as well as to achieve relative error guarantees (in contrast to absolute error guarantees). See the surveys [KV09], [HMT10], and [Mah11] for a thorough treatise of the large number of existing algorithms in this area.

It is natural to ask whether one can obtain space below n. The simple answer is no: $\Omega(nk)$ space is required to even represent the output in general. Low-rank approximation tries to capture the highest k *eigenvectors* of a matrix, which are k large-dimensional vectors. However, it is conceivable that one can achieve less space if we want to compute the top k *eigenvalues/singular values* only, in which case the output size $O(k)$. This question of estimating eigenvalues has been also asked in one of the first surveys on streaming algorithms [Mut05].

In this paper, we present the first such space-efficient algorithm that computes approximations to the top eigenvalues and singular values. Informally stated, we show how to approximate the top k eigenvalues, with additive error proportional to the residual error of the rank-k approximation. We achieve space that is proportional to k^2 (and error parameters).

We also show how to use our algorithm for computing the residual error of a rank-k approximation of a matrix, in space proportional to k^2. Such an algorithm may be applied, for example, in the following parameter selection scenario. Before applying one of the more expensive rank-k approximation algorithms, one may wish to know whether the target value of k will yield a good approximation, or one has to, say, increase k for a smaller residual error. Our algorithm would provide such an estimate at much smaller (space) cost, of roughly k^2, in contrast to roughly nk required by the best rank-k methods.

We remark that our results may be seen as a form of "heavy hitters" on the vector of eigenvalues/singular values of a matrix. We note that, in constrast to the

[*] MSR Silicon Valley.

[†] Princeton University. Supported by NSF CCF 0832797 and a Gordon Wu fellowship.

[1] Or "*big* datasets" in the ultra-modern speak.

standard heavy hitters, which achieve an equivalent error in space proportional k, our space requirement is proportional to k^2. It turns out that in our case, of matrices, dependence on k has to be at least $\Omega(k^2)$ [CW09]. Except for this difference in space requirement, our error guarantees are morally equivalent to the best guarantees for the (easier problem of) standard heavy hitter for vectors.

1.1 Results We now present our results formally. We state our eigenvalue/singular value reconstruction results as solutions to ℓ_1 and ℓ_2 heavy hitters problems. Our error bounds are in terms of the rank-k residual error (in constrast to the total eigenvalue mass), similar to the most desirable bounds one can obtain for a full-blown rank-k approximation. In the space bounds below, we do not include the random seed size, which we comment on later on.

THEOREM 1.1. (ℓ_1 HEAVY EIGEN-HITTERS) *Fix $\epsilon > 0$ and integers $n, k \geq 1$. Let A be a real symmetric $n \times n$ matrix, and let $\lambda_i(A)$ be the i^{th} largest eigenvalue of A in absolute value. Then there is a linear sketch of the matrix A, using space $O((k\epsilon^{-2} + \log n)^2)$, from which one can produce values $\tilde{\lambda}_i$, for $i \in [k]$, satisfying the following with at least $5/9$ success probability:*

$$|\lambda_i(A) - \tilde{\lambda}_i| \leq \epsilon|\lambda_i(A)| + \tfrac{1}{k} S_1^{k+1},$$

where S_1^{k+1} is the residual "ℓ_1 error":[2] *$S_1^{k+1} = \sum_{i>k} |\lambda_i(A)|$.*

We note that while the above statement applies to square symmetric matrices, it immediately extends to computing the approximate singular values of arbitrary matrices.

We also solve the following ℓ_2 heavy hitter problem for eigenvalues.

THEOREM 1.2. (ℓ_2 HEAVY EIGEN-HITTERS) *Fix $\epsilon > 0$ and integers $n, m, k \geq 1$. Let A be any $n \times m$ matrix, and let $s_i(A)$ be the i^{th} largest singular value of A. Then there is a linear sketch of the matrix A, using space $O((k\epsilon^{-2} + \log n)^2)$, from which one can produce values $\tilde{s}_i$, for $i \in [k]$, satisfying the following with at least $5/9$ success probability:*

$$|s_i^2(A) - \tilde{s}_i^2| \leq \epsilon s_i^2(A) + \tfrac{1}{k} (S_2^{k+1})^2,$$

where S_2^{k+1} is the residual "ℓ_2 error" (Frobenius norm): $S_2^{k+1} = \sqrt{\sum_{i>k} s_i^2(A)} = \min_{A' \text{ of rank } k} \|A - A'\|_F$.

We note that, in constrast to the ℓ_1 heavy eigen-hitters result from above, our ℓ_2 heavy eigen-hitter result does not reconstruct the sign of the eigenvalues in the case of a square symmetric matrix.

Finally, we state our result for estimating the residual error of a rank-k approximation.

THEOREM 1.3. (RESIDUAL ERROR ESTIMATION) *Fix $\epsilon > 0$ and integers $n, m, k \geq 1$. Let A be any $n \times m$ matrix, and let $s_i(A)$ be the i^{th} largest singular value of A. There is a linear sketch of the matrix A, using space $O((k\epsilon^{-3} + \log n)^2)$, from which one compute an $1 + \epsilon$ approximation to the ℓ_2 tail estimate of the singular values of A, namely $\sum_{i=k+1}^{n} s_i^2(A)$ (with at least $5/9$ success probability).*

As previously mentioned, in all these cases, the factor of $\Omega(k^2)$ in space is required to obtain the desired guarantee, as implied by the rank lower bound [CW09].

Our algorithms are essentially of the following kind: we sketch matrix A using random projection matrices G, H, of size $k^2\epsilon^{-O(1)} \times n$, to obtain sketch $S = GAH^T$, and then compute the top k singular values/eigenvalues of the resulting matrix S. In the case of residual error estimator, the algorithm just computes the residual error of rank-k approximation of S. Thus, one can view our results as a form of *bi-linear* dimensionality reduction of a matrix, one that preserves some coarse spectral structure. In particular, the classical Johnson-Lindenstauss lemma [JL84], a linear dimensionality reduction, can be seen as preserving the singular value of the $n \times 1$ matrix (a vector) A. Our results show what happens when we apply linear dimensionality reduction to a tensor of order 2.

We note that our algorithms are linear and therefore they work in the most general streaming setting. In particular, the stream elements are updates of the form (i, j, δ_{ij}), interpreted as "increase (i, j) matrix entry by $\delta_{ij} \in \mathbb{R}$". Similarly, the algorithms are also applicable in the case when the matrix is distributed across several machines.

Finally, we remark that our algorithms are randomized and use random seed of length $(\log n/\epsilon)^{O(1)}$ only (to generate the projection matrices G and H). In fact, the only necessary property of matrices G, H is that their spectrum is well-concentrated around 1 (matrices are very well conditioned).

1.2 Related work

Numerical linear algebra in the streaming model. As previously noted, there has been a lot of work on numerical linear algebra in the streaming model to address the need of efficient algorithms on massive datasets. Besides the best rank-k approximation, other

[2] We note that S_1^{k+1} can also be seen as the Schatten norm 1 of the error of best rank-k approximation: $S_1^{k+1} = \|A - A_k\|_{S_1}$ where A_k is the best rank-k approximation of A.

heavily studied problems have been approximate matrix multiplication, ℓ_p regression, and rank approximation, among others (see surveys [KV09], [HMT10], [Mah11] and the references therein).

Most related to ours is the result on rank approximation that shows that, in order to distinguish whether rank is k or more, one has to use space at least $\Omega(k^2)$, and at most $O(k^2 \log n)$ [CW09].

We also note that part of our interest in estimating eigenvalues also stems from an attack on estimating the Schatten norms of a matrix. In particular, Schatten norm 1 of a matrix, also called the nuclear norm, is the sum of the absolute values of the eigenvalues/singular values. On this front, we note that, in independent work, Li and Woodruff obtained lower bounds that are polynomial in n [LW12].

Heavy hitters. Our results can be seen from the prism of computing the heavy hitters of the vector Λ of eigenvalues/singular values. From this perspective, our ℓ_1 and ℓ_2 heavy (eigen-)hitters results are similar in spirit to those obtained by standard ℓ_1 and ℓ_2 heavy hitters of vectors, such as CountSketch [CCF02] and CountMin [CM05] (but not in terms of the algorithms). Namely, we recover well those eigenvalues whose values is at least a fraction $\approx 1/k$ of the total mass (ℓ_1 or ℓ_2) of all the eigenvalues (or, for that matter, of the residual spectral mass). In constrast to the standard vector heavy hitters, our algorithms do not have to recover *which* are the heavy "coordinates". This aspect adds an additional factor of $O(\log n)$ to the space for standard heavy hitters, which we do not incur (ignoring the random seed size).

We note that the standard heavy hitters of vectors have turned out to be quite a useful primitive — see for example the website for CountMin, an ℓ_1 heavy hitters sketch [CM10].

1.3 Techniques

We now briefly overview the technical aspects of our results.

One can try to approach the eigenvalue heavy hitters question by reasoning about a potential rank-k approximation to the input matrix A. For example, one such algorithm for rank approximation generates a sketch by multiplying the matrix A by a projection matrix G, of size $t \times n$, where $t = k(\epsilon^{-1}\log(n))^{O(1)}$ [Sar06]. Then, for some carefully constructed transformation π, [Sar06] proves the following guarantee: $\|A - \pi(GA)\|_F \leq (1+\epsilon)\min_{A_k \text{ of rank } k} \|A - A_k\|_F$, i.e., the residual ℓ_2 norm is well approximated. Intuitively, since GA approximates the top-k eigenspace, it may have some information about specifically the top k eigenvalues of A. Besides the issue that storing GA takes too much space (but can be dealt with applying the projection again), the question is whether one can prove that GA preserves the top spectrum of A specifically. Note that the above guarantee is not sufficient: the kth eigenvalue of A can be much less than $\|A - A_k\|_F$, in which case there is no hope to guarantee reconstruction of the kth eigenvalue. From a larger perspective, the difference is that we want point-wise guarantee for eigenvalues, whereas the above guarantee is "overall".

Our algorithm is nonetheless inspired by the above approach. In particular, our algorithm corresponds to taking two such projections G and H, of size $O(k/\epsilon^2)$ by n, and computing the sketch $S = GAG^T$ (for ℓ_1) or $S = GAH^T$ (for ℓ_2). Our main technical contribution is indeed to prove that this simple algorithm does the job: if we take top k eigenvalues/singular values of the resulting sketch S, we obtain good approximations of the eigenvalues/singular values of the original matrix A.

To prove our main theorems, we rely on several tools from matrix analysis. First, to bound the error incurred by the swarm of small eigenvalues, we use results on the distribution of singular values of random matrices (note that we essentially need to do so, even in the case when A is a partial identity matrix). Second, to obtain point-wise guarantee on eigenvalues/singular values, we deduce some eigenvalue interlacing laws for well-conditioned matrices, similar in spirit to the Cauchy interlacing theorem (which applies to all matrices). Third, to get a precise handle on the residual error of a rank-k approximation, we use Lidskii and dual Lidskii inequalities, which bound the eigenvalues of a sum of Hermitian matrices in terms of the eigenvalues of the component matrices.

We remark that we need to use only a small random seed for the generation of the projection matrices G, H. In particular we show limited independence suffices, following the proof of the concentration of the singular values of these random matrices [BY93].

2 Preliminaries

Consider a real matrix A of size n by n.

Definition 2.1. *For a matrix A, let $s_1(A), \ldots, s_n(A)$ be the singular values of A sorted by decreasing value. Define the p-Schatten norm of A to be $S_p(A) = (\sum_i s_i(A)^p)^{1/p}$. Define the p residual Schatten norm $S_p^j(A)$ as $S_p^j(A) = (\sum_{i>j} s_i(A)^p)^{1/p}$.*

Definition 2.2. *For a symmetric matrix A, let $\lambda_1(A), \ldots, \lambda_n(A)$ be the eigenvalues of A sorted in the decreasing absolute value (but preserving the signs).*

We will use the following inequalities on eigenvalues of Hermitian matrices.

Lemma 2.1. (Weyl inequality) *[Tao12, Section 1.3] Let M_1, M_2 be $t \times t$ Hermitian matrices. Then, for all $1 \leq i, j \leq t$, we have*

$$\lambda_{i+j-1}(M_1 + M_2) \leq \lambda_i(M_1) + \lambda_j(M_2).$$

Lemma 2.2. (Lidskii inequality) *[Tao12, Section 1.3.3] Let M_1, M_2 be $t \times t$ Hermitian matrices. For all $k \in [t]$,*

$$\sum_{j=1}^{t-k} \lambda_{k+j}(M_1 + M_2) \leq \sum_{j=1}^{t-k} \lambda_{k+j}(M_1) + \lambda_j(M_2).$$

Lemma 2.3. (Dual Lidskii inequality) *[Tao12, Section 1.3.3] Let M_1, M_2 be $t \times t$ Hermitian matrices. For all $k \in [t]$,*

$$\sum_{j=1}^{t-k} \lambda_{k+j}(M_1 + M_2) \geq \sum_{j=1}^{t-k} \lambda_{k+j}(M_1) + \lambda_{k+j}(M_2).$$

We will also use the following results on eigenvalues of random matrices. Since we are not aware of such explicit statements in the case of limited independence, we reproduce the proofs for limited independence in Appendix A, based on the proof from [BY93].

Fact 2.1. *For any $n \geq k$, $n \times k$ matrices with random entries whose moments match moments of entries of a matrix with independent $N(0,1)$ entries up to the $\Theta(\log^2 n/\epsilon^2)$th moment have singular values bounded from above by $(2+\epsilon)\sqrt{n}$ with probability $1 - 1/\operatorname{poly}(n)$.*

Fact 2.2. *For $k \leq \epsilon^2 n$ for some constant $\epsilon < 1$, the minimum singular value of a rectangular matrix of size $n \times k$ with entries whose moments match moments of entries of a matrix with independent $N(0,1)$ entries up to the $\Theta(\log^2 n/\epsilon^2)$th moment is bounded from above by $(1+2\epsilon)\sqrt{n}$ and from below by $(1-2\epsilon)\sqrt{n}$ with probability $1 - 1/\operatorname{poly}(n)$.*

Lemma 2.4. *Consider $m \geq n \geq k$. Let G be an $n \times m$ matrix with $\Theta(\log^2 n/\epsilon^2)$-wise independent entries and U be an $m \times k$ matrix satisfying $U^T U = I_k$ where I_k is the identity matrix of dimension k. Then GU satisfies the condition of the above two facts.*

Proof. Let V be an $m \times m$ unitary matrix ($VV^T = I_m$) whose first k columns form U. First notice that if G were fully independent, then the distribution of GV is the same as that of an $n \times m$ matrix H with independent $N(0,1)$ entries. Indeed, the density of the distribution of H at any $X \in \mathbb{R}^{n \times m}$ is $\frac{1}{(2\pi)^{nm/2}} \exp(-\|X\|_F^2/2)$. The density of the distribution of GV at any $X \in \mathbb{R}^{n \times m}$ is $\frac{1}{(2\pi)^{nm/2}} \exp(-\|XV^T\|_F^2/2) = \frac{1}{(2\pi)^{nm/2}} \exp(-\|X\|_F^2/2)$. Therefore, the moments of entries of G and GV are the same i.e. the moments of entries of GV are the same as those of independent $N(0,1)$.

Now consider the case where entries of G are only $\Theta(\log^2 n/\epsilon^2)$-wise independent. Since entries of GV are linear combinations of entries of G, by linearity of expectation, their moments up to the $\Theta(\log^2 n/\epsilon^2)$th moment are the same as those of GV when G is fully independent. Since GU is a submatrix of GV, the lemma follows.

Finally, we note that, our space bounds are in terms of words, that are $\Omega(\log n)$ bits, assuming that the entries of our input matrices also have bounded precision. In such a case, it is sufficient to generate the above bounded-independence matrices G with entries that have $1/n^{O(1)}$ precision.

3 Heavy hitters for eigenvalues and singular values

In this section, we show how to compute ℓ_1 and ℓ_2 heavy hitters of eigenvalues/singular values of real matrices, thereby proving Theorems 1.1 and 1.2. First, we show that we can easily reduce the problem of approximating singular values of general matrices to the problem of approximating eigenvalues of square symmetric matrices.

Lemma 3.1. *Assume that there is an algorithm for approximating large eigenvalues of a symmetric real matrix A, there is an algorithm for approximating large singular values of an arbitrary real matrix B.*

Proof. Consider the following block matrix.

$$A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$$

It is a well-known fact that the eigenvalues of this matrix are the singular values of B and their negations. Indeed, by the singular value decomposition, we can write B as $\sum_{i=1}^n \lambda_i u_i v_i^T$ where $\{u_i\}$ and $\{v_i\}$ are orthonormal sets of vectors and $\lambda_i \geq 0$. Observe that $u_i \circ v_i$ is an eigenvector of A with eigenvalue λ_i and $u_i \circ -v_i$ is an eigenvector of A with eigenvalue $-\lambda_i$.

From now on, we only consider symmetric matrices. Let $A = \sum_{i=1}^n \lambda_i u_i u_i^T = U\Lambda U^T$ where Λ is the diagonal matrix with diagonal entries equal to $\lambda_1, \ldots, \lambda_n$, U is a matrix with orthonormal columns and u_i is the ith column of U.

Before proving our main results, we need to establish some necessary lemmas that are somewhat similar to the Cauchy interlacing theorem.

3.1 Interlacing lemmas We prove two lemmas addressing a form of interlacing theorem. For example, the Cauchy interlacing theorem addresses the case when we consider a minor of a matrix, such as $I_{k,n}AI_{n,k}$ where $I_{k,n}$ is a truncated identity. Below, we address the case when we consider a matrix of the form GDG^T for a diagonal matrix D.

Lemma 3.1. *Let D be a $n \times n$ diagonal matrix. Let G be a $t \times n$ matrix.*

Then, for any $i \leq t$, we have that $\lambda_i(GDG^T) \leq \lambda_1(GD_{\geq i}G^T)$, where $D_{\geq i}$ is D with the largest $i-1$ diagonal values zeroed out.

Proof. Note that

$$\lambda_i(GDG^T) = \max_{S\subset\mathbb{R}^t,\dim(S)=i} \min_{x\in S} \frac{x^TGDG^Tx}{x^Tx}.$$

Let S be the subspace maximizing the above and let $S' = \{y \mid y = G^Tx, x \in S\}$ be the image of S under the linear transformation G^T. S' is a subspace of $\mathbb{R}^n$ of dimension at most i. Suppose the dimension of S' is smaller than i. Then there is some nonzero vector $x \in S$ such that $G^Tx = 0$. Thus, $\lambda_i(GDG^T) \leq 0 \leq \lambda_1(GD_{\geq i}G^T)$ and we obtain the desired conclusion. Henceforth, we assume the dimension of S' is i.

Now we consider the largest eigenvalue of $GD_{\geq i}G^T$:

$$\lambda_1(GD_{\geq i}G^T) = \max_{x\in\mathbb{R}^t} \frac{x^TGD_{\geq i}G^Tx}{x^Tx}.$$

Let P be the subspace spanned by the $n-i+1$ eigenvectors corresponding to the smallest $n-i+1$ eigenvalues of D. Since the dimension of S' is i, there is some non-zero vector $y \in S' \cap P$. By the definition of S', there is some $x \in S$ such that $y = G^Tx$. For this x, we have that $x^TGD_{\geq i}G^Tx = y^TD_{\geq i}y = y^TDy = x^TGDG^Tx$. We conclude that

$$\begin{aligned}\lambda_1(GD_{\geq i}G^T) \geq& \frac{x^TGD_{\geq i}G^Tx}{x^Tx}\\ \geq& \max_{S\subset\mathbb{R}^t,\dim(S)=i} \min_{x\in S} \frac{x^TGDxG^Tx}{x^Tx}\\ =&\lambda_i(GDG^T).\end{aligned}$$

Lemma 3.2. *For $t \geq 1$, fix D to be a diagonal matrix of size $k \times k$ where $k = O(\epsilon^2 t)$. Let G be a $t \times k$ matrix with the k largest singular values bounded between $1-\epsilon$ and $1+\epsilon$ for some $\epsilon \in (0,1)$.*

Then, the spectrum of GDG^T is a $1+O(\epsilon)$ multiplicative approximation of the spectrum of D i.e. when sorted by value, the i-th eigenvalue of GDG^T is a $1+O(\epsilon)$ multiplicative approximation of the i-th eigenvalue of D, for all $1 \leq i \leq k$.

Proof. Fix $i \in [k]$. As above, let $D_{\geq i}$ be D with largest i diagonal values zeroed out. Let j be the number of non-negative eigenvalues of D (i.e., $\lambda_j(D) \geq 0$ and, if $j < k$, also $\lambda_{j+1}(D) < 0$).

We also define $G_{\geq i}$ to be matrix G with rows restricted to the non-zeroed out entries of the diagonal matrix $D_{\geq i}$.

Suppose $i \leq j$. Then, using Lemma 3.1, we deduce that

$$\begin{aligned}\lambda_i(GDG^T) \leq& \lambda_1(GD_{\geq i}G^T)\\ =&\max_{x\in\mathbb{R}^t} x^TGD_{\geq i}G^Tx\\ \leq&\max_{x\in\mathbb{R}^t} x^TG_{\geq i}\cdot\lambda_i(D)\cdot G_{\geq i}^Tx\\ =&\lambda_i(D)\cdot\lambda_1(G_{\geq i}G_{\geq i}^T).\end{aligned}$$

By Fact 2.2, we conclude that $\lambda_i(GDG^T) \leq (1+\epsilon)\lambda_i(D)$.

We now continue with lower-bounding the ith eigenvalue. Let D_i be the matrix D where we zero out all but the biggest i values on the diagonal (i.e., we select the i highest eigenvalues). Let P_i be the subspace of $\mathbb{R}^k$ corresponding these i coordinates. By the condition on G, we have that

$$\max_{S\subset\mathbb{R}^t:\dim S=k} \min_{x\in S} \frac{x^TGG^Tx}{x^Tx} \geq 1-\epsilon.$$

Let S be the set that maximizes the above. There must be an i-dimensional subspace $S' \subset S$ such that, for any $x \in S'$, we have that $Gx \in P_i$. Having constructed the linear subspace S' of dimension i, we now conclude that

$$\begin{aligned}\lambda_i(GDG^T) \geq& \min_{x\in S'} \frac{x^TGDG^Tx}{x^Tx}\\ =&\min_{x\in S'} \frac{x^TGD_iG^Tx}{x^Tx}\\ \geq&\min_{x\in S'} \frac{\lambda_i(D)\cdot x^TGG^Tx}{x^Tx}\\ \geq&(1-\epsilon)\lambda_i(D).\end{aligned}$$

Hence, for any $i \leq j$, we have that $\lambda_i(GDG^T) = (1 \pm \epsilon)\lambda_i(D)$.

To prove the claim for negative eigenvalues, i.e., for $i > j$, consider the matrix $N = -D$. Since have that $\lambda_{k-i+1}(N) = -\lambda_i(D)$, we can apply the same argument as above to the $k-j$ largest eigenvalues of N and obtain that these eigenvalues are a $1 \pm \epsilon$ approximation of the eigenvalues $\{-\lambda_k(D), \ldots -\lambda_{j+1}(D)\}$.

This concludes the proof of Lemma 3.2.

3.2 ℓ_1 heavy hitters We are now ready to prove Theorems 1.1 and 1.2. Our sketch will be of the form $M = GAG^T$ where G is a t by n matrix with $\Theta((\log^2 n)/\epsilon^2)$-wise independent entries identically distributed as $N(0, 1/t)$. We first prove the following lemma on the trace of the sketch matrix.

Lemma 3.3. $\mathbb{E}[\operatorname{tr}(GAG^T)] = \sum_i \lambda_i = \operatorname{tr}(A)$ *and* $\operatorname{Var}[\operatorname{tr}(GAG^T)] = \sum_i \lambda_i^2/t$. *This is true even when the entries of G are 4-wise independent. Hence* $\operatorname{tr}(GAG^T)$ *is a* $1 + 3/\sqrt{t}$ *approximation to* $\operatorname{tr}(A)$ *with probability* 8/9.

Proof. Notice that $\mathbb{E}[tr(GAG^T)]$ and $\operatorname{Var}[tr(GAG^T)]$ only involve terms of degree at most 4 so the values of these quantities are the same when entries of G are 4-wise independent and when they are fully independent. When the entries are fully independent,

$$\mathbb{E}[\operatorname{tr}(GAG^T)] = \sum_{i,j} \lambda_i G_{j,i}^2 = \sum_i \lambda_i$$

$$\begin{aligned}\operatorname{Var}[\operatorname{tr}(GAG^T)] &= \operatorname{Var}[\operatorname{tr}(G\Lambda G^T)]\\ &= \sum_{i,j} \lambda_i^2 \operatorname{Var}[G_{j,i}^2]\\ &= \sum_i \lambda_i^2/t\end{aligned}$$

The final conclusion follows by Chebyshev's inequality:

$$\begin{aligned}\Pr[|\operatorname{tr}(GAG^T) - \operatorname{tr}(A)| > \tfrac{3}{\sqrt{t}} \operatorname{tr}(A)] &\leq \frac{\operatorname{Var}[\operatorname{tr}(GAG^T)]}{(3\operatorname{tr}(A)/\sqrt{t})^2}\\ &\leq \frac{\|\Lambda\|_F^2}{9\operatorname{tr}^2(A)}\\ &\leq \tfrac{1}{9}.\end{aligned}$$

The following lemma proves Theorem 1.1, which follows from applying the lemma for $\phi = \Theta(1/k)$.

Lemma 3.4. (ℓ_1 heavy hitters) *For* $n \geq 1$, *let* A *be a symmetric* $n \times n$ *matrix and* $\lambda_i(A)$ *be the i-th largest eigenvalue of A in absolute value. Fix* $\epsilon, \phi \in (0, 1)$. *Set* $t = \phi^{-1}\epsilon^{-2} + 2\log n$. *With probability at least* 9/10, *the* $\phi^{-1} - 1$ *largest eigenvalues of* GAG^T *in absolute value are* $1+\epsilon$ *multiplicative and* $O(\epsilon^2\phi S_1^{1/\phi}(A) + |\lambda_{1/\phi}|)$ *additive error approximations of the* $\phi^{-1} - 1$ *largest eigenvalues of* A *in absolute value.*

Proof. By the spectral theorem, we can decompose A as $A = U\Lambda U^T$ where Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i(A)$ and $U^TU = UU^T = I$. Let V_i be the set of eigenvalues of A in the range $s_{1/\phi}(A)2^{-i}, s_{1/\phi}(A)2^{-i+1}]$ and Λ_i be a diagonal matrix that contains entries of Λ whose absolute values corresponding to the singular values in V_i and the rest replaced with 0. Let λ_i be the i-th largest eigenvalue of A in absolute value. Let Λ_0 be a diagonal matrix containing entries of Λ whose absolute values are at least $|\lambda_{1/\phi}|$ i.e. $\lambda_1, \ldots, \lambda_{1/\phi}$. Let $G' = GU$ and G'_{V_i} be the same as G' but with the columns corresponding to singular values *not* in V_i replaced with 0. Then $GAG^T = G'(\sum_{i\geq 0} \Lambda_i)G'^T$. We have

$$\begin{aligned}\sum_{i\geq 1} \|G'\Lambda_i G'^T\|_2 \leq &\sum_{i=1}^{2\log n} |\lambda_{1/\phi}|2^{-i+1}\|G'_{V_i}G'^T_{V_i}\|_2 +\\ &\sum_{i\geq 2\log n} |\lambda_{1/\phi}|2^{-i+1}\|G'_{V_i}G'^T_{V_i}\|_2\\ \leq &\sum_{i=1}^{\log n} |\lambda_{1/\phi}|2^{-i+1}\|G'_{V_i}G'^T_{V_i}\|_2 + O(|\lambda_{1/\phi}|)\\ \leq &O(1)\sum_{i\geq 1} \frac{|\lambda_{1/\phi}|2^{-i+1}\max(|V_i|, t)}{t} +\\ &O(|\lambda_{1/\phi}|)\\ \leq &\frac{O(\sum_{i\geq 1/\phi} |\lambda_i|)}{t} + O(|\lambda_{1/\phi}|)\\ \leq &O(|\lambda_{1/\phi}| + \epsilon^2\phi S_1^{1/\phi}(A))\end{aligned}$$

The first inequality comes from the fact that $G'(|\lambda_{\epsilon^2 t}|2^{-i+1}I - \Lambda_i)G'^T$ and $G'(|\lambda_{\epsilon^2 t}|2^{-i+1}I + \Lambda_i)G'^T$ are symmetric psd. The second inequality comes from the fact that there are only n eigenvalues and for any column G_j of G, $\|G_j\|^2$ is bounded by $O(t)$ with probability $1 - 1/\operatorname{poly}(n)$. The third inequality comes from the fact that $\|G'_{V_i}{G'_{V_i}}^T\|_2$ is bounded by $(4 + O(\epsilon))\max(|V_i|, t)$ with probability at least $1 - 1/\operatorname{poly}(t)$ (by Fact 2.1 and Lemma 2.4).

By Lemma 3.2, the eigenvalues of $G'\Lambda_0 G'^T$ are $1 \pm \epsilon$ approximations of eigenvalues of Λ_0 i.e. the j-th largest eigenvalue of $G'\Lambda_0 G'^T$ in absolute value is a $1 \pm \epsilon$ approximation of λ_j. For any arbitrary $j \leq 1/\phi$, applying Weyl's inequality to the symmetric matrices $G'\Lambda_i G'^T$ and $G'\Lambda G'^T$,

$$\begin{aligned}\lambda_j(G'\Lambda G'^T) &\leq \lambda_j(G'\Lambda_0 G'^T) + \sum_{i\geq 1} \|G'\Lambda_i G'^T\|_2\\ &\leq \lambda_j + \epsilon|\lambda_j| + O(|\lambda_{1/\phi}| + \epsilon^2\phi S_1^{1/\phi}(A))\end{aligned}$$

Similarly, $\lambda_j(G'\Lambda G'^T) \geq \lambda_j - \epsilon|\lambda_j| - O(|\lambda_{1/\phi}| + \epsilon^2\phi S_1^{1/\phi}(A))$. This concludes the proof of the lemma.

3.3 ℓ_2 heavy hitters The following lemma proves Theorem 1.2, which follows from applying the lemma for $\phi = \Theta(1/k)$.

Lemma 3.5. (ℓ_2 heavy hitters) *Set $t = O(\phi^{-1}\epsilon^{-2} + \log n)$. Let G and H be t by n matrices with $\Theta((\log^2 n)/\epsilon^2)$-wise independent entries identically distributed as $N(0, 1/t)$. With probability at least 5/9, the top ϕ^{-1} singular values of GAH^T approximate the top ϕ^{-1} singular values of A with as follows for $1 \leq i \leq 1/\phi$:*

$$|s_i^2(GAH^T) - s_i^2(A)| \leq \epsilon s_i^2(A) + O(s_{1/\phi}^2(A) + \epsilon^2 \phi (S_2^{1/\phi}(\Lambda))^2)$$

where $S_2^j(A) = \sqrt{\sum_{i>j} s_i(A)^2}$.

Proof. We prove this lemma by essentially applying the ℓ_1 heavy hitters results twice. By Lemma 3.4, the top ϕ^{-1} eigenvalues of HA^TAH^T, which are the same as those of AH^THA^T, are approximation of the top ϕ^{-1} eigenvalues of A^TA with multiplicative error $1 \pm \epsilon$ and additive error $O(s_{1/\phi}(A^TA) + \epsilon^2\phi S_1^{1/\phi}(A^TA)) = O(s_{1/\phi}^2(A) + \epsilon^2 \phi S_2^{1/\phi}(A))$. By the singular value decomposition, we can decompose $A = U\Lambda V^T$ where Λ is diagonal and $U^TU = V^TV = I_n$. Let $\Lambda = \Lambda_l + \Lambda_s$ where Λ_l and Λ_s are diagonal matrices containing the $1/\phi$ entries with largest absolute values on the diagonal of Λ and the rest, respectively. By Lemma 3.3, $\operatorname{tr}(U\Lambda_s V^T H^T H V \Lambda_s U^T) = \operatorname{tr}(HV\Lambda_s^2 V^T H^T) = (1 \pm \epsilon)\operatorname{tr}(V\Lambda_s^2 V^T) = (1 \pm \epsilon)(S_2^{1/\phi}(A))^2$ with probability at least 8/9. Because $HV\Lambda_l^2 V^T H^T$ has rank at most $1/\phi$, by the Lidskii inequality (Fact 2.2) and the fact that AH^THA^T is psd, we have

$$\begin{aligned}
S_1^{1/\phi}(AH^THA^T) &= S_1^{1/\phi}(HA^TAH^T)\\
&= S_1^{1/\phi}(HV(\Lambda_l + \Lambda_s)^2 V^T H^T)\\
&= S_1^{1/\phi}(HV(\Lambda_l^2 + \Lambda_s^2) V^T H^T)\\
&\leq \operatorname{tr}(HV\Lambda_s^2 V^T H^T)\\
&\leq (1+\epsilon)(S_2^{1/\phi}(A))^2
\end{aligned}$$

By Lemma 3.4, the top ϕ^{-1} eigenvalues of $G(AH^THA^T)G^T$ are approximation of the top ϕ^{-1} eigenvalues of AH^THA^T with multiplicative error $1 \pm \epsilon$ and additive error $O(\epsilon^2\phi S_1^{1/\phi}(AH^THA^T) + s_{1/\phi}(AH^THA^T)) = O(\epsilon^2\phi(S_2^{1/\phi}(A))^2 + s_{1/\phi}^2(A))$ as noted above. The lemma for singular values of GAH^T follows.

4 Estimating Residual Rank-k Error

In this section we prove Theorem 1.3.

Proof. [Proof of Theorem 1.3]

The sketch is GAH^T, where G, H are t by n matrices with $t = 2k\epsilon^{-3}$, distributed $N(0, 1/t)$ with $O((\log n)^2/\epsilon^2)$ independence. The estimate of the residual error is the ℓ_2 norm of the $k+1$ to t singular values of the sketch GAH^T.

Lemma 4.1. *Let A be a $n \times m$ matrix, with $m \leq n$. Let H be a $t \times n$ matrix with $t = \Theta(k\epsilon^{-3} + \log n)$ and entries identically distributed according to $N(0, 1/t)$ with $\Theta(\log n)^2/\epsilon^2)$-independence. Then with probability at least 7/8, $S_2^k(HA)$ is a $1 \pm \epsilon$ approximation of $S_2^k(A)$.*

Proof. By the singular value decomposition, we have $A = U\Lambda V^T$ where $U^TU = V^TV = I_n$ and Λ is a diagonal matrix whose diagonal entries are the singular values of A, in order of non-increasing absolute value. Let $\Lambda = \Lambda_l + \Lambda_s$, where Λ_l is Λ restricted to the first $k + k/\epsilon$ diagonal values (all others zeroed out), and Λ_s contains the rest of the diagonal values.

Now set $M_1 = HU\Lambda_l^2U^TH^T$ and $M_2 = HU\Lambda_s^2U^TH^T$. Note that $HAA^TH^T = HU(\Lambda_s + \Lambda_l)^2U^TH^T = M_1 + M_2$. We will be applying Lidskii and the inverse Lidskii inequalities to M_1 and M_2.

First, by Lemma 3.4 applied to $U\Lambda_l^2U^T$ with $\phi = \epsilon/(\epsilon k + k)$, $HU\Lambda_l^2U^TH^T$, has spectrum that is a $1+\epsilon$ multiplicative approximation to the spectrum of Λ_l^2, namely $\{\lambda_1^2(\Lambda), \ldots \lambda_{k+k/\epsilon}^2(\Lambda))\}$ (note that $HU\Lambda_l^2U^TH^T$ has rank at most $k + k/\epsilon$ hence there are no other non-zero eigenvalues).

In particular, we obtain that $\sum_{j=1}^{t-k} \lambda_{k+j}(M_1)$ is a $1+\epsilon$ approximation to $\sum_{j=k+1}^{k+k/\epsilon} \lambda_j^2(\Lambda)$.

Now we analyze matrix M_2, starting by giving a tight bound for the trace norm. By Lemma 3.3, $\operatorname{tr}(HU\Lambda_s^2U^TH^T)$ is a $1 + \epsilon/3$ approximation to $\operatorname{tr}(U\Lambda_s^2U^T) = \operatorname{tr}(\Lambda_s^2) = \|\Lambda_s\|_F^2$, with probability 8/9.

Finally we want to bound the maximum eigenvalue of M_2. By Lemma 3.4, the maximum eigenvalue of M_2 is at most $(1+\epsilon) \cdot \lambda_{k+k/\epsilon+1}^2(\Lambda) + O(\lambda_{2k/\epsilon}^2(\Lambda) + \epsilon^3/k \sum_{j>k/\epsilon} \lambda_j^2(\Lambda)) \leq O(\frac{\epsilon}{k}\sum_{j>k} \lambda_j^2(\Lambda))$.

We can now complete the lemma. By the Lidskii inequality:

$$\begin{aligned}
(S_2^k(HA))^2 &\leq \sum_{j=1}^{t-k} \lambda_{k+j}(M_1) + \lambda_j(M_2)\\
&\leq (1+\epsilon)\sum_{j=k+1}^{k+k/\epsilon} \lambda_j^2(\Lambda) + (1 + O(\epsilon)) \sum_{j>k+k/\epsilon} \lambda_j^2(\Lambda)\\
&\leq (1 + O(\epsilon))(S_2^k(A))^2
\end{aligned}$$

By the dual Lidskii inequality:

$$\begin{aligned}
&(S_2^k(HA))^2\\
&\geq \sum_{j=1}^{t-k} \lambda_{k+j}(M_1) + \lambda_{k+j}(M_2)\\
&\geq (1-\epsilon)\sum_{j=k+1}^{k+k/\epsilon} \lambda_j^2(\Lambda)
\end{aligned}$$

$$+\left[(1-O(\epsilon))\sum_{j>k+k/\epsilon}\lambda_j^2(\Lambda)-k\cdot O(\frac{\epsilon}{k}(S_2^k(A))^2)\right]$$
$$\geq (1-O(\epsilon))(S_2^k(A))^2.$$

Hence

$$(1-O(\epsilon)))S_2^k(A))^2 \leq (S_2^k(HA))^2 \leq (1+O(\epsilon))(S_2^k(A))^2,$$

and by rescaling ϵ, we obtain the desired conclusion.

Applying Lemma 4.1 twice, we have with probability at least 3/4,

$$\begin{aligned}(S_2^k(GAH^T))^2 &\leq (1+\epsilon)(S_2^k(AH^T))^2\\ &= (1+\epsilon)(S_2^k(HA^T))^2\\ &\leq (1+\epsilon)^2(S_2^k(A^T))^2\\ &= (1+\epsilon)^2(S_2^k(A))^2.\end{aligned}$$

Similarly, $S_2^k(GAH^T) \geq (1+\epsilon)(S_2^k(A))$.

References

[BY93] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[CCF02] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Proceedings of the 29th International Colloquium on Automata Languages and Programming (ICALP)*, pages 693–703, 2002.

[CM05] Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[CM10] Graham Cormode and Muthu Muthukrishnan. Count-min sketch. 2010. `https://sites.google.com/site/countminsketch`.

[CW09] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC*, pages 205–214, 2009.

[FKV04] Alan Frieze, Ravindran Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J.ACM*, 51(6), 2004.

[HMT10] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *arXiv:0909.4061*, 2010. `http://arxiv.org/abs/0909.4061`.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of lipshitz mapping into hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[KV09] Ravindran Kannan and Santosh Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3–4):157–288, 2009.

[LW12] Yi Li and David P. Woodruff. The sketching complexity of matrix rank and Schatten norms. Manuscript, 2012.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–124, 2011.

[Mut05] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[PTRV98] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: a probabilistic analysis. In *PODS'98*, 1998.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science*, 2006.

[Tao12] Terence Tao. *Topics in Random Matrix Theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012.

A Bounds for eigenvalues of random matrices

In this section we provide proofs for Facts 2.1 and 2.2. Such proofs have appeared before; we reproduce them here to track explicitly the independence required for the guarantees to hold.

Consider a random $k \times n$ matrix X where $k \leq n$ and the entries of X have moments up to the $\Theta(\log^2 n/\epsilon^2)$th moment matching those of independent identically distributed random variables with $\mathbb{E}[X_{11}] = 0$ and $\mathbb{E}[X_{11}^2] = 1$. For simplicity, we also assume that for any $i = O(1/\epsilon^2)$, $\mathbb{E}[X_{11}^i] \leq b(i)$ with $b(i)$ not depending on n (for instance, for standard Gaussian random variables, which is the case we want, $\mathbb{E}[X_{11}^i] \leq (i-1)!!$). Let $S = (1/n)XX^T$. First we show that to get the desired facts, one only needs to consider the case $k \geq \epsilon^2 n$.

Lemma A.1. *Consider a symmetric psd matrix A of rank k and a vector v. Let $B = A - vv^T$. Let $\lambda_i(M)$ denote the i-th largest eigenvalue of M. We have $\lambda_i(A) \geq \lambda_i(B) \geq \lambda_{i+1}(A)$ for all $1 \leq i \leq k-1$.*

Proof. For any vector x, we have $x^TBx = x^TAx - (x^Tv)^2 \leq x^TAx$ so $\lambda_i(A) \geq \lambda_i(B)$. Let S be the subspace that is the intersection of the span of the $i+1$ largest eigenvectors of A and the subspace orthogonal to v. The dimension of S is at least i and for any $x \in S$, we have $x^TBx = x^TAx$ so $\lambda_i(B) \geq \lambda_{i+1}(A)$.

Corollary A.1. *Let X be a $k \times n$ matrix, H be the first $k-1$ rows of H and X_k be the kth row of X. If X^TX has k eigenvalues in the range $[a,b]$ then $H^TH = X^TX - X_k^TX_k$ has $k-1$ singular values in the range $[a,b]$.*

From now on, assume $k \geq \epsilon^2 n$. Let $y = k/n$. We follow the proof described in [BY93]. Since the whole proof is long, we only sketch the changes needed. Let T be the same as S but with all the diagonal entries replaced with 0. Define $T(0) = I$, $T(1) = T$, and $T(l) = T_{ab}(l)$ be a $k \times k$ matrix with

$$T_{ab}(l) = n^{-l} \sum X_{av_1} X_{u_1 v_1} X u_1 v_2 X_{u_2 v_2} \cdots X_{u_{l-1} v_l} X_{b v_l}$$

where the summation is over $a \neq u_1, u_1 \neq u_2, \ldots, u_{l-1} \neq b$ and $v_1 \neq v_2, \ldots, v_{l-1} \neq v_l$.

We use w.h.p. as a shorthand for with probability at least $1 - 1/\operatorname{poly}(n)$. W.h.p., all entries of X are bounded by $O(\log n)$. We will use this fact through out the section.

Lemma A.2. *[BY93, Lemma 1] Assume the entries of X are $4ml$-wise independent random variables distributed as $N(0,1)$ with $m = \Theta(\log^2 n)$ and $l = O(1/\epsilon^2)$. W.h.p., we have $\|T(l)\| \le (1+o(1))(2l+1)(l+1)y^{(l-1)/2}$.*

Proof. [Sketch of proof] Since the entries of X are $4ml$-wise independent, $\operatorname{tr}(T^{2m}(l))$ follows the same distribution as when they are fully independent. Let $\delta = n^{-0.49}$. Notice that $ml(2l+1)^2 < \sqrt{k}\delta$, $ml\delta^{1/3} = o(y^{1/12}\log n)$, and $m = \omega(\log n)$ so the rest of the original proof goes through.

The following weakened version of their lemma is enough to obtain a proof in our case.

Lemma A.3. *[BY93, weak version of Lemma 2 (sufficiency direction)] Let $f = O(1/\epsilon^2), \alpha > 0.5, \beta \ge 0$ and $M > 0$ be constants. Let Y be a $Mn^\beta \times n$ matrix with $\Omega(\log n)$-wise independent identically distributed random entries with $Y_{ij} = Z_{ij}^f$ and Z_{ij} follows the same distribution as X_{11}. Let c be a constant that is equal to $\mathbb{E}[Y_{11}]$ if $\alpha \le 1$ (otherwise it can take any value). Then w.h.p.,*

$$\max_{j \le Mn^\beta} \left| n^{-\alpha} \sum_{i=1}^{n} (Y_{ij} - c) \right| = o(1)$$

Proof. We use the following lemma, also from [BY93].

Lemma A.4. *[BY93, Lemma A.1] Suppose $X_1, \ldots, X_n$ are g-wise independent random variables with mean 0 and finite g-moment, where g is a positive even integer. Then for $C(g) = 2^g \cdot g!$,*

$$\mathbb{E}[|\sum_{i=1}^{n} X_i|^g] \le C\left(n\mathbb{E}[X_1^g] + n^{g/2}(\mathbb{E}[X_1^2])^{g/2}\right)$$

Wlog assume $c = 0$. Applying Lemma A.4 with $g > (\beta + 10)/(\alpha - 0.5)$ and Markov inequality, we have

$$\Pr[|n^{-\alpha}\sum_{i=1}^{n} Y_{i1}| > n^{(0.5-\alpha)/2}] \le n^{(-\alpha-0.5)g/2}\mathbb{E}[|\sum_{i=1}^{n} Y_i|^g]$$
$$\le C(\epsilon)n^{(0.5-\alpha)g} \le M^{-1}n^{-\beta-10}$$

Union bound over all $j \le Mn^\beta$ gives us the desired claim with probability $1 - n^{-10}$.

The rest of the original proof goes through by bounding the operator norm of $(T - yI)^t$ with $t = O(1/\epsilon^2)$.

Lemma A.5. *[BY93, Theorem 1] Whp, the singular values of X are bounded from above by $1 + (1+\epsilon)\sqrt{y}$ and from below by $1 - (1+\epsilon)\sqrt{y}$.*

Proof. Consider $t = \Theta(1/\epsilon^2)$. We use the following lemma.

Lemma A.6. *[BY93, Lemma 8] W.h.p. we have*

$$(T - yI)^t = \sum_{r=0}^{t} (-1)^{r+1} T(r) \sum_{i=0}^{[(t-r)/2]} C_i(t,r) y^{t-r-i} + o(1)$$

where $|C_i(t,r)| \le 2^t$.

As shown in the original proof, by Lemma A.3, $\|S - I - T\|_2 = \max_i |n^{-1}\sum_{j=1}^{n}(X_{ij}^2 - 1)| = o(1)$ w.h.p. so we only need to show $\|T - yI\|_2 \le (1+\epsilon)2\sqrt{y}$.

As also shown in the original proof, by Lemma A.2 and A.6, w.h.p., we have $\|T - yI\|_2^t \le Ct^4 2^t y^{t/2}$ so $\|T - yI\|_2 \le (1+\epsilon)2\sqrt{y}$.