# Beating Randomized Response on Incoherent Matrices 

Moritz Hardt* Aaron Roth ${ }^{\dagger}$

October 30, 2018


#### Abstract

Computing accurate low rank approximations of large matrices is a fundamental data mining task. In many applications however the matrix contains sensitive information about individuals. In such case we would like to release a low rank approximation that satisfies a strong privacy guarantee such as differential privacy. Unfortunately, to date the best known algorithm for this task that satisfies differential privacy is based on naive input perturbation or randomized response: Each entry of the matrix is perturbed independently by a sufficiently large random noise variable, a low rank approximation is then computed on the resulting matrix.

We give (the first) significant improvements in accuracy over randomized response under the natural and necessary assumption that the matrix has low coherence. Our algorithm is also very efficient and finds a constant rank approximation of an $m \times n$ matrix in time $O(m n)$. Note that even generating the noise matrix required for randomized response already requires time $O(m n)$.


[^0]
## 1 Introduction

Consider a large $m \times n$ matrix $A$ in which rows correspond to individuals, columns correspond to movies, and the non-zero entry in $A(i, j)$ represent the rating that individual $i$ has given to movie $j$. Such a data set shares two important characteristics with many other data sets:

1. It can be represented as a matrix with very different dimensions. There are many more people than movies, so $n \gg m$
2. It is composed of sensitive information: the rating that an individual gives to a particular movie (and the very fact that he watched said movie) can be possibly compromising information.

Nevertheless, although we want to reveal little about the existence of individual ratings in this data set, it might be extremely useful to be able to allow data analysts to mine such a matrix for statistical information. Even while protecting the privacy of individual entries, it might still be possible to release another matrix that encodes a great deal of information about the original data set. For example, we might hope to be able to recover the cut structure of the corresponding rating graph, perform principal component analysis (PCA), or apply some other data mining technique.

Indeed, this example is not merely theoretical. Data of exactly this form was released by Netflix as part of their competition to design improved recommender systems. Spectral methods such as PCA were commonly used on this dataset, and privacy concerns were acknowledged: Netflix attempted to "anonymize" the dataset in an ad-hoc way. Following this supposedly anonymized release, Naranyanan and Shmatikov [NS08] were able to re-identify many individuals in the dataset by crossreferencing the reviews with publicly available reviews in the internet movie database. As a result of their work, a planned second Netflix challenge was canceled. The story need not have ended this way however - the formal privacy guarantee known as differential privacy could have prevented the attack of [NS08], and indeed, McSherry and Mironov [MM09] demonstrated that many of the recommender systems proposed in the competition could have been implemented in a differentially private way. [MM09] make use of private low-rank matrix approximations using input perturbation methods. In fact, it is not possible to generically improve on input perturbation methods for all matrices without violating blatant non privacy [DN03]. Nevertheless, in this paper, we give the first algorithms for low rank matrix approximation with performance guarantees that are significantly better than input perturbation, under certain commonly satisfied conditions which are already assumed in prior work on non-private low-rank matrix approximation.

In this paper, we consider the problem of privately releasing accurate low-rank approximations to datasets that can be represented as matrices. Such matrix approximations are one of the most fundamental building blocks for statistical analysis and data mining, with key applications including latent semantic indexing and principle component analysis. We provide theorems bounding the accuracy of our approximations as compared to the optimal low rank approximations in the Frobenius norm. The classical Eckart-Young theorem asserts that the optimal rank- $k$ approximation of a matrix $A$ (in either the Frobenius or Spectral norms) is obtained by computing the singular value decomposition $A=U \Sigma V^{T}$, and releasing the truncated $\operatorname{SVD} A_{k}=U \Sigma_{k} V^{T}$, where in $\Sigma_{k}$, all but the top $k$ singular values have been zeroed out. Computing the SVD of a matrix takes time $O\left(\mathrm{mn}^{2}\right)$. In addition to offering privacy guarantees, our algorithm is also extremely efficient: it requires only elementary matrix operations and simple noisy perturbations, and for constant $k$ takes time only $O(m n)$. This represents a happy confluence of the two goals of privacy and efficiency. Normally, the two are at odds, and differentially private algorithms tend to be (much) less efficient than their non-private counterparts. In this
case, however, we will see that some algorithms for fast approximate low-rank matrix approximation are much more amenable to a private implementation than their slower counterparts.

Computing low rank matrix approximations privately has been considered at least since [BDMN05], and to date, no algorithm has improved over simple input perturbation, which achieves an error (when compared with the best rank $k$ approximation $\left.A_{k}\right)$ in Frobenius norm of $\Theta(\sqrt{k(n+m)})$. Although this error is optimal without making any assumptions on the matrix, this error can be prohibitive when the best rank $k$ approximation is actually very good: when $\left\|A-A_{k}\right\|_{F} \ll \sqrt{k(n+m)}$. That is, exactly in the case when a low rank approximation to the matrix would be most useful. We give an algorithm which improves over input perturbation under the conditions that $m \ll n$ and that the coherence of the matrix is small: roughly, that no single row of the matrix is too significantly correlated with any of the right singular vectors of the matrix. Equivalently, no left singular vector has large correlation with one of the standard basis vectors. Low coherence is a commonly studied and satisfied condition. For example, Candes and Tao, motivated by the same Netflix Prize dataset re-identified by [NS08], consider the problem of matrix completion under low coherence conditions [CT10]. They show that matrix completion is possible under low coherence assumptions, and that several reasonable random matrix models exhibit a strong incoherence property. Notably, [CT10] were not concerned with privacy at all: they viewed low coherence as a natural assumption satisfied for datasets resembling the Netflix prize data that could be leveraged to obtain stronger utility guarantees. This represents a second happy confluence of the goals of data privacy and utility: low coherence is an assumption that others already make free of privacy concerns in order to improve the state of the art in data analysis. We show that the same assumption can simultaneously be leveraged for data privacy. In retrospect, low coherence is also an extremely natural condition in the context of privacy, although one that has not previously been considered in the literature. If a matrix fails to have low coherence, then intuitively the data of individual rows of the matrix is encoded closely in individual singular vectors. If it does have low coherence, no small set of singular vectors can be used to encode any row of the matrix with high accuracy, and intuitively, low rank approximations reveal less local information about particular entries of the matrix.

The problem we solve is the following: Given a matrix $A$ and a target rank $k$ we privately compute and release a rank $O(k)$ matrix $B$ such that $\|A-B\|_{F}$ is not much larger than $\left\|A-A_{k}\right\|_{F}$, where $A_{k}$ is the optimal rank $k$ approximation to $A$, and $\|\cdot\|_{F}$ is the Frobenius norm. The quality of the approximation depends on several factors, including $n, m$, the desired rank $k$, and the coherence of the matrix. Our approach improves over input perturbation when the matrix coherence is small.

Our algorithm promises $(\varepsilon, \delta)$-differential privacy [DMNS06] with respect to changes of any single row of magnitude 1 in the $\ell_{2}$-norm. This is only stronger than the standard notion of changing any single entry in the matrix by a unit amount. In the very special case of the matrix representing a (possibly unbalanced) graph, this captures (for example) the addition or removal of a single edge. Therefore in this case our algorithm is promising edge privacy rather than vertex privacy. From a privacy point of view, this is less desirable than vertex privacy, but is still a strong guarantee which is appropriate in many settings. We note that edge privacy is well studied with respect to graph problems (see, e.g. [NRS07, GLM ${ }^{+}$10, GRU11]), and we do not know of any algorithms with non-trivial guarantees on graphs that promise vertex privacy, nor any algorithms in the more general case of matrices that promise privacy with respect to entire rows.

### 1.1 Our results

We start with our first algorithm that improves over randomized response on matrices of small $C$ coherence. We say that an $m \times n$ matrix $A$ has coherence $C$, if no row has Euclidean norm more than $C \cdot\|A\|_{F} / \sqrt{m}$, i.e., more than $C$ times the the typical row norm. This parameter varies between 1 and $\sqrt{m}$, since no row can have Euclidean norm more than $\|A\|_{F}$. Intuitively the condition says that no single row contributes too significantly to the Frobenius norm of the matrix.

Theorem 1.1 (Informal version of Theorem 6.2). There is an ( $\varepsilon, \delta$ )-differentially private algorithm which given a matrix $A \in \mathbb{R}^{m \times n}$ of coherence $C$ such that $n \geqslant m$ computes a rank $2 k$ matrix $B$ such that with probability 9/10,

$$
\|A-B\|_{F} \leqslant O\left(\left\|A-A_{k}\right\|_{F}\right)+O_{\varepsilon, \delta}\left(\sqrt{k m}+\sqrt{k n} \cdot \frac{\sqrt{C k\|A\|_{F}}}{(n m)^{1 / 4}}\right) .
$$

Moreover, the algorithm runs in time $O(\mathrm{kmn})$.
Hidden in the $O_{\varepsilon, \delta}$-notation is a factor of $O(\log (k / \delta) / \varepsilon)$ that depends on the privacy parameters. Usually, $\delta \ll 1 / k$ so that $\log (k / \delta) \leqslant 2 \log (1 / \delta)$. To understand the error bound note that the first term is proportional to the best possible approximation error $\left\|A-A_{k}\right\|_{F}$ of any rank $k$ approximation. In particular, this term is optimal up to constant factors. The second term expresses a more interesting phenomenon. Recall that we assume $n \gg m$ so that $\sqrt{k n}$ would usually dominate $\sqrt{k m}$ except that the the $\sqrt{k n}$ term is multiplied by a factor which can be very small if the matrix has low coherence and is not too dense. For example, when $k=O(1), C=O(1)$ and $\|A\|_{F}=O(\sqrt{n})$, the error is roughly $O\left(\sqrt{m}+\sqrt{n} / m^{1 / 4}\right)$ which can be as small as $O\left(n^{3 / 8}\right)$ depending on the magnitude of $m$. However, already in a much wider range of parameters we observe an error of $o(\sqrt{\mathrm{kn}})$. In fact, in Section C we illustrate why the Netflix data satisfies the assumptions made here and why they are likely to hold in other recommender systems.

When $\|A\|_{F} \geqslant \sqrt{n}$, the previous theorem cannot improve on randomized response by more than a factor of $O\left(\mathrm{~m}^{1 / 4}\right)$. Our next theorem uses a stronger but standard notion of coherence known as $\mu_{0}$-coherence. We defer a formal definition of $\mu_{0}$-coherence to Section 5, but we remark that this parameter varies between 1 and $m$. Using this notion we are able to obtain improvements roughly of order $O(\sqrt{m})$.

Theorem 1.2 (Informal version of Theorem 6.3). There is an ( $\varepsilon, \delta)$-differentially private algorithm which given a matrix $A \in \mathbb{R}^{m \times n}$ with $n \geqslant m$ and of $\mu_{0}$-coherence $\mu$ and rank $r \geqslant 2 k$ computes a rank $2 k$ matrix $B$ such that with probability $9 / 10$,

$$
\|A-B\|_{F} \leqslant O\left(\left\|A-A_{k}\right\|_{F}\right)+O_{\varepsilon, \delta}\left(\sqrt{k m}+\sqrt{k n} \cdot \sqrt{\frac{\mu k r}{m}}\right) .
$$

Moreover, the algorithm runs in time $O(\mathrm{kmn})$.
The hidden factor here is the same as before. Note that when $\mu k r=\operatorname{polylog}(n)$, the theorem can lead to an error bound $\tilde{O}\left(n^{1 / 4}\right)$ depending on the magnitude of $m$. Note that this is roughly the square root of what randomized response would give. But again under much milder assumptions on the coherence, the error remains $o(\sqrt{k n})$. Notably, Candes and Tao [CT10] work with a stronger incoherence assumption than what is needed here. Nevertheless they show that even their stronger assumption is satisfied in a number of reasonable random matrix models. A slight disadvantage of
the error bound in Theorem 1.2 is that the actual rank $r$ of the matrix enters the picture. Theorem 1.2 hence cannot improve over Theorem 1.1 when the matrix has very large rank. We do not know if the dependence on $r$ in the above bound is inherent or rather an artifact of our analysis.

Finally, we remark that while our result depends on the $\mu_{0}$-coherence of the input matrix, our algorithm does not require knowledge or estimation of the $\mu_{0}$-coherence of the input matrix. The only parameters provided to the algorithm are the target rank and the privacy parameters.

Reconstruction attacks and tightness of our results. As it turns out, existing work on "blatant non-privacy" and reconstruction attacks [DN03] demonstrates that our results are essentially tight under the given assumptions. To draw this connection, let us first observe why input perturbation cannot be improved without any assumption on the matrix. To be more precise, by input perturbation we refer to the method which simply perturbs each entry of the matrix with independent Gaussian noise of magnitude $O\left(\varepsilon^{-1} \sqrt{\log (1 / \delta)}\right)$, which is sufficient to achieve $(\varepsilon, \delta)$-differential privacy with respect to unit $\ell_{2}$ perturbations of the entire matrix. To obtain a rank $k$ approximation to the original matrix, one can then simply compute the exactly optimal rank $k$ approximation to the perturbed matrix using the singular value decomposition, which as one can show introduces error $O_{\varepsilon, \delta}(\sqrt{\mathrm{km}}+\sqrt{\mathrm{kn}})$ compared to the optimal rank $k$ approximation to the original matrix in the Frobenius norm. First, let us observe that it is not possible in general to have an algorithm which guarantees error in the Frobenius norm of $o(\sqrt{k n})$ for every matrix $A$, without violating blatant non-privacy ${ }^{1}$, as defined by [DN03]. This is because there is a simple reduction which starts with an $(\varepsilon, \delta)$-differentially private algorithm for computing rank $k$ approximations to matrices $A \in \mathbb{R}^{m \times n}$ and gives an $(\varepsilon, \delta)$-differentially private algorithm which can be used to reconstruct almost every entry in any database $D \in\{0,1\}^{n^{\prime}}$ for $n^{\prime}=k \cdot n$. It is known that $(\varepsilon, \delta)$-private mechanisms do not admit such reconstruction attacks, and so the result is a lower bound. The reduction follows from the fact that we can always encode a bit-valued database $D \in\{0,1\}^{n^{\prime}}$ for $n^{\prime}=k \cdot n$ as $k$ rows of an $m \times n$ matrix for any $m \geqslant k$, simply by zeroing out all additional $m-k$ rows. Note that the resulting matrix only has rank $k$, and so the optimal rank $k$ approximation to this matrix has zero error. If we could recover a matrix $A^{\prime}$ such that $\left\|A-A^{\prime}\right\|_{F}=o(\sqrt{k n})$, this would mean that for a typical nonzero row $A_{i}$ of the matrix with $i \in[k]$, we would have $\left\|A_{i}-A_{i}^{\prime}\right\|_{2}=o(\sqrt{n})$, and $\left\|A_{i}-A_{i}^{\prime}\right\|_{1} \leqslant o(n)$. Then, by simply rounding the entries, we could reconstruct the original database $D$ in almost all of its entries, giving blatant non-privacy as defined by [DN03].

What is happening in the above example? Intuitively, the problem is that in the rank $k$ matrix we construct from $D$, the $k$ nonzero rows of the matrix are encoded accurately by only $k$ right singular vectors. On the other hand, low coherence implies that any $k$ right singular vectors poorly represent a set of only $k$ rows. Hence, there is hope to circumvent the above impediment using a low coherence assumption on the matrix. Indeed, this is precisely what Theorem 1.1 and Theorem 1.2 demonstrate. Nevertheless, reconstruction attacks still lead to lower bounds even under low coherence assumptions. Indeed, using the above ideas, the next proposition shows that Theorem 1.1 is essentially tight up to a factor of $O(\sqrt{k})$. Since in many applications $k=O(1)$, this discrepancy between our upper bound and the lower bound is often insignificant.

Proposition 1.3. Any algorithm which given an $m \times n$ matrix $A$ of coherence $C$ outputs a rank $k$

[^1]matrix B such that with high probability
$$
\|A-B\|_{F} \leqslant o\left(\sqrt{k n} \cdot \frac{\sqrt{C\|A\|_{F}}}{(n m)^{1 / 4}}\right)
$$
cannot satisfy ( $\varepsilon, \delta)$-differential privacy for sufficiently small constants $\varepsilon, \delta$.
Informal proof. For the sake of contradiction, suppose there exists such an algorithm $\mathcal{M}$ that satisfies $(\varepsilon, \delta)$-differential privacy. Then consider a randomized algorithm $\mathcal{M}^{\prime}:\{0,1\}^{n^{\prime}} \rightarrow \mathbb{R}^{n^{\prime}}$ which takes a data set $D \in\{0,1\}^{n^{\prime}}$ containing a sensitive bit for $n^{\prime}=k n$ individuals and encodes it as the $m \times n$ matrix $A_{D}$ which contains $D$ in its first $k$ rows and is 0 everywhere else. $\mathcal{M}^{\prime}(D)$ then computes $\mathcal{M}\left(A_{D}\right)$ and outputs the projection of $\mathcal{M}\left(A_{D}\right)$ onto the first $k$ rows (thought of as a vector of length $n^{\prime}=k n$ ).

We claim that $\mathcal{M}^{\prime}$ is $(\varepsilon, \delta)$-differentially privacy. This is because the map from $D$ to $A_{D}$ is sensitivity preserving and the post-processing computed on $M\left(A_{D}\right)$ preserves $(\varepsilon, \delta)$-differential privacy of $M$.

On the other hand, we claim that $\mathcal{M}^{\prime}$ is blatantly non-private. To see this note that the matrix $A_{D}$ has coherence $C=\sqrt{m / k}$ and $\|A\|_{F} \leqslant \sqrt{k n}$ so that one can check that $\|A-\mathcal{M}(A)\|_{F} \leqslant o(\sqrt{k n})$ with high probability. This implies that $\left\|D-\mathcal{M}^{\prime}(D)\right\|_{2} \leqslant o\left(\sqrt{n^{\prime}}\right)$ with high probability. We therefore also have $\left\|D-\mathcal{M}^{\prime}(D)\right\|_{1} \leqslant o\left(n^{\prime}\right)$. But in this case we can compute a data set $D^{\prime}$ from the output of $M^{\prime}(D)$ such that $\left\|D-D^{\prime}\right\|_{0}=o\left(n^{\prime}\right)$ by rounding. This is the definition of a reconstruction attack showing that $\mathcal{M}^{\prime}$ is blatantly non-private. Since $(\varepsilon, \delta)$-differential privacy is known to prevent blatant non-privacy ${ }^{2}$ for sufficiently small $\varepsilon, \delta>0$, this presents the contradiction we sought.

A similar proof shows that error $o(\sqrt{n} \cdot \sqrt{\mu / m})$ (where $\mu$ is the $\mu_{0}$-coherence of the matrix) cannot be achieved with $(\varepsilon, \delta)$-differential privacy. This shows that also Theorem 1.2 is tight up to the exact dependence on $k$ and $r$. We leave it as an intriguing open problem to determine the exact interplay between coherence and the other parameters.

### 1.2 Techniques and proof overview

Our algorithm is based on a random-projection algorithm of Halko, Martinsson and Tropp [HMT11], which involves two steps: range finding and projection. The range finding algorithm first computes $k$ Gaussian measurements of $A$, which we denote by $Y=A \Omega$. Here, $A$ is $m \times n$ and $\Omega$ is $n \times k$. These measurements can be thought of as a random projection of the matrix into a lower dimensional representation, i.e., $Y$ is $m \times k$. The crux of the analysis in [HMT11] is in arguing that $Y$ already captures most of the range of $A$. Hence, all that remains to be done is to compute the orthonormal projection operator $P_{Y}$ into the span of $Y$, and to compute the projection $P_{Y} A$. Note that $P_{Y} A$ is now a $k$-dimensional approximation of $A$ and since $Y$ closely approximated the range of $A$, it must be a good approximation, say, in the Frobenius norm.

The motivation of [HMT11] was to obtain a fast low rank approximation algorithm. Indeed, [HMT11] give a detailed theoretical analysis and empirical evaluation of the algorithm's performance.

Step 1: Privacy preserving range finder and projection. We will leverage the algorithm of [HMT11] to obtain improved accuracy bounds in the privacy setting. As a first step, we need to be able to carry out the range finding and projection step in a privacy preserving manner. Our analysis proceeds by observing that the projection of $A$ to $Y$ approximately preserves all of the $\ell_{2}$ row-norms of $A$, and so

[^2]we can apply a Gaussian perturbation to $Y$, rather than to $A$. (An $m \times k$ standard Gaussian matrix has Frobenius norm $O(\sqrt{\mathrm{~km}})$, which is now independent of $n$ ). The formal presentation of this part of the argument appears in Section 4. This step provides an approximation to the range of $A$ which might already be useful for some applications, but has not yet achieved our goal of computing a low rank approximation to $A$ itself. For this, we need the projection step discussed next.

Step 2: Controlling the projection matrix using low coherence. We then show that under our low-coherence assumption on $A$, the entries of the projection matrix into the range of $Y, P_{Y}$, must be small in magnitude. Finally, when $P_{Y}$ has small entries, the final projection step, of computing $P_{Y} A$ has low sensitivity, and although we must now again add a Gaussian perturbation of dimension $m \times n$, the magnitude of the perturbation in each entry can be smaller than would have been necessary under naive input perturbation.

In order to obtain bounds on the $\ell_{\infty}$-norm of the projection operator we make crucial use of the low-coherence assumption. Here we describe the proof strategy that leads to Theorem 1.2. Theorem 1.1 is somewhat easier to show and follows along similar lines. The first observation is that the Gaussian measurements taken by the range finding algorithm are mostly linear combinations of the top left singular vectors of the matrix. But when the matrix $A$ has low coherence, then its top left singular vectors must have very small correlation with the standard basis. This means that the top singular vectors must have small coordinates. As a result each of the Gaussian measurements we take must have small $\ell_{\infty}$-norm relative to the magnitude of the measurement. Some complications arise as we must add noise to the matrix $Y$ for privacy reasons and then orthonormalize it using the Gram-Schmidt orthonormalization algorithm. A key observation is that the noise matrix is generated independently of $Y$. As a result, it must be the case that all columns of the noise matrix have very small inner product with the columns of $Y$. A careful technical argument uses this observation in order to show that the effect of noise can be controlled throughout the Gram-Schmidt orthonormalization. The result is a projection matrix in which the magnitude of each entry is small whenever the coherence of $A$ was small to begin with.

The exact proof strategy depends on the notion of coherence that we work with. Both notions we consider in this paper are presented and analyzed in Section 5. We then also show that small $\mu_{0}$-coherence is indeed a stronger assumption than small $C$-coherence.

### 1.3 Related Work

### 1.3.1 Differential Privacy

We use as our privacy solution concept the by now standard notion of differential privacy, developed in a series of papers [BDMN05, $\mathrm{CDM}^{+} 05$, DMNS06], and first defined by Dwork, McSherry, Nissim, and Smith [DMNS06]. The problem of privately computing low-rank approximations to matrix valued data was one of the first problems studied in the differential privacy literature, first considered by Blum, Dwork, McSherry, and Nissim [BDMN05], who give an input perturbation based algorithm for computing the singular value decomposition by directly computing the eigenvector decomposition of a perturbed covariance matrix. Computing low rank approximations is an extremely useful primitive for differentially private algorithms, and indeed, McSherry and Mironov [MM09] used the algorithm given in [BDMN05] in order to implement and evaluate differentially private versions of recommendation algorithms from the Netflix prize competition.

Finding differentially private low-rank approximation algorithms with superior theoretical performance guarantees to input perturbation methods has remained an open problem. Beating input perturbation methods for arbitrary symmetric matrices was recently explicitly proposed as an open problem in [GRU11], who showed that such algorithms would lead to the first efficient algorithm for privately releasing synthetic data useful for graph cuts which improves over simple randomized response. Our work does not resolve this open question because our results only improve over input perturbation methods for matrices with unbalanced dimensions which satisfy a low-coherence assumption, but is the first algorithm to improve over [BDMN05] under any condition.

Comparison to recent results of Kapralov, McSherry and Talwar. In a recent independent and simultaneous work, Kapralov, McSherry, and Talwar [KMT11] give a new polynomial-time algorithm for computing privacy-preserving rank 1 approximations to symmetric, positive-semidefinite matrices. Their algorithm achieves $(\varepsilon, 0)$-differential privacy under unit spectral norm perturbations to the matrix. Their algorithm outputs a vector $v$ such that for all $\alpha>0, \mathbb{E}\left[v^{T} A v\right] \geqslant(1-\alpha)\|A\|-$ $O(n \log (1 / \alpha) /(\varepsilon \alpha))$ (where $\|\cdot\|$ denotes the spectral norm) and they show that this is nearly tight for $(\varepsilon, 0)$-differential privacy guarantees. Our results are therefore strictly incomparable. In this work, the goal is to achieve error $o(\sqrt{k n})$ (i.e. $o(\sqrt{n})$ for rank-1 approximations) assuming low coherence, (a stronger error bound) under ( $\varepsilon, \delta$ )-differential privacy (a weaker privacy guarantee) and without making any assumptions about symmetry or positive-semidefiniteness.

### 1.3.2 Fast Computation of Low Rank Matrix Approximations

There is also an extensive literature on randomized algorithms for computing approximately optimal low rank matrix approximations, motivated by improving the running time of the exact singular value decompositions. This literature originated with the work of Papadimitriou et al [PTRV98] and Frieze, Kannan, and Vempala [FKV04], who gave algorithms based on random projections and column sampling (in both cases with the goal of decreasing the dimension of the matrix). Achlioptas and McSherry [AM01] give fast algorithms for computing low rank approximations based on randomly perturbing the original matrix (which can be done to induce sparsity). Although [AM01] pre-dated the privacy literature, some of the algorithms presented in it can be viewed as privacy preserving, because perturbing the actual matrix with appropriately scaled Gaussian noise is a privacy preserving operation sometimes referred to as randomized response. When appropriately scaled (for privacy) Gaussian noise is added to an $m \times n$ matrix, it results in an algorithm for approximating the best rank $k$ approximation up to an additive error of $O(\sqrt{k(m+n)})$ in the Frobenius norm.

Our algorithms are most closely related to the very recent work of Halko, Martinsson, and Tropp [HMT11], who give fast algorithms for computing low rank approximations based on two steps: range finding, and projection. As already discussed, in the first step, these algorithms project the matrix $A$ into an $m \times k$ matrix $Y$ which approximately captures the range of $A$. Then $A$ is projected into the range of $Y$, which yields a rank $k$ matrix which gives a good approximation to $A$ if a good rank- $k$ approximation exists. We will further discuss the algorithm of [HMT11] and our modifications in the course of the paper.

### 1.3.3 Low Coherence Conditions

Low coherence conditions have been recently studied in a number of papers for a number of matrix problems, and is a commonly satisfied condition on matrices. Recently, Candes and Recht [CR09]
and Candes and Tao [CT10] considered the problem of matrix completion. Matrix completion is the problem of recovering all entries of a matrix from which only a subset of the entries which have been randomly sampled. This problem is inspired by the Netflix prize recommendation problem, in which a matrix is given, with individuals on the rows, movies on the columns, and in which the matrix entries correspond to individual movie ratings. The matrix provides only a small number of movie ratings per individual, and the challenge is to predict the missing entries in the matrix. Clearly accurate matrix completion is impossible for arbitrary matrices, but [CR09, CT10] show the remarkable result that it is possible under low coherence assumptions. Candes and Tao [CT10] also show that almost every matrix satisfies a low coherence condition, in the sense that randomly generated matrices will be low coherence with extremely high probability.

Talwalkar and Rostamizadeh recently used low-coherence assumptions for the problem of (nonprivate) low-rank matrix approximation [TR 10]. A common heuristic for speeding the computation of low-rank matrix approximations is to compute on only a small randomly chosen subset of the columns, rather than on the entire matrix. [TR10] showed that under low-coherence assumptions similar to those of [CR09, CT10], the spectrum of a matrix is in fact well approximated by a small number of randomly sampled columns, and give formal guarantees on the approximation quality of the sampling based Nyström method of low-rank matrix approximation.

## Acknowledgments

We thank Boaz Barak, Anupam Gupta, and Jon Ullman for many helpful discussions. We thank Sham Kakade for pointing us to the paper of [HMT11]. We thank Michael Kapralov, Frank McSherry and Kunal Talwar for communicating their result to us, and for useful conversations about low rank matrix approximations. We wish to thank Microsoft Research New England where part of this research was done.

## 2 Preliminaries

We view our dataset as a real valued matrix $A \in \mathbb{R}^{m \times n}$. We sometimes denote the $i$-th of a matrix by $A_{(i)}$. Let

$$
\begin{equation*}
\mathcal{N}=\left\{P \in \mathbb{R}^{m \times n} \text { : there exists an index } i \in[m] \text { such that }\left\|P_{(i)}\right\|_{2} \leqslant 1 \text { and }\left\|P_{(j)}\right\|_{2}=0 \text { for all } j \neq i\right\} \tag{1}
\end{equation*}
$$

denote the set of matrices that take 0 at all values, except possibly in a single row, which has Euclidean norm at most 1 .

Definition 2.1. We say that two matrices $A, A^{\prime} \in \mathbb{R}^{m \times n}$ are neighboring if $\left(A-A^{\prime}\right) \in \mathcal{N}$.
We use the by now standard privacy solution concept of differential privacy:
Definition 2.2. An algorithm $M: \mathbb{R}^{m \times n} \rightarrow R$ (where $R$ is some arbitrary abstract range) is ( $\varepsilon, \delta$ )differentially private if for all pairs of neighboring databases $A, A^{\prime} \in \mathbb{R}^{m \times n}$, and for all subsets of the range $S \subseteq R$ :

$$
\operatorname{Pr}\{M(A) \in S\} \leqslant \exp (\varepsilon) \operatorname{Pr}\left\{M\left(A^{\prime}\right) \in S\right\}+\delta
$$

We make use of the following useful facts about differential privacy.

Fact 2.3. If $M: \mathbb{R}^{m \times n} \rightarrow R$ is $(\varepsilon, \delta)$-differentially private, and $M^{\prime}: R \rightarrow R^{\prime}$ is an arbitrary randomized algorithm mapping $R$ to $R^{\prime}$, then $M^{\prime}(M(\cdot)): \mathbb{R}^{m \times n} \rightarrow R^{\prime}$ is $(\varepsilon, \delta)$-differentially private.

The following useful theorem of Dwork, Rothblum, and Vadhan tells us how differential privacy guarantees compose.

Theorem 2.4 (Composition [DRV10]). Let $\varepsilon, \delta \in(0,1), \delta^{\prime}>0$. If $M_{1}, \ldots, M_{k}$ are each $(\varepsilon, \delta)$ differentially private algorithms, then the algorithm $M(A) \equiv\left(M_{1}(A), \ldots, M_{k}(A)\right)$ releasing the concatenation of the results of each algorithm is $(k \varepsilon, k \delta)$-differentially private. It is also $\left(\varepsilon^{\prime}, k \delta+\delta^{\prime}\right)$ differentially private for:

$$
\varepsilon^{\prime}<\sqrt{2 k \ln \left(1 / \delta^{\prime}\right)} \varepsilon+2 k \varepsilon^{2}
$$

We denote the 1 -dimensional Gaussian distribution of mean $\mu$ and variance $\sigma^{2}$ by $N\left(\mu, \sigma^{2}\right)$. We use $N\left(\mu, \sigma^{2}\right)^{d}$ to denote the distribution over $d$-dimensional vectors with i.i.d. coordinates sampled from $N\left(\mu, \sigma^{2}\right)$. We write $X \sim D$ to indicate that a variable $X$ is distributed according to a distribution $D$. We note the following useful fact about the Gaussian distribution.
Fact 2.5. If $g_{i} \sim N\left(\mu_{i}, \sigma_{i}^{2}\right)$, then $\sum g_{i} \sim N\left(\sum_{i} \mu_{i}, \sum_{i} \sigma_{i}^{2}\right)$.
The following theorem is well known folklore. We include a proof in the appendix for completeness.

Theorem 2.6 (Gaussian Mechanism). Let $x, y \in \mathbb{R}^{d}$ be any two vectors such that $\|x-y\|_{2} \leqslant c$. Let $Y \in \mathbb{R}^{d}$ be an independent random draw from $N\left(0, \rho^{2}\right)^{d}$, where $\rho=c \varepsilon^{-1} \sqrt{\log (1.25 / \delta)}$. Then for any $S \subseteq \mathbb{R}^{d}:$

$$
\operatorname{Pr}\{x+Y \in S\} \leqslant \exp (\varepsilon) \operatorname{Pr}[y+Y \in S]+\delta
$$

Vector and matrix norms. We denote by $\|\cdot\|_{p}$ the $\ell_{p}$-norm of a vector and sometimes use $\|\cdot\|$ as a shorthand for the Euclidean norm. Given a real $m \times n$ matrix $A$, we will work with the spectral norm $\|A\|_{2}$ and the Frobenius norm $\|A\|_{F}$ defined as

$$
\begin{equation*}
\|A\|_{2}=\max _{\|x\|=1}\|A x\| \quad \text { and } \quad\|A\|_{F}=\sqrt{\sum_{i, j} a_{i j}^{2}} \tag{2}
\end{equation*}
$$

For any $m \times n$ matrix $A$ of rank $r$ we have $\|A\|_{2} \leqslant\|A\|_{F} \leqslant \sqrt{r} \cdot\|A\|_{2}$. For a matrix $Y$ we denote by $P_{Y}$ the orthonormal projection operator onto the range of $Y$.
Fact 2.7. $P_{Y}=Y\left(Y^{*} Y\right) Y^{-1}$
Fact 2.8 (Submultiplicativity). For any $m \times n$ matrix $A$ and $n \times r$ matrix $B$ we have $\|A B\|_{F} \leqslant\|A\|_{F} \cdot\|B\|_{F}$.
Theorem 2.9 (Weyl). For any $m \times n$ matrices $A$, $E$, we have $\left|\sigma_{i}(A+E)-\sigma_{i}(A)\right| \leqslant\|E\|_{2}$, where $\sigma_{i}(M)$ denotes the $i$-th singular value of a matrix $M$. where $\sigma_{i}(M)$ denotes the $i$-th singular value of a matrix $M$.

## 3 Low-rank approximation via Gaussian measurements

We will begin by presenting an algorithm of Halko, Martinsson and Tropp [HMT11] as described in Figure 1. The algorithm produces a rank $r+p$ approximation that already for $p \geqslant 2$ closely matches the best rank $r$ approximaton of the matrix in Frobenius norm. The guarantees of the algorithm are detailed in Theorem 3.1.

Input: Matrix $A \in \mathbb{R}^{m \times n}$, target rank $r$, oversampling parameter $p$.

1. Range finder: Let $\Omega$ be an $n \times k$ standard Gaussian matrix where $k=p+r$. Compute the $n \times k$ measurement matrix $Y=A \Omega$. Compute the orthonormal projection operator $P_{Y}$.
2. Projection: Compute the projection $B=P_{Y} A$.

Output: Matrix $B$ of rank $k$.
Figure 1: Base algorithm for computing a low-rank approximation

Theorem 3.1 ([HMT11]). Suppose that $A$ is a real $m \times n$ matrix with singular values $\sigma_{1} \geqslant \sigma_{2} \geqslant \ldots$. Choose a target rank $r \geqslant 2$ and an oversampling parameter $p \geqslant 2$ where $r+p \leqslant \min \{m, n\}$. Draw an $n \times(r+p)$ standard Gaussian matrix $\Omega$, construct the sample matrix $Y=A \Omega$ and let $B=P_{Y} A$. Then the expected approximation error in Frobenius norm satisfies

$$
\begin{equation*}
\mathbb{E}\|A-B\|_{F} \leqslant\left(1+\frac{r}{p-1}\right)^{1 / 2} \sqrt{\sum_{j>r} \sigma_{j}^{2}} . \tag{3}
\end{equation*}
$$

In particular, for $p=r+1$ we have

$$
\begin{equation*}
\mathbb{E}\|A-B\|_{F} \leqslant \sqrt{2} \cdot\left\|A-A_{r}\right\|_{F} \tag{4}
\end{equation*}
$$

When applying the the theorem we will use Markov's inequality to argue that the error bounds hold with sufficiently high probability up to a constant factor loss. As shown in [HMT11], much better bounds on the failure probability are possible. We will omit the precise bounds here for the sake of simplicity.

## 4 Privacy-preserving sub-routines: Range finder and projection

In order to give a privacy preserving variant of the above algorithm, we will first need to carefully bound the sensitivity of the range finder and of the projection step, and bound the effect of the necessary perturbations. We do this for each step in this section.

### 4.1 Privacy-preserving range finder

In this section we present a privacy-preserving algorithm which finds a set of vectors $Y$ whose span contains most of the spectrum of a given matrix $A$.

Lemma 4.1. The algorithm in Figure 2 satisfies ( $\varepsilon, \delta)$-differential privacy.
Proof. We argue that outputting $\tilde{Y}$ preserves $(\varepsilon, \delta)$-differential privacy. That outputting $W$ preserves $(\varepsilon, \delta)$-differential privacy follows from the fact that differential privacy holds under arbitrary postprocessing.

Consider any two neighboring matrices $A, A^{\prime} \in \mathbb{R}^{m \times n}$ differing in their $i^{\prime}$ th row, and let $Y=A \Omega$ and $Y^{\prime}=A^{\prime} \Omega$. Define $e \in \mathbb{R}^{n}$ to be $e^{T}=A_{(i)}-A_{(i)}^{\prime}$. Note that by the definition of neighboring, we must have $\|e\|_{2} \leqslant 1$. Observe that for each row $j \neq i$, we have $Y_{(j)}=Y_{(j)}^{\prime}$, and define $\widehat{e} \in \mathbb{R}^{k}$ to be

Input: Matrix $A \in \mathbb{R}^{m \times n}$, target rank $r$, oversampling parameter $p$ such that $r+p \leqslant \min \{m, n\}$, privacy parameters $\varepsilon, \delta \in(0,1)$.

1. Let $\Omega$ be an $n \times k$ standard Gaussian matrix where $k=p+r$.
2. Compute the $n \times k$ measurement matrix $Y=A \Omega$.
3. Let $N \sim N\left(0, \rho^{2}\right)^{m \times k}$ where $\rho=2 \varepsilon^{-1} \sqrt{2 k \log (4 k / \delta)}$.
4. Let $\tilde{Y}=Y+N$.
5. Orthonormalize the columns of $\tilde{Y}$ and let the result be $W$.

Output: Orthonormal $m \times k$ matrix $W$.
Figure 2: Privacy-preserving range finder
$\widehat{e}=Y_{(i)}-Y_{(i)}^{\prime}=e^{T} \Omega$. First, we will give a high-probability bound on $\|e\|_{2}$. Observe that for each $j \in[k], \widehat{e}_{j}$ is distributed like a standard Gaussian:

$$
\widehat{e}_{j} \sim \sum_{\ell=1}^{n} e_{\ell} \cdot N(0,1)=N\left(0, \sum_{\ell=1}^{n} e_{\ell}^{2}\right)=N(0,1),
$$

where we used Fact 2.5. Therefore, we have for any $t \geqslant 1$, by standard Gaussian tail bounds,

$$
\operatorname{Pr}\left\{\left|\widehat{e}_{j}\right| \geqslant t\right\} \leqslant 2 \exp \left(-\frac{t^{2}}{2}\right)
$$

Taking a union bound over all $k$ coordinates we have:

$$
\operatorname{Pr}\left\{\max _{j \in[k]}\left|\widehat{e}_{j}\right| \geqslant \sqrt{2 \log (4 k / \delta)}\right\} \leqslant \frac{\delta}{2}
$$

In particular, we have except with probability $\delta / 2$,

$$
\begin{equation*}
\|e\|_{2} \leqslant \sqrt{2 k \log (4 k / \delta)} \tag{5}
\end{equation*}
$$

Note that we have set $\rho$ such that conditioned on Equation 5 (which holds with probability at least 1 $\delta / 2)$ we have the following by Theorem 2.6: For every set $S \subseteq \mathbb{R}^{m \times k}, \operatorname{Pr}\{\tilde{Y} \in S\} \leqslant \exp (\varepsilon) \operatorname{Pr}\left\{\tilde{Y}^{\prime} \in S\right\}+$ $\delta / 2$. Hence, without any conditioning we can say:

$$
\operatorname{Pr}\{\tilde{Y} \in S\} \leqslant \exp (\varepsilon) \operatorname{Pr}\left\{\tilde{Y}^{\prime} \in S\right\}+\delta
$$

which completes the proof of privacy.
Theorem 4.2. Let $A$ be an $m \times n$ matrix with singular values $\sigma_{1} \geqslant \sigma_{2} \geqslant \ldots$. Then, given $A$ and valid parameters $r, p, \varepsilon, \delta$, the algorithm in Figure 2 returns a matrix $W$ such that $W$ satisfies $(\varepsilon, \delta)$ differential privacy, and moreover we have the error bound

$$
\begin{equation*}
\mathbb{E}\left\|A-W W^{T} A\right\|_{F} \leqslant\left(1+\frac{r}{p-1}\right)^{1 / 2} \sqrt{\sum_{j>r}\left(\sigma_{j}+\rho\right)^{2}} . \tag{6}
\end{equation*}
$$

Proof. Privacy follows from Lemma 4.1. Let us therefore argue the second part of the theorem. Consider the $m \times(n+m)$ matrix

$$
A^{\prime}=\left[A \mid \rho I_{m \times m}\right]
$$

where $I_{m \times m}$ is the $m \times m$ identity matrix. Let $\Omega^{\prime}$ denote a random $(m+n) \times k$ Gaussian matrix. Note that

$$
\tilde{Y} \sim A^{\prime} \Omega^{\prime}
$$

That is, $\tilde{Y}$ is distributed the same way as $Y^{\prime}=A^{\prime} \Omega^{\prime}$. Here, we're using the fact that $\rho N(0,1)=$ $N\left(0, \rho^{2}\right)$.

On the other hand, by Theorem 3.1, we know that $Y^{\prime}$ is a good range for $A^{\prime}$ in the sense that

$$
\begin{equation*}
\mathbb{E}\left\|A^{\prime}-P_{Y^{\prime}} A^{\prime}\right\|_{F} \leqslant\left(1+\frac{r}{p-1}\right)^{1 / 2} \sqrt{\sum_{j>r} \sigma_{j}^{\prime 2}} \tag{7}
\end{equation*}
$$

Here, $\sigma_{j}^{\prime}$ denotes the $j$-th largest singular value of $A^{\prime}$.
Claim 4.3. $\left\|A-P_{Y^{\prime}} A\right\|_{F} \leqslant\left\|A^{\prime}-P_{Y^{\prime}} A^{\prime}\right\|_{F}$
Proof. The claim is immediate, because we can obtain $A$ from $A^{\prime}$ by truncating the last $m$ columns. Hence, the approximation error can only decrease.

Claim 4.4. For all $j$, we have $\left|\sigma_{j}-\sigma_{j}^{\prime}\right| \leqslant \rho$.
Proof. Consider the matrix $A_{0}=\left[A \mid 0_{m \times m}\right]$ where $0_{m \times m}$ is the all zeros matrix. Note that

$$
A^{\prime}=A_{0}+E \quad \text { with } \quad E=\left[0_{m \times n} \mid \rho I_{m \times m}\right] .
$$

Also, $\sigma_{j}=\sigma_{j}\left(A_{0}\right)$, since we just appended an all zeros matrix. On the other hand,

$$
\|E\|_{2}=\left\|\rho I_{m \times m}\right\|_{2}=\rho
$$

Hence, by Weyl's perturbation bound (Theorem 2.9)

$$
\left|\sigma_{j}-\sigma_{j}^{\prime}\right|=\left|\sigma_{j}\left(A_{0}\right)-\sigma_{j}^{\prime}\right| \leqslant\|E\|_{2}=\rho
$$

Combining the previous claims with (7), we have

$$
\mathbb{E}\left\|A-P_{Y^{\prime}} A\right\|_{F} \leqslant \mathbb{E}\left\|A^{\prime}-P_{Y^{\prime}} A^{\prime}\right\|_{F} \leqslant\left(1+\frac{r}{p-1}\right)^{1 / 2} \sqrt{\sum_{j>r}\left(\sigma_{j}+\rho\right)^{2}}
$$

Since $Y^{\prime}$ and $\tilde{Y}$ are identically distributed, the same claim is true when replacing $P_{Y^{\prime}}$ by $P_{\tilde{Y}}$. Furthermore, $P_{\tilde{Y}}=W W^{T}$ and so the claim follows.

Corollary 4.5. Let $A \in \mathbb{R}^{m \times n}$ be as in the previous theorem. Assume that $m \leqslant n$ and run the algorithm with $p \geqslant r+1$. Then, with probability 99/100,

$$
\left\|A-W W^{T} A\right\|_{F} \leqslant O\left(\sqrt{\sum_{j>r} \sigma_{j}^{2}}+\sqrt{\rho^{2} m}\right)
$$

Proof. By Markov's inequality and the previous theorem, we have with probability 99/100,

$$
\left\|A-P_{W} A\right\|_{F} \leqslant O\left(\sqrt{\sum_{j>r}\left(\sigma_{j}+\rho\right)^{2}}\right) .
$$

But note that $\left(\sigma_{j}+\rho\right)^{2}=\sigma_{j}^{2}+2 \sigma_{j} \rho+\rho^{2} \leqslant 3 \sigma_{j}^{2}+3 \rho^{2}$. This is because either $\sigma_{j}>\rho \geqslant 1$ and thus $\sigma_{j}^{2} \geqslant \sigma_{j} \rho$ or else $\rho \geqslant \sigma$ in which case $\rho^{2} \geqslant \sigma_{j} \rho$. The claim follows by using that $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$ for non-negative $a, b \geqslant 0$.

### 4.2 Privacy-preserving projections

In the previous section we showed a privacy-preserving algorithm that finds a small number of orthonormal vectors $W$ such that $A$ is well-approximated by $W W^{T} A$. To obtain a privacy preserving low-rank approximation algorithm we still need to show how to carry out the projection step in a privacy-preserving fashion. We analyze the error of the projection step in terms of the magnitude of the maximum entry of each column of $W$. This serves to bound the sensitivity of the matrix multiplication operation. The smaller the entries of $W$, the smaller the over all error that we incur.

Input: Matrix $A \in \mathbb{R}^{m \times n}$, matrix $W \in \mathbb{R}^{m \times k}$ whose columns have norm at most 1 , privacy parameters $\varepsilon, \delta \in(0,1)$.

1. Let $W=\left[w_{1}\left|w_{2}\right| \cdots \mid w_{k}\right]$ and for each $i \in[k]$ let $\alpha_{i}=\left\|w_{i}\right\|_{\infty}$ denote the maximum magnitude entry in $w_{i}$.
2. Let $N$ be a random $k \times n$ matrix where $N_{i j} \sim N\left(0, \alpha_{i}^{2} \rho^{2}\right)$ for $i \in[k], j \in[n]$ and $\rho=2 \varepsilon^{-1} \sqrt{8 k \ln (4 k / \delta) \ln (2 / \delta)}$.
3. Compute the matrix $B=W\left(W^{T} A+N\right)$.

Output: Matrix $B$ of rank $k$.
Figure 3: Privacy-preserving projection

Lemma 4.6. The output B of the algorithm satisfies $(\varepsilon, \delta)$-differential privacy.
Proof. We will argue that releasing $W^{T} A+N$ preserves $(\varepsilon, \delta)$-differential privacy. That releasing $B$ preserves differential privacy follows from the fact that differential privacy does not degrade under arbitrary post-processing. Fix any two neighboring matrices $A, A^{\prime}$ differing in their $i$ 'th row. Let $E=A-A^{\prime}$ and let $e^{T}=A_{(i)}-A_{(i)}^{\prime}=E_{(i)}$. Recall by the definition of neighboring, $\|e\|_{2} \leqslant 1$, and for all other $j \neq i,\left\|E_{j}\right\|_{2}=0$. For any $j \in[k]$, consider the $j$ 'th row of $W^{T} E$ :

$$
\left\|\left(W^{T} E\right)_{(j)}\right\|_{2}=\sqrt{\sum_{\ell=1}^{n} W_{\ell, j}^{2} \cdot e_{\ell}^{2}} \leqslant \alpha_{j}\|e\|_{2}=\alpha_{j} .
$$

Hence, by Theorem 2.6, releasing $\left(W^{T} E\right)_{(j)}+g^{T}$ where $g \sim N\left(0, \alpha_{j}^{2} \rho^{2}\right)^{n}$ preserves $\left(\frac{\varepsilon}{\sqrt{8 k \ln (2 / \delta}}, \frac{\delta}{2 k}\right)$ differential privacy. Finally, we apply Theorem 2.4 to see that releasing each of the $k$ rows of $W^{T} E$
preserves $\left(\varepsilon^{\prime}, k(\delta / 2 k)+\delta / 2\right)=\left(\varepsilon^{\prime}, \delta\right)$-differential privacy for:

$$
\varepsilon^{\prime} \leqslant \sqrt{2 k \ln (2 / \delta)} \cdot \frac{\varepsilon}{\sqrt{8 k \ln (2 / \delta)}}+2 k\left(\frac{\varepsilon}{\sqrt{8 k \ln (2 / \delta)}}\right)^{2} \leqslant \varepsilon
$$

as desired.
Theorem 4.7. The algorithm above returns a matrix $B$ such that $B$ satisfies $(\varepsilon, \delta)$-differential privacy and moreover with probability 99/100,

$$
\|A-B\|_{F} \leqslant\left\|A-W W^{T} A\right\|_{F}+O\left(\sqrt{k \sum_{i=1}^{k} \alpha_{i}^{2} \rho^{2} n}\right) .
$$

In particular if $\max _{i} \alpha_{i}=\alpha$, we have with the same probability,

$$
\|A-B\|_{F} \leqslant\left\|A-W W^{T} A\right\|_{F}+O\left(\frac{\alpha k \log (k / \delta) \sqrt{n}}{\varepsilon}\right) .
$$

Proof.

$$
\|A-B\|_{F}=\left\|A-W\left(W^{T} A+N\right)\right\|_{F}=\left\|A-W W^{T} A-W N\right\|_{F} \leqslant\left\|A-W W^{T} A\right\|+\|W N\|_{F}
$$

But $\|W\|_{F}=\sqrt{k}$ so that, by Fact 2.8,

$$
\|W N\|_{F} \leqslant\|W\|_{F}\|N\|_{F}=\sqrt{k} \cdot\|N\|_{F} .
$$

On the other hand, by Jensen's inequality and linearity of expectation,

$$
\mathbb{E}\|N\|_{F} \leqslant \sqrt{\mathbb{E}\|N\|_{F}^{2}}=\sqrt{\sum_{i, j} \mathbb{E} N_{i j}^{2}}=\sqrt{\sum_{i=1}^{k} c_{k}^{2} \rho^{2} n} .
$$

The claim now follows from Markov's inequality.
Note that the quantities $\alpha_{i}$ are always bounded by 1 , since all $w_{i}$ 's are unit vectors. In the next section, we will show that under certain incoherence assumptions, we will have (or will be able to enforce) the condition that the $\alpha_{i}$ values are bounded significantly below 1 .

## 5 Incoherent matrices

Intuitively speaking, a matrix is incoherent if its left singular vectors have low correlation with the standard basis vectors. There are multiple ways to formalize this intuition. Here, we will work with two natural notions of coherence. In both cases we will be able to show that we can findin a privacy-preserving way-projection operators that have small entries. As demonstrated in the previous section, this directly leads to improvements over randomized response.

### 5.1 C-coherent matrices

In this section we work with matrices $A$ in which row norms do not deviate by too much from the typical row norm. Another way to look at this condition is that coordinate projections provide little spectral information about the matrix $A$. From this angle the condition we need can be interpreted as low coherence in the sense that the singular vectors of $A$ that correspond to large singular values must be far from the standard basis.

Definition 5.1 ( $C$-coherence). We say that a matrix $A \in \mathbb{R}^{m \times n}$ is $C$-coherent if

$$
\max _{i \in[m]}\left\|e_{i}^{T} A\right\| \leqslant C \cdot \frac{\|A\|_{F}}{\sqrt{m}} .
$$

Note that we have $1 \leqslant C \leqslant \sqrt{m}$.
The next lemma shows that sparse vectors have poor correlation with the matrix in the above sense. We say a vector $w$ is $\ell$-sparse if it has at most $\ell$ nonzero coordinates.

Lemma 5.2. Let $A \in \mathbb{R}^{m \times n}$ be a $C$-coherent matrix. Let $w$ be an $\ell$-sparse unit vector in $\mathbb{R}^{m}$. Then,

$$
\left\|\omega^{T} A\right\| \leqslant \frac{C \sqrt{\ell}\|A\|_{F}}{\sqrt{m}} .
$$

Proof. Since $w$ is $\ell$-sparse we can write it as $w=\sum_{i=1}^{\ell} \alpha_{i} e_{i}$ where $e_{1}, \ldots, e_{\ell}$ are $\ell$ distinct standard basis vectors and $\sum_{i} \alpha_{i}^{2}=1$. Hence,

$$
\left\|w^{T} A\right\|=\left\|\sum_{i=1}^{\ell} \alpha_{i} e_{i}^{T} A\right\| \leqslant \sum_{i=1}^{\ell} \alpha_{i}\left\|e_{i}^{T} A\right\| \leqslant \frac{C \sqrt{\ell}\|A\|_{F}}{\sqrt{m}} .
$$

In the last step we used the Cauchy-Schwarz inequality and the fact that $A$ is $C$-coherent.
Lemma 5.3. Let $\alpha>0$. Let $A \in \mathbb{R}^{m \times n}$ be a $C$-coherent matrix. Let $w \in \mathbb{R}^{m}$ be a unit vector and suppose $w_{\alpha}$ is the vector obtained from $w$ by zeroing all coordinates greater than $\alpha$. Then,

$$
w_{\alpha}^{T} A=w^{T} A+e
$$

where $e$ is a vector of norm

$$
\|e\| \leqslant \frac{C\|A\|_{F}}{\alpha \sqrt{m}} .
$$

Proof. Note that $w-w_{\alpha}$ is an $\ell$-sparse vector with $\ell \leqslant 1 / \alpha^{2}$. Here we used that $w$ is a unit vector and hence there can be at most $1 / \alpha^{2}$ coordinates larger than $\alpha$. The lemma now follows directly from Lemma 5.2.

The next lemma is a straightforward extension of the previous one for the case where we multiply $A$ by a matrix $W$ rather than a single vector.

Lemma 5.4. Let $\alpha>0$. Let $A \in \mathbb{R}^{m \times n}$ be a C-coherent matrix. Let $W \in \mathbb{R}^{m \times k}$ be a matrix whose columns have unit length. Suppose $W_{\alpha}$ is the matrix obtained from $W$ by zeroing all entries greater than $\alpha$. Then,

$$
W W_{\alpha}^{T} A=W W^{T} A+E
$$

where $E$ is a matrix of Frobenius norm

$$
\|E\|_{F} \leqslant \frac{C k\|A\|_{F}}{\alpha \sqrt{m}}
$$

Proof. By the previous lemma, we have $W_{\alpha}^{T} A=W^{T} A+E^{\prime}$, where every row of $E^{\prime}$ has Euclidean norm $C\|A\|_{F} / \alpha \sqrt{m}$. Hence, $\left\|E^{\prime}\right\|_{F} \leqslant C \sqrt{k}\|A\|_{F} / \alpha \sqrt{m}$. But then

$$
W W_{\alpha}^{T} A=W W^{T} A+W E^{\prime}
$$

Put $E=W E^{\prime}$ and note that $\|E\|_{F} \leqslant\|W\|_{F}\left\|E^{\prime}\right\|_{F}=\sqrt{k}\left\|E^{\prime}\right\|_{F}$. The lemma follows.
The previous lemma quantifies what happens if we replace $W W^{T} A$ by $W W_{\alpha}^{T} A$. Working with $W_{\alpha}^{T}$ for small $\alpha$ instead of $W^{T}$ will decrease the sensitivity of the computation of $W_{\alpha}^{T} A$. On the other hand, by the previous lemma we have an expression for the error resulting from the truncation step.

### 5.2 Strong coherence

Here we introduce and work with the notion of $\mu_{0}$-coherence which is a standard notion of coherence. As we will see in Section 5.3, it is a stronger notion than $C$-coherence. Consequently, the results we will be able to obtain using $\mu_{0}$-coherence are stronger than our previous results on $C$-coherence in certain aspects.
Definition 5.5 ( $\mu_{0}$-coherence). Let $U$ be an $m \times r$ matrix with orthonormal columns and $r \leqslant n$. Recall, that $P_{U}=U U^{T}$. The $\mu_{0}$-coherence of $U$ is defined as

$$
\begin{equation*}
\mu_{0}(U)=\frac{m}{r} \max _{1 \leqslant j \leqslant m}\left\|P_{U} e_{j}\right\|^{2}=\frac{m}{r} \max _{1 \leqslant j \leqslant m}\left\|U_{(j)}\right\|^{2} \tag{8}
\end{equation*}
$$

Here, $e_{j}$ denotes the $j$-th $m$-dimensional standard basis vector and $U_{(j)}$ denotes the $j$-th row of $U$.
The $\mu_{0}$-coherence of an $m \times n$ matrix $A$ of rank $r$ given in its singular value decomposition $U \Sigma V^{T}$ where $U \in \mathbb{R}^{m \times r}$ is defined as $\mu_{0}(U)$.

Fact 5.6. $1 \leqslant \mu_{0}(U) \leqslant m$
Proof. Since $U$ is orthonormal, there must always exists a row of square norm $r / m$. On the other hand, no row of $U$ has squared norm larger than $r$.

The above notion is used extensively throughout the literature in the context of matrix completion and low rank approximation, e.g., in Candes and Recht [CR09], Keshavan et al. [KMO10], Talwalkar and Rostamizadeh [TR10], Mohri and Talwalkar [MT11]. Motivated by the Netflix problem, Candes and Tao [CT10] study matrix completion for matrices satisfying a stronger incoherence assumption than small $\mu_{0}$-coherence.

Our goal from here on is to show that if we run our range finding algorithm from Section 4.1 on a low-coherence matrix it will produce a projection matrix with small entries. This result (presented in Lemma 5.11) requires several technical lemmas.

The first technical step is a lemma showing that vectors that lie in the range of an incoherent matrix must have small $\ell_{\infty}$-norm.

Lemma 5.7. Let $U$ be an orthonormal $m \times r$ matrix. Suppose $w \in \operatorname{range}(U)$ and $\|w\|=1$. Then,

$$
\|w\|_{\infty}^{2} \leqslant \frac{r}{m} \cdot \mu_{0}(U) .
$$

Proof. Let $u_{1}, \ldots, u_{r}$ denote the columns of $U$. By our set of assumptions, there exist $\alpha_{1}, \ldots, \alpha_{r} \in \mathbb{R}$ such that

$$
w=\sum_{i=1}^{r} \alpha_{i} u_{i} \quad \text { and } \quad \sum_{i} \alpha_{i}^{2}=1
$$

Therefore, denoting the $j$-th entry of $w$ by $w_{j}$ and the $j$-th entry of $u_{i}$ by $u_{i j}$, we have

$$
\begin{aligned}
\left|w_{j}\right|^{2}=\left(\sum_{i=1}^{r} \alpha_{i} u_{i j}\right)^{2} & \leqslant\left(\sum_{i=1}^{r} \alpha_{i}^{2}\right)\left(\sum_{i=1}^{r} u_{i j}^{2}\right) \\
& =\left\|U_{(j)}\right\|^{2} .
\end{aligned}
$$

(by Cauchy-Schwarz)

In particular $\|w\|_{\infty}^{2} \leqslant \max _{j \in[m]}\left\|U_{(j)}\right\|^{2}$. On the other hand, using Definition 5.5,

$$
\|w\|_{\infty}^{2} \leqslant \max _{j \in[m]}\left\|U_{(j)}\right\|^{2}=\frac{r}{m} \cdot \mu_{0}(U) .
$$

The lemma follows.
We will need the following geometric lemma: If we start with a small orthonormal set of vectors of low coherence and we append few random unit vectors, then the span of the resulting set of vectors has a low coherence basis.

Lemma 5.8. Let $u_{1}, \ldots, u_{r} \in \mathbb{R}^{m}$ be orthonormal vectors. Pick unit vectors $n_{1}, \ldots, n_{k} \in \mathbb{S}^{m-1}$ uniformly at random. Assume that

$$
\begin{equation*}
m \geqslant c_{0} k(r+k) \log (r+k) \tag{9}
\end{equation*}
$$

where $c_{0}$ is a sufficiently large constant. Then, there exists a set of orthonormal vectors $v_{1}, \ldots, v_{r+k} \in$ $\mathbb{R}^{m}$ such that $\operatorname{span}\left\{v_{1}, \ldots, v_{r+k}\right\}=\operatorname{span}\left\{u_{1}, \ldots, u_{r}, n_{1}, \ldots, n_{k}\right\}$ and furthermore, with probability 99/100,

$$
\mu_{0}\left(\left[v_{1}|\cdots| v_{r+k}\right]\right) \leqslant 2 \mu_{0}\left(\left[u_{1}|\cdots| u_{k}\right]\right)+O\left(\frac{k \log m}{r}\right)
$$

Proof. We will construct the basis iteratively using the Gram-Schmidt orthonormalization algorithm starting with the partial orthonormal basis $u_{1}, \ldots, u_{r}$. The algorithm works as follows: At iteration $i$ we have obtained a partial orthonormal basis $v_{1}, \ldots, v_{t}$ where $t=r+i-1$. We then pick a random unit vector $v \in \mathbb{S}^{m-1}$ and let $v^{\prime}=\sum_{i=1}^{t} v_{i} v_{i}^{T} v$. Put

$$
v_{t+1}=\frac{v-v^{\prime}}{\left\|v-v^{\prime}\right\|}
$$

Let $V_{t}=\left[v_{1}|\cdots| v_{t}\right]$ and $V_{t+1}=\left[V_{t} \mid v_{t+1}\right]$. Our goal is to bound $\left\|v_{t+1}\right\|_{\infty}^{2}$ as this will directly lead to a bound on $\mu_{0}\left(V_{t+1}\right)$ in terms of $V_{t}$. Summing up this bound over $t$ will lead to a bound on $\mu_{0}\left(V_{r+k}\right)$ which is what the lemma is asking for.

Let us start with a two simple claims that follow from measure concentration on the sphere. The first one bounds the $\ell_{\infty}$-norm of a random unit vector.

Claim 5.9. $\|v\|_{\infty}^{2} \leqslant O\left(\frac{\log m}{m}\right)$ with probability $1-1 / 200 k$.
Proof. It is not hard to show that for every $i \in[m]$, the coordinate projection $f_{i}(v)=v_{i}$ is a Lipschitz function on the sphere. Moreover, the median of $f_{i}$ is 0 by spherical symmetry. By measure concentration (Theorem B.1), $\operatorname{Pr}\left\{\left|f_{i}\right|>\varepsilon\right\} \leqslant O\left(\exp \left(-\varepsilon^{2} m / 2\right)\right.$ ). Setting $\varepsilon=O(\sqrt{(\log m+\log k) / m})=$ $O(\sqrt{\log (m) / m})$ and taking a union bound over all $m$ coordinates completes the proof.

The second claim we need bounds the Euclidean norm of $v^{\prime}$.
Claim 5.10. $\left\|v^{\prime}\right\|^{2} \leqslant O\left(\frac{r+k}{m}\right)$ with probability $1-1 / 200 k$.
Proof. Proceeding as in proof of the previous claim, we note that for each $i \in[t], f_{i}(v)=\left\langle v_{i}, v\right\rangle$ is a Lipschitz function on the sphere with median 0 . Applying Theorem B. 1 with $\varepsilon=O(\sqrt{(\log t+\log k) / m})$, it follows that with probability $1-1 / 200 \mathrm{kt}$,

$$
f_{i}(v)^{2} \leqslant O\left(\frac{\log k+\log t}{m}\right)
$$

Taking a union bound over all $i \in[t]$ we have with probability $1-1 / 200 k$,

$$
\left\|v^{\prime}\right\|^{2}=\sum_{i=1}^{t}\left\langle v_{i}, v\right\rangle^{2} \leqslant O\left(\frac{t(\log k+\log t)}{m}\right)=O\left(\frac{(r+k) \log (r+k))}{m}\right),
$$

where we used that $t \leqslant r+k$.
On the one hand, note that $v^{\prime}$ is in the span of $v_{1}, \ldots, v_{t}$ by definition. Hence, Lemma 5.7 directly implies that

$$
\begin{equation*}
\left\|v^{\prime}\right\|_{\infty}^{2} \leqslant \frac{t}{m} \cdot\left\|v^{\prime}\right\|^{2} \cdot \mu_{0}\left(V_{t}\right) . \tag{10}
\end{equation*}
$$

Hence, combining Equation 10 with Claim 5.10, we have with probability $1-1 / 200 k$,

$$
\begin{equation*}
\left\|v^{\prime}\right\|_{\infty}^{2} \leqslant O\left(\frac{t(r+k) \log (r+k)}{m^{2}} \cdot \mu_{0}\left(V_{t}\right)\right) . \tag{11}
\end{equation*}
$$

On the other hand, we can bound $\left\|v_{t+1}\right\|_{\infty}^{2}$ as follows:

$$
\left\|v_{t+1}\right\|_{\infty}^{2}=\frac{\left\|v-v^{\prime}\right\|_{\infty}^{2}}{\left\|v-v^{\prime}\right\|^{2}} \leqslant \frac{\|v\|_{\infty}^{2}+2\|v\|_{\infty}\left\|v^{\prime}\right\|_{\infty}+\left\|v^{\prime}\right\|_{\infty}^{2}}{\left\|v-v^{\prime}\right\|^{2}} \leqslant \frac{3\left(\|v\|_{\infty}^{2}+\left\|v^{\prime}\right\|_{\infty}^{2}\right)}{\left\|v-v^{\prime}\right\|^{2}} .
$$

By Claim 5.10 we have that with probability $1-1 / 200 k$,

$$
\left\|v-v^{\prime}\right\|^{2}=\|v\|^{2}+\left\|v^{\prime}\right\|^{2}-2\left\langle v, v^{\prime}\right\rangle \geqslant 1-2\left\|v^{\prime}\right\|^{2} \geqslant 1-O\left(\frac{(r+k) \log (r+k)}{m}\right) .
$$

In the first inequality above we used that $\left\langle v, v^{\prime}\right\rangle=\sum_{i=1}^{t}\left\langle v_{i}, v\right\rangle^{2}=\left\|v^{\prime}\right\|^{2}$. We then applied Claim 5.10 in the second inequality. By Equation $9, m$ is sufficiently large so that

$$
\begin{equation*}
\frac{1}{\left\|v-v^{\prime}\right\|^{2}} \leqslant O(1) . \tag{12}
\end{equation*}
$$

Combining Equation 11 with Equation 12 and applying Claim 5.9, we conclude that with with probability at least $1-1 / 100 k$,

$$
\left\|v_{t+1}\right\|_{\infty}^{2} \leqslant O\left(\frac{\log m}{m}+\frac{t(r+k) \log (r+k)}{m^{2}} \cdot \mu_{0}\left(V_{t}\right)\right)
$$

But when the above bound on $\left\|v_{t+1}\right\|_{\infty}^{2}$ holds, then we must have

$$
\begin{equation*}
\mu_{0}\left(V_{t+1}\right) \leqslant \mu_{0}\left(V_{t}\right)+\frac{m}{t+1}\left\|v_{t+1}\right\|_{\infty}^{2} \leqslant\left(1+O\left(\frac{(r+k) \log (r+k)}{m}\right)\right) \mu_{0}\left(V_{t}\right)+O\left(\frac{\log m}{t}\right) \tag{13}
\end{equation*}
$$

Taking a union bound over all $k$ steps, we find that with probability $99 / 100$, Equation 13 is true at all steps of the Gram-Schmidt algorithm. Assuming that this event occurs, we have:

$$
\begin{aligned}
\mu_{0}\left(V_{r+k}\right) & \leqslant\left(1+O\left(\frac{(r+k) \log (r+k)}{m}\right)\right)^{k} \mu_{0}\left(V_{r}\right)+O\left(\frac{k \log m}{r}\right) \\
& \leqslant 2 \mu_{0}\left(V_{r}\right)+O\left(\frac{k \log m}{r}\right) \quad \quad(\text { since } m \gg k(r+k) \log (r+k) \text { by Equation 9) }
\end{aligned}
$$

This finishes our proof of the lemma since $\mu_{0}\left(V_{r}\right)=\mu_{0}\left(\left[u_{1}|\cdots| u_{r}\right]\right)$ by definition.
The choice of failure probability in the previous lemma was rather arbitrary and stronger bounds can be achieved. We finally arrive at the main lemma in this section.

Lemma 5.11. Let $A$ be an $m \times n$ matrix of rank $r$. Let $\Omega \sim N(0,1)^{n \times k}$ with $k \leqslant r$ denote a random standard Gaussian matrix and define $Y=A \Omega$. Assume that $m \geqslant c_{0} k r \log r$ for sufficiently large constant $c_{0}$. Further, let $\sigma>0$ and $N \sim N\left(0, \sigma^{2}\right)^{m \times k}$ denote a random Gaussian matrix with i.i.d. entries sampled from $N\left(0, \sigma^{2}\right)$. Put $\tilde{Y}=A \Omega+N$ and let $w_{1}, \ldots, w_{k}$ be an orthonormal basis for the range of $\tilde{Y}$. Then, with probability 99/100,

$$
\max _{i \in[k]}\left\|w_{i}\right\|_{\infty} \leqslant \sqrt{\frac{4 r}{m} \cdot \mu_{0}(A)}+O\left(\sqrt{\frac{k \log m}{m}}\right) .
$$

Proof. Let $U$ denote the left singular factor of $A$. Let $u_{1}, \ldots, u_{r}$ denote the columns of $U$. We have,

$$
\operatorname{span}\left(\left\{w_{1}, \ldots, w_{k}\right\}\right)=\operatorname{range}(\tilde{Y}),
$$

since $w_{1}, \ldots, w_{k}$ is an orthonormal basis for the range of $\tilde{Y}$ by construction. On the other hand,

$$
\operatorname{range}(Y) \subseteq \operatorname{range}(A)=\operatorname{span}\left(\left\{u_{1}, \ldots, u_{r}\right\}\right) .
$$

Since $\tilde{Y}=Y+N$ this implies that $\operatorname{range}(\tilde{Y}) \subseteq \operatorname{span}\left\{u_{1}, \ldots, u_{r}, n_{1}, \ldots n_{k}\right\}$, where $n_{1}, \ldots, n_{k}$ are the columns of $N$ normalized such that $\left\|n_{i}\right\|=1$. By assumption $m$ is large enough so that we can apply Lemma 5.8. Thus we obtain orthonormal vectors $v_{1}, \ldots, v_{r+k}$ satisfying

$$
\operatorname{range}(\tilde{Y}) \subseteq \operatorname{span}\left\{v_{1}, \ldots, v_{r+k}\right\}
$$

and the matrix $V$ whose columns are $v_{1}, \ldots, v_{r+k}$ has coherence

$$
\mu_{0}(V) \leqslant 2 \mu_{0}(U)+O\left(\frac{k \log m}{r}\right)
$$

with probability 99/100. In particular, $w_{i} \in \operatorname{range}(V)$ for all $i \in[k]$. Therefore, by Lemma 5.7, we have that

$$
\max _{i \in[k]}\left\|w_{i}\right\|_{\infty}^{2} \leqslant \frac{r+k}{m} \cdot \mu_{0}(V) \leqslant \frac{2(r+k)}{m} \cdot \mu_{0}(U)+O\left(\frac{(r+k) k \log m}{r m}\right) .
$$

Since $k \leqslant r$ and $\mu_{0}(A)=\mu_{0}(U)$, we conclude that

$$
\max _{i \in[k]}\left\|w_{i}\right\|_{\infty} \leqslant \sqrt{\frac{4 r}{m} \cdot \mu_{0}(A)}+O\left(\sqrt{\frac{k \log m}{m}}\right) .
$$

The lemma follows.
Remark 5.12. We remark that the previous lemma is essentially tight. Indeed, under the given assumption on $A$ there could be a left singular vector of $\ell_{\infty}$-norm $\sqrt{r \mu_{0}(A) / m}$. The above lemma implies that we are never more than a $O(\sqrt{\log m})$-factor away from this bound.

### 5.3 Relation between $C$-coherence and $\mu_{0}$-coherence

Here we show that the assumption of small $\mu_{0}$-coherence is strictly stronger than that of small $C$ coherence assuming the rank of the matrix is not too large.
Lemma 5.13. Let $A$ be an $m \times n$ matrix of rank $r$. Then, $A$ is $C$-coherent where

$$
C \leqslant \sqrt{r \mu_{0}(A)}
$$

Proof. Let the SVD of $A$ be $U \Sigma V^{T}$ and denote the right singular vectors by $v_{1}, \ldots, v_{r}$. Extend them arbitrarily to an orthonormal basis of $\mathbb{R}^{n}$, denoted $v_{1}, \ldots, v_{n}$. We then have for every $j \in[m]$,

$$
\begin{equation*}
\left\|e_{j}^{T} A\right\|^{2}=\sum_{i=1}^{n}\left\langle e_{j}^{T} A, v_{i}\right\rangle^{2}=\sum_{i=1}^{r}\left(\sigma_{i}\left\langle e_{j}, u_{i}\right\rangle\right)^{2} \leqslant\left(\sum_{i=1}^{r}\left|\sigma_{i}\left\langle e_{j}, u_{i}\right\rangle\right\rangle\right)^{2}, \tag{14}
\end{equation*}
$$

where we used that the $\ell_{2}^{2}$-norm of a vector is bounded by the $\ell_{1}^{2}$-norm. On the other hand,

$$
\begin{equation*}
\left(\sum_{i=1}^{r} \mid \sigma_{i}\left\langle e_{j}, u_{i}\right\rangle\right)^{2} \leqslant\left(\sum_{i=1}^{r}\left|\sigma_{i}\right|\left|\left\langle e_{j}, u_{i}\right\rangle\right|\right)^{2} \leqslant\left(\sum_{i=1}^{r} \sigma_{i}^{2}\right)\left(\sum_{i=1}^{r}\left\langle e_{j}, u_{i}\right\rangle^{2}\right)=\|A\|_{F}^{2} \cdot\left\|U_{j}\right\|^{2} \tag{15}
\end{equation*}
$$

where we used Cauchy-Schwarz in the inequality. It follows that

$$
\max _{j \in[m]}\left\|e_{j}^{T} A\right\|^{2}=\|A\|_{F}^{2} \max _{j \in[m]}\left\|U_{j}\right\|^{2}=\|A\|_{F}^{2} \frac{r \mu_{0}(A)}{m}
$$

Taking square roots on both sides and rearranging, we find

$$
\frac{\sqrt{m}}{\|A\|_{F}} \cdot \max _{j \in[m]}\left\|e_{j}^{T} A\right\| \leqslant \sqrt{r \mu_{0}(A)} .
$$

Note that the left hand side is exactly the smallest $C$ for which $A$ is $C$-coherent. This proves the lemma.

Recall that Lemma 5.2 showed that the singular vectors corresponding to large singular values of a $C$-coherent matrix $A$ cannot be too sparse. In particular, the top singular vectors must have small $\mu_{0}$-coherence as a result. However, we cannot rule out that there are singular vectors corresponding to small singular values that do have large coordinates.

## 6 Privacy-preserving low rank approximations

In this section we compose the range finder, projection and truncation step to get a private low rank approximation algorithm suitable for matrices of low coherence.

Input: Matrix $A \in \mathbb{R}^{m \times n}$, target rank $r \geqslant 2$, oversampling parameter $p \geqslant 2$, pruning parameter $\alpha>0$, privacy parameters $\varepsilon, \delta \in(0,1)$.

1. Range finder: Run the range finder (Figure 2) on $A$ with sampling parameter $k=p+r$ and privacy parameters ( $\varepsilon / 2, \delta / 2$ ). Let the output be denoted by $W$.
2. Pruning: Let $W^{\prime}$ be the matrix obtained from $W$ by zeroing out all entries larger than $\alpha$.
3. Projection: Run the projection algorithm (Figure 3) on input $A, W^{\prime}$ and privacy parameters $(\varepsilon / 2, \delta / 2)$. Let $B$ denote the output of the projection algorithm.

Output: Matrix $B$ of rank $k=(r+p)$.
Figure 4: The private find and project algorithm (PFP) for computing privacy-preserving low-rank approximations

## Lemma 6.1. The PFP algorithm satisfies ( $\varepsilon, \delta$ )-differential privacy.

Proof. This follows directly from composition and the privacy guarantee achieved by the subroutines.

The next theorem details the performance of PFP on $C$-coherent matrices. In particular, it shows that in a natural range of parameters it improves significantly over randomized response (input perturbation).

Theorem 6.2 (Approximation for $C$-coherent matrices). There is an $(\varepsilon, \delta)$-differentially private algorithm that given a $C$-coherent matrix $A \in \mathbb{R}^{m \times n}$ and parameters $r \geqslant 2, p \geqslant 2$ produces a rank $k=r+p$ matrix $B$ such that with probability $9 / 10$,

$$
\begin{equation*}
\|A-B\|_{F} \leqslant O\left(\sqrt{1+\frac{r}{p-1}} \cdot\left\|A-A_{r}\right\|_{F}+\frac{\sqrt{k m} \log (k / \delta)}{\varepsilon}+\sqrt{C\|A\|_{F}} k\left(\frac{n}{m}\right)^{1 / 4} \frac{\log (k / \delta)^{1 / 2}}{\varepsilon^{1 / 2}}\right) . \tag{16}
\end{equation*}
$$

In particular, the second error term is o $(\sqrt{k n \log (k / \delta)} / \varepsilon)$, whenever

$$
\begin{equation*}
m=o(n) \quad \text { and } \quad \frac{C k\|A\|_{F} \sqrt{\log (k / \delta)}}{\sqrt{n}}=o(\sqrt{m}) . \tag{17}
\end{equation*}
$$

We generally think of $C, k$ as small compared to both $m$ and $n$. Equation 17 states that the algorithm outperforms randomized response whenever $m$ is not too large compared to $n$ and not too small compared to the rank $k$, the Frobenius norm of $A$ divided by $\sqrt{n}$, and the coherence parameter $C$. These two conditions are naturally satisfied for a wide range of parameters. For example, when $\|A\|_{F}=O(\sqrt{k n})$ (so that randomized response no longer provides non-trivial error) and $C=O(1)$ (i.e., the matrix is very incoherent), then the requirement on $m$ is just that

$$
\omega\left(k^{3}\right) \leqslant m \leqslant o(n) .
$$

The proof of Theorem 6.2 is a straightforward combination of our previous error bounds for range finding, pruning and projection.

Proof of Theorem 6.2. We run PFP with the given set of parameters $r, p, \varepsilon, \delta$ and a suitable choice of the pruning parameter $\alpha>0$. Before fixing $\alpha$, we claim that the error of the algorithm satisfies, with probability $9 / 10$,

$$
\|A-B\|_{F} \leqslant O\left(\sqrt{1+\frac{r}{p-1}} \cdot\left\|A-A_{r}\right\|_{F}+\frac{C k\|A\|_{F}}{\alpha \sqrt{m}}+(\sqrt{k m}+\alpha k \sqrt{n}) \cdot \frac{\log (k / \delta)}{\varepsilon}\right)
$$

Here, the first term follows from Theorem 3.1 and an application of Markov's inequality to argue that the bound holds except with sufficiently small constant probability. The other terms follow from Theorem 4.7 (error bound of the projection algorithm), Corollary 4.5 (error bound of the range finder), and, Lemma 5.4 (error bound for the pruning step with parameter $\alpha$ ). We can now optimize $\alpha$ so as to achieve the geometric mean between the two terms that it appears in (as $\alpha$ and $1 / \alpha$ ). Running PFP with this choice of $\alpha$ directly results in the error bound stated in Equation 16. Equation 17 is now easily verified by equating the $O(\cdot)$-term in Equation 16 with $o(\sqrt{k n \log (k / \delta)} / \varepsilon)$ and rearranging.

Since all sub-routines fail with probability at most $1 / 100$, we can take a union bound to conclude that the algorithm fails to satisfy the error bound with probability at most $1 / 10$.

We will next analyze the performance of PFP on $\mu_{0}$-incoherent matrices. In this case no truncation is necessary, since we argued that the projection matrix with high probability already has very small entries. The error bound here is stronger in certain aspects as we will discuss in a moment.

Theorem 6.3 (Approximation for $\mu_{0}$-coherent matrices). There is an $(\varepsilon, \delta)$-differentially private algorithm that given a rank $R$ matrix $A \in \mathbb{R}^{m \times n}$ and parameters $r \geqslant 2, p \geqslant 2$ such that $k=r+p \leqslant R$ and $m \geqslant \omega(R k \log R)$ produces a rank $k$ matrix $B$ such that with probability $9 / 10$,

$$
\begin{equation*}
\|A-B\|_{F} \leqslant O\left(\sqrt{\frac{r}{p-1}} \cdot\left\|A-A_{r}\right\|_{F}+\left(\sqrt{k m}+\sqrt{\frac{k R \mu_{0}(A)+k^{2} \log m}{m}} \sqrt{k n}\right) \cdot \frac{\log (k / \delta)}{\varepsilon}\right) \tag{18}
\end{equation*}
$$

In particular, the error is $o(\sqrt{k n \log (k / \delta)} / \varepsilon)$, whenever

$$
\begin{equation*}
m=o(n) \quad \text { and } \quad R k\left(\mu_{0}(A)+\log m\right) \sqrt{\log (k / \delta)}=o(m) \tag{19}
\end{equation*}
$$

Just as in the previous theorem we get a range for $m$ in which the algorithm improves over randomized response. Here, we need the coherence of $A$ to be small compared to $m$. We also observe a dependence on the rank of the matrix. This means the algorithm presents no improvement if the matrix is close to being full rank. Recall that $\mu_{0}(A)$ can be as small as $O(1)$. In particular, in the natural case where $\mu_{0}(A), k, R$ all are small compared to $m$, e.g., $m^{0.3}$, the requirement in Equation 19 reduces to $m=o(n)$.

Note that Theorem 6.3 is quantitatively stronger than Theorem 6.2 in the following regime: When $k, R, C, \mu_{0}(A)$ are all small (e.g., $n^{o(1)}$ ), $m \leqslant \sqrt{n}$ and $\|A\|_{F}^{2} \geqslant n$, then Theorem 6.3 improves over randomized response by a factor of roughly $\sqrt{m}$, whereas Theorem 6.2 achieves an $m^{1 / 4}$-factor improvement.

Proof of Theorem 6.3. We run PFP with the given set of parameters $r, p, \varepsilon, \delta$ and $\alpha=1$. Note that this choice of $\alpha$ implies that we never modify the matrix returned by the range finder. We claim that the error of the algorithm is with probability $9 / 10$,

$$
\|A-B\|_{F} \leqslant O\left(\sqrt{1+\frac{r}{p-1}} \cdot\left\|A-A_{r}\right\|_{F}+\left(\sqrt{k m}+\sqrt{\frac{k R \mu_{0}(A)+k^{2} \log m}{m}} \sqrt{k n}\right) \cdot \frac{\log (k / \delta)}{\varepsilon}\right)
$$

which is what we stated in the theorem. The first error term follows as before from Theorem 3.1 and Markov's inequality so that it holds with probability 99/100. The term of $O(\sqrt{k m \log (k / \delta)} / \varepsilon)$ follows from Corollary 4.5. To understand the remaining terms that by Lemma 5.11 we have that the matrix $W=\left[w_{1}|\cdots| w_{k}\right]$ returned by the range finder satisfies with probability $99 / 100$,

$$
\alpha=\max _{i \in[k]}\left\|w_{i}\right\|_{\infty} \leqslant \sqrt{\frac{4 R}{m} \cdot \mu_{0}(A)}+O\left(\sqrt{\frac{k \log m}{m}}\right) .
$$

In applying Lemma 5.11 we needed that $m \geqslant c_{0} k R \log R$ for sufficiently large constant which is satisfied by our assumption. Hence, Theorem 4.7 ensures that the error resulting from the projection operation is at most $O(\alpha k \sqrt{n \log (k / \delta)} / \varepsilon)$. Expanding $\alpha$ in the latter bound gives the stated error term. Equation 17 is now easily verified by equating the $O(\cdot)$-term in Equation 18 with $o(\sqrt{k n \log (k / \delta)} / \varepsilon)$ and rearranging.

Again, we can take a union bound over the failure probabilities of the sub-routines to bound the probability that our algorithm fails to satisfy the stated bound by $1 / 10$.

## References

[AM01] Dimitris Achlioptas and Frank McSherry. Fast computation of low rank matrix approximations. In Proc. 33rd Symposium on Theory of Computing (STOC), pages 611-618. ACM, 2001.
[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In Proc. 24th Symposium on Principles of Database Systems (PODS), pages 128-138. ACM, 2005.
$\left[\mathrm{CDM}^{+} 05\right]$ Shuchi Chawla, Cynthia Dwork, Frank Mcsherry, Adam Smith, and Hoeteck Wee. Toward privacy in public databases. In Proc. 2nd TCC, pages 363-385. Springer, 2005.
[CR09] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. Foundations of Computional Mathematics, 9:717-772, December 2009.
[CT10] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory, 56(5):2053-2080, 2010.
[De11] Anindya De. Lower bounds in differential privacy. CoRR, abs/1107.2183, 2011.
[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Proc. 3rd TCC, pages 265-284. Springer, 2006.
[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In Proc. 22nd PODS, pages 202-210. ACM, 2003.
[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In Proc. 51st Foundations of Computer Science (FOCS). IEEE, 2010.
[FKV04] Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. Journal of the ACM (JACM), 51:1025-1041, November 2004.
[ $\left.\mathrm{GLM}^{+} 10\right]$ Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. Differentially private combinatorial optimization. In Proceedings of the Twenty First Annual ACM-SIAM Symposium on Discrete Algorithms. ACM-SIAM, 2010.
[GRU11] Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release. $C o R R$, abs/1107.3731, 2011.
[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review, 53(2):217-288, 2011.
[KMO10] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. IEEE Transactions on Information Theory, 56(6):2980-2998, 2010.
[KMT11] Michael Kapralov, Frank McSherry, and Kunal Talwar. Personal communication. 2011.
[Mat02] Jiri Matousek. Lectures on Discrete Geometry. Springer-Verlag New York, Inc., 2002.
[MM09] Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the net. In Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 627-636. ACM, 2009.
[MT11] Mehryar Mohri and Ameet Talwalkar. Can matrix coherence be efficiently and accurately estimated? In Proc. 14th Conference on Artificial Intelligence and Statistics (AISTATS), 2011.
[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In Proc. 39th Symposium on Theory of Computing (STOC), pages 75-84. ACM, 2007.
[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In IEEE Symposium on Security and Privacy (S\&PP), pages 111-125. IEEE, 2008.
[PTRV98] Christos Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In Proc. 17th PODS, pages 159-168. ACM, 1998.
[TR10] Anand Talwalkar and Afshin Rostamizadeh. Matrix coherence and the nystrom method. In Proc. 26th Conference on Uncertainty in Artificial Intelligence (UAI), 2010.

## A Privacy of the Gaussian Mechanism

Theorem A. 1 (Gaussian Mechanism). Let $x, y \in \mathbb{R}^{d}$ be any two vectors such that $\|x-y\|_{2} \leqslant c$. Let $Y \in \mathbb{R}^{d}$ be an independent random draw from $N\left(0, \rho^{2}\right)^{d}$, where $\rho=c \varepsilon^{-1} \sqrt{\log 1.25 / \delta}$. Then for any $S \subseteq \mathbb{R}^{d}:$

$$
\operatorname{Pr}[x+Y \in S] \leqslant \exp (\varepsilon) \operatorname{Pr}[y+Y \in S]+\delta
$$

Proof. For a set $S \subseteq \mathbb{R}^{d}$, write $S-x$ to denote the set $\{s-x: s \in S\}$ and $S Q$ to denote $\{s Q: s \in S\}$. Write $S_{i}=\left\{s_{i}: s \in S\right\}$ to denote the projection of the set onto the $i$ 'th coordinate of its elements.

First we consider the one dimensional case, where $x, y \in \mathbb{R}$ and $\|x-y\|_{2}=|x-y| \leqslant c$. Without loss of generality, we may take $x=0$ and $y=c$. Let $T \subseteq S$ be the set $T=\left\{z \in S: z<\frac{\rho^{2} \varepsilon}{c}-\frac{c}{2}\right\}$ First, we argue that $\operatorname{Pr}[x+Y \in S \backslash T]=\operatorname{Pr}[Y \in S \backslash T] \leqslant \delta$. This follows directly from the tail bound:

$$
\operatorname{Pr}[Y \geqslant t] \leqslant \frac{\rho}{\sqrt{2 \pi}} \exp \left(-t^{2} / 2 \rho^{2}\right)
$$

Observing that:

$$
\operatorname{Pr}[Y \in S \backslash T] \leqslant \operatorname{Pr}\left[Y \geqslant \frac{\rho^{2} \varepsilon}{c}-\frac{c}{2}\right]
$$

and plugging in our choice of $\rho=c \varepsilon^{-1} \sqrt{\log 1.25 / \delta}$ completes the claim. Next we show that conditioned on the event that $Y \notin S \backslash T$, we have: $\operatorname{Pr}[x+Y \in S] \leqslant \exp (\varepsilon) \operatorname{Pr}[y+Y \in S]$. Conditioned on this event we have:

$$
\left|\ln \left(\frac{\operatorname{Pr}[Y \in S]}{\operatorname{Pr}[Y \in S-c]}\right)\right| \leqslant \max _{z \in T}\left|\ln \left(\frac{\operatorname{Pr}[Y=z]}{\operatorname{Pr}[Y=z-c]}\right)\right|=\left|\ln \left(\frac{\exp \left(-z^{2} / 2 \rho^{2}\right)}{\exp \left(-(z+c)^{2} / 2 \rho^{2}\right)}\right)\right|
$$

where here $\operatorname{Pr}[Y=t]$ denotes the probability density function of $N\left(0, \rho^{2}\right)$ at $t$. This quantity is bounded by $\varepsilon$ whenever:

$$
z \leqslant \frac{\rho^{2} \varepsilon}{c}-\frac{c}{2}
$$

i.e. whenever $z \in T$. Therefore:

$$
\operatorname{Pr}[x+Y \in S] \leqslant \exp (\varepsilon) \operatorname{Pr}[y+Y \in S]+\delta
$$

which completes the proof in the 1-dimensional case.
For the multi-dimensional case, we will take advantage of the rotational invariance of the Gaussian distribution to rotate any Euclidean length $c$-perturbation into a length $c$ standard basis vector, reducing it to the 1 -dimensional case.

Consider any two vectors $x, y \in \mathbb{R}^{d}$ such that $\|x-y\|_{2} \leqslant c$. Let $Q \in \mathbb{R}^{d \times d}$ be the orthonormal (rotation) matrix such that $(x-y) Q=c^{\prime} \cdot e_{1}$ where $e_{1} \in \mathbb{R}^{d}$ is the 1 st standard basis vector $e_{1}=$ $(1,0, \ldots, 0)$, and $c^{\prime}=\|x-y\|_{2} \leqslant c$. We will use the fact that for any orthonormal matrix $Q$, and for any $Y \sim N\left(0, \rho^{2}\right)^{d}, Y Q \sim N\left(0, \rho^{2}\right)^{d}$ : i.e. spherically symmetric Gaussian distributions are invariant under rotation. We have:

$$
\operatorname{Pr}[x+Y \in S]=\operatorname{Pr}[(x+Y) Q \in S Q]=\operatorname{Pr}[x Q+Y Q \in S Q]=\operatorname{Pr}[Y \in S Q-x Q]
$$

We want to bound:

$$
\left|\ln \left(\frac{\operatorname{Pr}[Y \in S Q-x Q]}{\operatorname{Pr}[Y \in S Q-y Q]}\right)\right|
$$

Now note that we have chosen $Q$ such that $(S Q-x Q)_{i}=(S Q-y Q)_{i}$ for all $i>1\left(\right.$ Because $(x Q)_{i}=$ $(y Q)_{i}$ for all $y>1$ ). Therefore, we have:

$$
\left|\ln \left(\frac{\operatorname{Pr}[Y \in S Q-x Q]}{\operatorname{Pr}[Y \in S Q-y Q]}\right)\right|=\left|\ln \left(\frac{\operatorname{Pr}\left[Y_{1} \in(S Q)_{1}-(x Q)_{1}\right]}{\operatorname{Pr}\left[Y_{1} \in(S Q)_{1}-(y Q)_{1}\right]}\right)\right|
$$

Note that by rotational invariance, we have: $\operatorname{Pr}\left[(z Q)_{1} \geqslant t\right]=\operatorname{Pr}\left[z_{1} \geqslant t\right]$ for any vector $z \in \mathbb{R}^{d}$, and so we are now again in the 1 -dimensional case, in which the theorem is already proven.

## B Measure concentration on the sphere

In Section 5 we used the following classical result regarding concentration of Lipschitz functions on the sphere. A proof can be found for example in Matousek's text book [Mat02].

Theorem B. 1 (Lévy's lemma). Let $f: \mathbb{S}^{d-1} \rightarrow \mathbb{R}$ be a Lipschitz function in the sense that

$$
|f(x)-f(y)| \leqslant\|x-y\|_{2}
$$

and define the median of $f$ as $\operatorname{med}(f)=\sup \left\{t \in \mathbb{R}: \operatorname{Pr}\{f \leqslant t\} \leqslant \frac{1}{2}\right\}$. Then,

$$
\operatorname{Pr}\{|f-\operatorname{med}(f)|>\varepsilon\} \leqslant 4 \exp \left(-\varepsilon^{2} d / 2\right),
$$

where probability probability and expectation are computed with respect to the uniform measure on the sphere.

## C The Netflix Data

In this section we illustrate why the data set released by Netflix satisfies the assumptions underlying Theorem 1.1. That is, the matrix is unbalanced, sparse and $C$-coherent (Definition 5.1) for very small $C$. Indeed, according to information released by Netflix, the data set has the following properties:

1. There are $x=100,480,507$ movie ratings, $m=17,770$ movies and $n=480,189$ users. In particular, the data set is very sparse in that only a $x / m n \approx 0.011$ fraction of the matrix is nonzero. Also note that $m \ll n$.
2. The most rated movie in the data set is Miss Congeniality with $t=227,715$ ratings (followed by Independence Day with 216,233). Hence, the maximum number of entries in one row is only a $t / x \approx 0.0022$ fraction of the total number of nonzero entries. Moreover, all entries of the matrix are in $\{1, \ldots, 5\}$ and thus very small numbers.

We conclude that, indeed, the Netflix matrix is sparse and the maximum norm of any row takes up only a tiny fraction of the total norm of the matrix. We further believe that these properties are likely to hold in other recommender systems. Indeed, the average number of ratings per user should be small (thus resulting in a sparse matrix), and no item should be rated almost as often as all other items taken together (thus resulting in low coherence).


[^0]:    *IBM Research Almaden. Email: mhardt@us.ibm. com
    ${ }^{\dagger}$ Department of Computer and Information Sciences, University of Pennsylvania. Email: aaroth@cis.upenn.edu

[^1]:    ${ }^{1}$ An algorithm $\mathcal{M}$ is blatantly non-private if for every database $D \in\{0,1\}^{n^{\prime}}$ it is possible to reconstruct a $1-o(1)$ fraction of the entries of $D$ exactly, given only the output of the mechanism $\mathcal{M}(D)$.

[^2]:    ${ }^{2}$ See, e.g., the proof of Theorem 4.1 in [De11].

