2014 IEEE Annual Symposium on Foundations of Computer Science

Understanding Alternating Minimization for Matrix Completion

Moritz Hardt
IBM Research Almaden
San Jose, CA, USA
Email: mhardt@us.ibm.com

***Abstract*—Alternating minimization is a widely used and empirically successful heuristic for matrix completion and related low-rank optimization problems. Theoretical guarantees for alternating minimization have been hard to come by and are still poorly understood. This is in part because the heuristic is iterative and non-convex in nature. We give a new algorithm based on alternating minimization that provably recovers an unknown low-rank matrix from a random subsample of its entries under a standard incoherence assumption. Our results reduce the sample size requirements of the alternating minimization approach by at least a quartic factor in the rank and the condition number of the unknown matrix. These improvements apply even if the matrix is only close to low-rank in the Frobenius norm. Our algorithm runs in nearly linear time in the dimension of the matrix and, in a broad range of parameters, gives the strongest sample bounds among all subquadratic time algorithms that we are aware of.**

Underlying our work is a new robust convergence analysis of the well-known Power Method for computing the dominant singular vectors of a matrix. This viewpoint leads to a conceptually simple understanding of alternating minimization. In addition, we contribute a new technique for controlling the coherence of intermediate solutions arising in iterative algorithms based on a smoothed analysis of the QR factorization. These techniques may be of interest beyond their application here.

I. Introduction

Alternating minimization is an empirically successful heuristic for the matrix completion problem in which the goal is to recover an unknown low-rank matrix from a subsample of its entries. Matrix completion has received a tremendous amount of attention over the past few years due to its fundamental role as an optimization problem and its applicability in number of areas including collaborative filtering and quantum tomography. Alternating minimization has been used early on in the context of matrix completion [1], [2] and continues to play an important role in practical approaches to the problem. The approach also formed an important component in the winning submission for the Netflix Prize [3].

Given a subset Ω of entries drawn from an unknown matrix A, Alternating minimization starts from a poor approximation $X_0 Y_0^\top$ to the target matrix and gradually improves the approximation quality by fixing one of the factors and minimizing a certain objective over the other factor. Here, X_0, Y_0 each have k columns where k is the target rank of the factorization. The least squares objective is the typical choice. In this case, at step ℓ we solve the optimization problem

$$X_\ell = \arg\min_X \sum_{(i,j)\in\Omega} \left[A_{ij} - (XY_{\ell-1}^\top)_{ij}\right]^2 .$$

This optimization step is then repeated with X_ℓ fixed in order to determine Y_ℓ as

$$Y_\ell = \arg\min_X \sum_{(i,j)\in\Omega} \left[A_{ij} - (X_\ell Y^\top)_{ij}\right]^2 .$$

Separating the factors X_ℓ and Y_ℓ is what makes the optimization step tractable. This basic update step is usually combined with an initialization procedure for finding X_0, Y_0, as well as methods for modifying intermediate solutions, e.g., truncating large entries. More than a specific algorithm we think of alternating minimization as a framework for solving a non-convex low-rank optimization problem.

A major advantage of alternating minimization over alternatives is that each update is computationally cheap and has a small memory footprint as we only need to keep track of $2k$ vectors. In contrast, the *nuclear norm* approach to matrix completion [4], [5], [6] requires solving a semidefinite program. The advantage of the nuclear norm approach is that it comes with strong theoretical guarantees under certain assumptions on the unknown matrix and the subsample of its entries. There are two (by now standard) assumptions which together imply that nuclear norm minimization succeeds. The first is that the subsample Ω includes each entry of A uniformly at random with probability p. The second assumption is that the first k singular vectors of A span an *incoherent* subspace. Informally coherence measures the correlation of the subspace with any standard basis vector. More formally, the coherence of a k-dimensional subspace of $\mathbb{R}^n$ is at most μ if the projection of each standard basis vector has norm at most $\sqrt{\mu k/n}$. The space spanned by the top k singular space of various random matrix models typically satisfies this property with small μ. But also real-world matrices tend to exhibit incoherence when k is reasonably small.

Theoretical results on matrix completion primarily apply to the *nuclear norm* semidefinite program which is prohibitive to execute on realistic instance sizes. There certainly has been progress on practical algorithms for solving related convex programs [7], [8], [9], [10], [11]. Unfortunately, these algorithms are not known to achieve the same type of recovery guarantees attained by exact nuclear norm

0272-5428/14 $31.00 © 2014 IEEE
DOI 10.1109/FOCS.2014.75

minimization. This raises the important question if there are fast algorithms for matrix completion that come with guarantees on the required sample size comparable to those achieved by nuclear norm minimization. In this work we make progress on this problem by proving strong sample complexity bounds for alternating minimization. Along the way our work helps to give a theoretical justification and understanding for why alternating minimization works.

A. Our results

We begin with our result on the *exact* matrix completion problem where the goal is to recover an unknown rank k matrix M from a subsample Ω of its entries where each entry is included independently with probability p. Here and in the following we will always assume that $M = U\Lambda U^\top$ is a symmetric $n \times n$ matrix with singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_k$. Our result generalizes straightforwardly to rectangular matrices as we will see.

Our algorithm will output a pair of matrices (X, Y) where X is an orthonormal $n \times k$ matrix that approximates U in the strong sense that $\|(I - UU^\top)X\| \leqslant \varepsilon$. Moreover, the matrix $XY^\top$ is close to M in Frobenius norm. To state the theorem we formally define the coherence of U as $\mu(U) \stackrel{\text{def}}{=} \max_{i\in[n]}(n/k)\|e_i^\top U\|_2^2$ where e_i is the i-th standard basis vector.

Theorem I.1. *Given a sample of size $\widetilde{O}(pn^2)$ drawn from an unknown $n \times n$ matrix $M = U\Lambda U^\top$ of rank k by including each entry with probability p, our algorithm outputs with high probability a pair of matrices (X, Y) such that $\|(I - UU^\top)X\| \leqslant \varepsilon$ and $\|M - XY^\top\|_F \leqslant \varepsilon\|M\|_F$ provided that*

$$pn \geqslant k(k + \log(n/\varepsilon))\mu(U)\left(\|M\|_F/\sigma_k\right)^2 . \qquad (1)$$

Our result should be compared with two remarkable recent works by Jain, Netrapalli and Sanghavi [12] and Keshavan [13] who gave rigorous sample complexity bounds for alternating minimization. [12] obtained the bound $pn \geqslant k^7(\sigma_1/\sigma_k)^6\mu(U)^2$ and Keshavan obtained the incomparable bound $pn \geqslant k(\sigma_1/\sigma_k)^8\mu(U)$ that is superior when the matrix has small *condition number* σ_1/σ_k. Since $\|M\|_F \leqslant \sqrt{k}\sigma_1$ our result improves upon [12] by at least a factor of $k^4(\sigma_1/\sigma_k)^4\mu(U)$ and improves on [13] as soon as $\sigma_1/\sigma_k \gg k^{1/3}$. The improvement is larger when $\|M\|_F = O(\sigma_1)$ which we expect if the singular values decay rapidly.

Theorem I.1 is a special case of Theorem VI.1. We remark that the number of least squares update steps is bounded by $O(\log(n/\varepsilon)\log n)$. The cost of performing these update steps is up to a logarithmic factor what dominates the worst-case running time of our algorithm. It can be seen that the least squares problem can be solved in time $O\left(nk^3 + |\Omega| \cdot k\right)$ which is is linear in $n + |\Omega|$ and polynomial in k. The number of update steps enters the sample complexity since we assume (as in previous work) that fresh samples are used in each step. However, the logarithmic dependence on $1/\varepsilon$ guarantees exponentially fast convergence and allows us to obtain any inverse polynomial error with only a constant factor overhead in sample complexity.

Noisy matrix completion: In noisy matrix completion the unknown matrix is only close to low-rank, typically in Frobenius norm. Our results apply to any matrix of the form $A = M + N$, where $M = U\Lambda U^\top$ is a matrix of rank k as before and $N = (I - UU^\top)A$ is the part of A not captured by the dominant singular vectors. Here, N can be an arbitrary deterministic matrix that satisfies the following constraints:

$$\max_{i\in[n]}\|e_i^\top N\|^2 \leqslant \frac{\mu_N}{n}\cdot\sigma_k^2 \quad \text{and} \quad \max_{ij\in[n]}|N_{ij}| \leqslant \frac{\mu_N}{n}\cdot\|A\|_F . \qquad (2)$$

Here, e_i denotes the i-th standard basis vector so that $\|e_i^\top N\|$ is the Euclidean norm of the i-th row of N. The conditions state no entry and no row of N should be too large compared to the Frobenius norm of N. We can think of the parameter μ_N as an analog to the coherence parameter $\mu(U)$ that we saw earlier. Since N could be close to full rank, denoting by V the space spanned by the columns of N, the parameter $\mu(V)$ is no longer meaningful. If the rank of V is k, then our assumptions roughly reduce to what is implied by requiring $\mu(V) \leqslant \mu_N$. From here on we let $\mu^* = \max\{\mu(U), \mu_N, \log n\}$.

Theorem I.2. *Given a sample of size $\widetilde{O}(pn^2)$ drawn from an unknown $n \times n$ matrix $A = M + N$ where $M = U\Lambda U^\top$ has rank k and $N = (I - UU^\top)A$ satisfies* (**??**)*, our algorithm outputs with high probability (X, Y) such that $\|(I - UU^\top)X\| \leqslant \varepsilon$ and $\|M - XY^\top\|_F \leqslant \varepsilon\|A\|_F$ provided that*

$$pn \geqslant \frac{k(k + \log(n/\varepsilon))\mu^*}{(1 - \sigma_{k+1}/\sigma_k)^5}\left(\frac{\|M\|_F + \|N\|_F/\varepsilon}{\sigma_k}\right)^2 .$$

The theorem is a strict generalization of the noise-free case which we recover by setting $N = 0$ in which case the separation parameter $\gamma_k := 1 - \sigma_{k+1}/\sigma_k$ is equal to 1. The result follows from Theorem VI.1 that gives a somewhat stronger sample complexity bound. Compared to our noise-free bound, there are two new parameters that enter the sample complexity. The first one is the separation parameter γ_k. The second is the quantity $\|N\|_F/\varepsilon$. To interpret this quantity, suppose that that A has a good low-rank approximation in Frobenius norm, formally, $\|N\|_F \leqslant \varepsilon\|A\|_F$ for $\varepsilon \leqslant 1/2$, then it must also be the case that $\|N\|_F/\varepsilon \leqslant 2\|M\|_F$. Our algorithm then finds a good rank k approximation with at most $\widetilde{O}(k^3(\sigma_1/\sigma_k)^2\mu^* n)$ samples assuming $\gamma_k = \Omega(1)$. Hence, assuming that A has a good rank k approximation in Frobenius norm and that σ_k and σ_{k+1} are well-separated, our bound recovers the noise-free bound up to a constant factor.

Note that if we're only interested in the second error bound $\|M - XY^\top\|_F \leqslant \varepsilon\|M\|_F + \|N\|_F$, we we can eliminate the dependence on the condition number in the sample complexity entirely. The reason is that any singular

value smaller than $\varepsilon\sigma_1/k$ can be treated as part of the noise matrix. Assuming the condition number is at least k to begin with we can always find two singular values that have separation at least $\Omega(k)$. This ensures that the sample requirement is polynomial in k without any dependence on the condition number and gives us the following corollary.

Corollary I.3. *Under the assumptions of Theorem I.2, if* $\sigma_1 \geqslant k\sigma_k/\varepsilon$, *then we can find* X, Y *such that* $\|M - XY^\top\|_F \leqslant \varepsilon\|A\|_F$ *provided that* $pn \geqslant \mathrm{poly}(k)\mu^*$.

The previous corollary is remarkable, because small error in Frobenius norm is the most common error measure in the literature on matrix completion. The result shows that in this error measure, there is no dependence on the condition number. The result is tight for $k = O(1)$ up to constant factors even information-theoretically as we will discuss below.

The approach of Jain et al. was adapted to the noisy setting by Gunasekar et al. [14] showing roughly same sample complexity in the noisy setting under some assumptions on the noise matrix. We achieve the same improvements over [14] as we did compared to [12] in the noise-free case. Moreover, our assumptions in (**??**) are substantially weaker than the assumption of [14]. The latter work required the largest entry of N in absolute value to be bounded by $O(\sigma_k/n\sqrt{k})$. This directly implies that each row of N has norm at most $O(\sigma_k/\sqrt{kn})$ and that $\|N\|_F \leqslant O(\sigma_k/\sqrt{k})$. Moreover under this assumption we would have $\gamma_k \geqslant 1 - o_k(1)$. Keshavan's result [13] also applies to the noisy setting, but it requires $\|N\| \leqslant (\sigma_k/\sigma_1)^3$ and $\max_i \|e_i^\top N\| \leqslant \sqrt{\mu(U)k/n}\|N\|$. In particular this bound does not allow $\|N\|_F$ to grow with $\|M\|_F$. Since neither result allows arbitrarily small singular value separation, we cannot use these results to eliminate the dependence on the condition number as is possible using our technique.

Remark on required sample complexity and assumptions: It is known that information-theoretically $\Omega(k\mu(U)n)$ measurements are necessary to recover the unknown matrix [6] and this bound is achieved (up to log-factors) by the nuclear norm semidefinite program. Compared with the information-theoretic optimum our bound suffers a factor $O(k(\|M\|_F/\sigma_k)^2)$ loss. While we do not know if this loss is necessary, there is a natural barrier. If we denote by $P_\Omega(A)$ the matrix in which all unobserved entries are 0 and the others are scaled by $1/p$, then $\Omega(k\mu(U)(\|M\|_F/\sigma_k)^2 n)$ samples are necessary to ensure that $P_\Omega(A)$ preserves the k-th singular value to within constant relative error. Formally, $\|P_\Omega(A) - A\|_2 \leqslant 0.1\sigma_k$. While this is not a necessary requirement for alternating least squares, it represents the current bottleneck for finding a good initial matrix.

It is also known that without an incoherence assumption the matrix completion problem can be ill-posed and recovery becomes infeasible even information-theoretically [6]. Moreover, even on incoherent matrices it was recently shown that already the exact matrix completion problem remains computationally hard to approximate in a strong sense [15]. This shows that additional assumptions are needed beyond incoherence to make the problem tractable.

II. Proof overview and techniques

Robust convergence of subspace iteration: An important observation of [12] is that the update rule in alternating minimization can be analyzed as a noisy update step of the well known *power method* for computing eigenvectors, also called *subspace iteration* when applied to multiple vectors simultaneously. The noise term that arises depends on the sampling error induced by the subsample of the entries. We further develop this point of view by giving a new robust convergence analysis of the power method.

To illustrate the technique, consider a model of numerical linear algebra in which an input matrix A can only be accessed through noisy matrix vector products of the form $Ax + g$, where x is a chosen vector and g is a possibly adversarial noise term. Our goal is to compute the dominant singular vectors $u_1, \ldots, u_k$ of the matrix A. Subspace iteration starts with an initial guess, an orthonormal matrix $X_0 \in \mathbb{R}^{n\times k}$ typically chosen at random. The algorithm then repeatedly computes $Y_\ell = AX_{\ell-1} + G_\ell$, followed by an orthonormalization step in order to obtain X_ℓ from Y_ℓ. Here, G_ℓ is the noise variable added to the computation.

Theorem III.8 characterizes the convergence behavior of this general algorithm. An important component of our analysis is the choice of a suitable potential function that decreases at each step. Here we make use of the tangent of the *largest principal angle* between the subspace U spanned by the first k singular vectors of the input matrix and the k-dimensional space spanned by the columns of the iterate X_ℓ. Principal angles are a very useful tool in numerical analysis that we briefly recap in Section III. Our analysis shows that the algorithm essentially converges at the rate of $(\sigma_{k+1} + \Delta)/(\sigma_k - \Delta)$ for some $\Delta \ll \sigma_k$ under suitable conditions on the noise matrix G_ℓ.

Least squares update: The least squares update works as follows:

$$Y_\ell = \arg\min_Y \|P_\Omega(A - X_{\ell-1}Y^\top)\|_F^2 . \tag{3}$$

Since we can focus on symmetric matrices without loss of generality, there is no need for an alternating update in which the left and right factor are flipped. We therefore drop the term "alternating". We can express the optimal Y_ℓ as $Y_\ell = AX_{\ell-1} + G_\ell$ using gradient information about the least squares objective. The error term G_ℓ has an intriguing property. Its norm $\|G_\ell\|$ depends on the quantity $\|V^\top X_{\ell-1}\|$ which coincides with the sine of the largest principal angle between U and $X_{\ell-1}$. This property ensures that as the algorithm begins to converge the norm of the error term starts to diminish. Near exact recovery is now possible (assuming the matrix has rank at most k). A novelty in our approach

is that we obtain strong bounds on $\|G_\ell\|$ by computing $O(\log n)$ independent copies of Y_ℓ (using fresh samples) and taking the componentwise median of the resulting matrices. The resulting procedure called MEDIANLS is analyzed in Section IV.

A difficulty with iterating the least squares update in general is that it is unclear how well it converges from a random initial matrix X_0. In our analysis we therefore use an initialization procedure that finds a matrix X_0 that satisfies $\left\|V^\top X_0\right\| \leqslant 1/4$. Our initialization procedure is based on (approximately) computing the first k singular vectors of $P_\Omega(A)$. To rule out large entries in the vectors we truncate the resulting vectors. While this general approach is standard, our truncation procedure first applies a random rotation to the vectors that leads to a tighter analysis than the naive approach.

Smooth orthonormalization: A key novelty in our approach is the way we argue about the coherence of each iterate X_ℓ. Ideally, we would like to argue that $\mu(X_\ell) = O(\mu^*)$. A direct approach would be to argue that X_ℓ was obtained from Y_ℓ using the QR-factorization and so $X_\ell = Y_\ell R^{-1}$ for some invertible R. This gives the bound $\|e_i^\top X_\ell\| \leqslant \|e_i^\top Y_\ell\| \cdot \|R^{-1}\|$ that unfortunately is quite lossy and leads to a dependence on the condition number.

We avoid this problem using an idea that's closely related to the *smoothed analysis* of the QR-factorization. Sankar, Spielman and Teng [16] showed that while the perturbation stability of QR can be quadratic, it is constant after adding a sufficiently large amount of Gaussian noise. In the context of smoothed analysis this is usually interpreted as saying that there are "few bad inputs" for the QR factorization. In our context, the matrix Y_ℓ is already the outcome of a noisy operation $Y_\ell = AX_{\ell-1} + G_\ell$ and so there is no harm in actually adding a Gaussian noise matrix H_ℓ to Y_ℓ provided that the norm of that matrix is no larger than that of G_ℓ. Roughly speaking, this will allow us to argue that there is no dependence on the condition number when applying the QR-factorization to Y_ℓ. There are some important complications. The magnitude of Y_ℓ may be too large to apply the smoothed analysis argument directly to Y_ℓ. Instead we observe that the columns of X_ℓ are contained in the range S of the $n \times 2k$ matrix $[U \mid (NX_{\ell-1} + G_\ell + H_\ell)]$. This is because $Y_\ell = AX_{\ell-1} + G_\ell + H_\ell$ and $AX_{\ell-1} = MX_{\ell-1} + NX_{\ell-1}$ where $M = U\Lambda U^\top$ and $N = (I - UU^\top)A$. Since S has dimension at most $2k$ it suffices to argue that this space has small coherence. Moreover we can choose H_ℓ to be roughly on the same order as $NX_{\ell-1}$ and G_ℓ so that the smoothed analysis argument leads to an excellent bound bound on the smallest singular value of $NX_{\ell-1} + G_\ell + H_\ell$. To prove that the coherence is small we need to exhibit a basis for S. This requires us to argue about the related matrix $(I - UU^\top)(NX_{\ell-1} + G + H_\ell)$ since we need to orthonormalize the last k vectors against the first when constructing a basis. Another minor complication is that we don't know the magnitude of G_ℓ so we need to find the right scaling of H_ℓ on the fly. We call the resulting procedure that SMOOTHQR and analyze its guarantees in Section V.

Putting things together: The final algorithm that we analyze is quite simple to describe as shown in Figure 1. The algorithm makes use of an initialization procedure INITIALIZE that we defer to Section VII. In Section VI we prove our main theorem. At a high-level, the theorem is proved by induction. The main inductive hypothesis is that the coherence of the ℓ-th solution X_ℓ is small, i.e., bounded in terms of the coherence parameter μ^*. Given that the the coherence is small we can control the magnitude of the noise term $G_{\ell+1}$ using matrix concentration inequalities. Given that $G_{\ell+1}$ is small in spectral norm, our results on the noisy power method show that the algorithm makes progress towards convergence. To ensure that the inductive hypothesis continues to hold we use our analysis of the smooth orthonormalization.

The generalization of our result to rectangular matrices follows from a standard "dilation" argument available in the full version. The description of the algorithm also uses a helper function called SPLIT that's used to split the subsample into independent pieces of roughly equal size while preserving the distributional assumption that our theorems use.

Input: Observed set of indices $\Omega \subseteq [n] \times [n]$ of an unknown symmetric matrix $A \in \mathbb{R}^{n \times n}$ with entries $P_\Omega(A)$, number of iterations $L \in \mathbb{N}$, error parameter $\varepsilon > 0$, target dimension k, coherence parameter μ.

Algorithm SALTLS$(P_\Omega(A), \Omega, L, k, \varepsilon, \mu)$:

1) $(\Omega_0, \Omega') \leftarrow$ SPLIT$(\Omega, 2)$, $(\Omega_1, \ldots, \Omega_L) \leftarrow$ SPLIT(Ω', L)
2) $X_0 \leftarrow$ INITIALIZE$(P_{\Omega_0}(A), \Omega_0, k, \mu)$
3) For $\ell = 1$ to L:
   a) $Y_\ell \leftarrow$ MEDIANLS$(P_{\Omega_\ell}(A), \Omega_\ell, X_{\ell-1}, L, k)$
   b) $X_\ell \leftarrow$ SMOOTHQR$(Y_\ell, \varepsilon, \mu)$

Output: Pair of matrices (X_{L-1}, Y_L)

Figure 1: Smoothed alternating least squares (SALTLS)

A. Further discussion of related work

There is a vast literature on the topic that we cannot completely survey here. Most closely related is the work of Jain et al. [12] that suggested the idea of thinking of alternating least squares as a noisy update step in the Power Method. Our approach takes inspiration from this work by analyzing least squares using the noisy power method. However, our analysis is substantially different in both how convergence and low coherence is argued. The approach of Keshavan [13] uses a rather different argument.

As an alternative to the nuclear norm approach, Keshavan, Montanari and Oh [17], [18] present two approaches, a

spectral approach and an algorithm called OPTSPACE. The spectral approach roughly corresponds to our initialization procedure and gives similar guarantees. OPTSPACE requires a stronger incoherence assumption, has larger sample complexity in terms of the condition number, namely $(\sigma_1/\sigma_k)^6$, and requires optimizing over the Grassmanian manifold. However, the requirement on N achieved by OPTSPACE can be weaker than ours in the noisy setting. In the exact case, our algorithm has a much faster convergence rate (logarithmic dependence on $1/\varepsilon$ rather than polynomial).

There are a number of fast algorithms for matrix completion based on either (stochastic) gradient descent [19] or (online) Frank-Wolfe [9], [20]. These algorithms generally minimize squared loss on the *observed* entries subject to a nuclear norm constraint and in general do not produce a matrix that is close to the true unknown matrix on all entries. In contrast, our algorithm guarantees convergence *in domain*, that is, to the unknown matrix itself. Moreover, our dependence on the error is logarithmic whereas in these algorithms it is polynomial.

Privacy-preserving spectral analysis: Our work is also closely related to a line of work on differentially private singular vector computation [21], [22], [23]. These papers each consider algorithms based on the power method where noise is injected to achieve a privacy guarantee called Differential Privacy. Hardt and Roth [21], [22], [23] observed that incoherence could be used to obtain improved guarantees. This requires controlling the coherence of the iterates produced by the noisy power method which leads to similar problems as the ones faced here. What's simpler in the privacy setting is that the noise term is typically Gaussian leading to a cleaner analysis. Our work uses a similar convergence analysis for noisy subspace iteration that was used in a concurrent work by the author [22].

B. Preliminaries and Notation

We denote by $A^\top$ the transpose of a matrix (or vector) A. We use the notation $x \gtrsim y$ do denote that the relation $x \geqslant Cy$ holds for a sufficiently large absolute constant $C > 0$ independent of x and y. We let $\mathcal{R}(A)$ denote the range of the matrix A.

Definition II.1 (Coherence)**.** The *μ-coherence* of a k-dimensional subspace U of $\mathbb{R}^n$ is defined as $\mu(U) \stackrel{\text{def}}{=} \max_{i\in[n]} \frac{n}{k} \|P_U e_i\|_2^2$, where e_i denotes the i-th standard basis vector.

III. ROBUST CONVERGENCE OF SUBSPACE ITERATION

Figure 2 presents our basic template algorithm. The algorithm is identical to the standard subspace iteration algorithm except that in each iteration ℓ, the computation is perturbed by a matrix G_ℓ. The matrix G_ℓ can be adversarially and adaptively chosen in each round. We will analyze under which conditions on the perturbation we can expect the algorithm to converge rapidly.

Input: Matrix $A \in \mathbb{R}^{n\times n}$, number of iterations $L \in \mathbb{N}$, target dimension k

1) Let $X_0 \in \mathbb{R}^{n\times k}$ be an orthonormal matrix.
2) For $\ell = 1$ to L:
   a) Let $G_\ell \in \mathbb{R}^{n\times k}$ be a perturbation.
   b) $Y_\ell \leftarrow AX_{\ell-1} + G_\ell$
   c) $X_\ell \leftarrow \mathrm{GS}(Y_\ell)$

Output: Matrix X_L with k orthonormal columns

Figure 2: Noisy Subspace Iteration (NSI)

Principal angles are a useful tool in analyzing the convergence behavior of numerical eigenvalue methods. We will use the largest principal angle between two subspaces as a potential function in our convergence analysis.

Definition III.1. Let $X, Y \in \mathbb{R}^{n\times k}$ be orthonormal bases for subspaces $\mathcal{X}, \mathcal{Y}$, respectively. Then, the sine of the *largest principal angle* between $\mathcal{X}$ and $\mathcal{Y}$ is defined as $\sin\theta(\mathcal{X},\mathcal{Y}) \stackrel{\text{def}}{=} \left\|(I - XX^\top)Y\right\|$.

We use some standard properties of the largest principal angle.

Proposition III.2 ([24])**.** *Let $\mathcal{X}, \mathcal{Y}, X, Y$ be as in Definition III.1 and let $X_\perp$ be an orthonormal basis for the orthogonal complement of $\mathcal{X}$. Then, we have* $\cos\theta(\mathcal{X},\mathcal{Y}) = \sigma_k(X^\top Y)$. *and assuming $X^\top Y$ is invertible,* $\tan\theta(\mathcal{X},\mathcal{Y}) = \|X_\perp^\top Y(X^\top Y)^{-1}\|$

From here on we will always assume that A has the spectral decomposition $A = U\Lambda_U U^\top + V\Lambda_V V^\top$, where $U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{n\times(n-k)}$ corresponding to the first k and last $n-k$ eigenvectors respectively. We will let $\sigma_1 \geqslant \ldots \geqslant \sigma_n$ denote the singular values of A which coincide with the absolute eigenvalues of A sorted in non-increasing order.

Our convergence analysis tracks the tangent of the largest principal angles between the subspaces $\mathcal{R}(U)$ and $\mathcal{R}(X_\ell)$. The next lemma shows a natural condition under which the potential decreases multiplicatively in step ℓ. We think of this lemma as a local convergence guarantee, since it assumes that the cosine of the largest principal angle between $\mathcal{R}(U)$ and $\mathcal{R}(X_{\ell-1})$ is already lower bounded by a constant.

Lemma III.3 (One Step Local Convergence)**.** *Let $\ell \in \{1,\ldots,L\}$. Assume that* $\cos\theta_k(U, X_{\ell-1}) \geqslant \frac{1}{2} > \frac{\|U^\top G_\ell\|}{\sigma_k}$. *Then,*

$$\tan\theta(U, X_\ell) \leqslant \tan\theta(U, X_{\ell-1}) \cdot \frac{\sigma_{k+1} + \frac{2\|V^\top G_\ell\|}{\tan\theta(U,X_{\ell-1})}}{\sigma_k - 2\|U^\top G_\ell\|} .$$

The next lemma essentially follows by iterating the previous lemma.

Lemma III.4 (Local Convergence)**.** *Let* $0 \leqslant \varepsilon \leqslant 1/4$. *Let* $\Delta = \max_{1 \leqslant \ell \leqslant L} \|G_\ell\|$ *and* $\gamma_k = 1 - \sigma_{k+1}/\sigma_k$. *Assume that* $\|V^\top X_0\| \leqslant 1/4$ *and* $\sigma_k \geqslant 8\Delta/\gamma_k\varepsilon$. *Then,*

$$\left\|V^\top X_L\right\| \leqslant \max\left\{\varepsilon, 2 \cdot \left\|V^\top X_0\right\| \cdot \exp(-\gamma_k L/2)\right\} .$$

Proof: Our first claim shows that once the potential function is below ε at step $\ell - 1$, it cannot increase beyond ε.

Claim III.5. *Let* $\ell \geqslant 1$. *Suppose that* $\tan\theta(U, X_{\ell-1}) \leqslant \varepsilon$. *Then,* $\tan\theta(U, X_\ell) \leqslant \varepsilon$.

Proof: By our assumption, $\cos\theta_k(U, X_{\ell-1}) \geqslant \sqrt{1-\varepsilon^2} \geqslant 15/16$. Together with the lower bound on σ_k, the assumptions for Lemma III.3 are met. Hence, using our assumptions,

$$\tan\theta(U, X_\ell) \leqslant \frac{(1-\gamma_k)\sigma_k\varepsilon + 2\Delta}{\sigma_k - 2\Delta} \leqslant \varepsilon .$$

■

Our second claim shows that if the potential is at least ε at step $\ell - 1$, it will decrease by a factor $1 - \gamma_k/2$.

Claim III.6. *Let* $\ell \geqslant 1$ *Suppose that* $\tan\theta(U, X_{\ell-1}) \in [\varepsilon, 1/2]$. *Then,*

$$\tan\theta(U, X_\ell) \leqslant (1 - \gamma_k/2)\tan\theta(U, X_{\ell-1}) .$$

Proof: Using the assumption of the claim we have $\cos\theta(U, X_{\ell-1}) \geqslant \frac{1}{\tan\theta(U, X_{\ell-1})} \geqslant 1/2 > \Delta/\sigma_k$. We can therefore apply Lemma III.3 to conclude

$$\begin{aligned}\tan\theta(U, X_\ell) &\leqslant \tan\theta(U, X_{\ell-1}) \cdot \frac{(1-\gamma_k)\sigma_k + 2\Delta}{\sigma_k - 2\Delta} \\ &\leqslant \tan\theta(U, X_{\ell-1}) \cdot \frac{(1-\gamma_k)(1+\gamma_k/4)}{1-\gamma_k/4} \\ &\leqslant \tan\theta(U, X_{\ell-1})(1-\gamma_k/2)\end{aligned}$$

■

The two previous claims together imply that

$$\tan\theta(U, X_L) \leqslant \max\left\{\tan\theta(U, X_0)(1-\gamma_k/2)^L, \varepsilon\right\} ,$$

provided that $\tan\theta(U, X_0) \leqslant 1/2$. This is the case since we assumed that $\sin\theta(U, X_0) \leqslant 1/4$. Note that $(1-\gamma_k/2)^L \leqslant \exp(-\gamma_k L/2)$. It remains to observe that $\left\|V^\top X_L\right\| \leqslant \tan\theta(U, X_L)$ and further $\tan\theta(U, X_0) \leqslant 2\left\|V^\top X_0\right\|$ by our assumption on X_0. ■

In our application later on the error terms $\|G_\ell\|$ decrease as ℓ increases and the algorithm starts to converge. We need a convergence bound for this type of shrinking error. The next definition expresses a condition on G_ℓ that allows for a useful convergence bound.

Definition III.7 (Admissible)**.** Let $\gamma_k = 1 - \sigma_{k+1}/\sigma_k$. We say that the pair of matrices $(X_{\ell-1}, G_\ell)$ is *ε-admissible for* NSI if

$$\|G_\ell\| \leqslant \frac{1}{32}\gamma_k\sigma_k\|V^\top X_{\ell-1}\| + \frac{\varepsilon}{32}\gamma_k\sigma_k. \tag{4}$$

We say that a family of matrices $\{(X_{\ell-1}, G_\ell)\}_{\ell=1}^L$ is *ε-admissible for* NSI if each member of the set is ε-admissible. We will use the notation $\{G_\ell\}$ as a shorthand for $\{(X_{\ell-1}, G_\ell)\}_{\ell=1}^L$.

We have the following convergence guarantee for admissible noise matrices.

Theorem III.8. *Let* $\gamma_k = 1 - \sigma_{k+1}/\sigma_k$. *Let* $\varepsilon \leqslant 1/2$. *Assume that the family of noise matrices* $\{G_\ell\}$ *is* $(\varepsilon/2)$*-admissible for* NSI *and that* $\left\|V^\top X_0\right\| \leqslant 1/4$. *Then, we have* $\left\|V^\top X_L\right\| \leqslant \varepsilon$ *for any* $L \geqslant 4\gamma_k^{-1}\log(1/\varepsilon)$.

Proof: We prove by induction that for every $t \geqslant 0$ after $L_t = 4t\gamma_k^{-1}$ steps, we have $\left\|V^\top X_{L_t}\right\| \leqslant \max\left\{2^{-(t+1)}, \varepsilon\right\}$. The base case ($t = 0$) follows directly from the assumption that $\left\|V^\top X_0\right\| \leqslant 1/4$. We turn to the inductive step. By induction hypothesis, we have $\left\|V^\top X_{L_t}\right\| \leqslant \max\left\{2^{-(t+1)}, \varepsilon\right\}$. We apply Lemma III.4 with "$X_0 = X_{L_t}$" and error parameter $\max\left\{2^{-t+2}, \varepsilon\right\}$ and $L = L_{t+1} - L_t$. The conditions of the lemma are satisfied as can be easily checked using the assumption that $\{G_\ell\}$ is $\varepsilon/2$-admissible. Using the fact that $L_{t+1} - L_t = 4/\gamma_k$, the conclusion of the lemma gives $\left\|V^\top X_{L_{t+1}}\right\| \leqslant \max\left\{\varepsilon, 2 \cdot \max\left\{\varepsilon, 2^{-(t+1)}\right\}\exp\left(-\frac{\gamma_k(L_{t+1}-L_t)}{2}\right)\right\} \leqslant \max\left\{\varepsilon, 2^{-(t+2)}\right\}$. ■

IV. Least squares update rule

Input: Target dimension k, observed set of indices $\Omega \subseteq [n] \times [n]$ of an unknown symmetric matrix $A \in \mathbb{R}^{n \times n}$ with entries $P_\Omega(A)$, orthonormal matrix $X \in \mathbb{R}^{n \times k}$.
Algorithm $\mathrm{LS}(P_\Omega(A), \Omega, X, L, k)$:
$Y \leftarrow \arg\min_{Y \in \mathbb{R}^{n \times k}} \left\|P_\Omega(A - XY^\top)\right\|_F^2$
Output: Pair of matrices (X, Y)

Figure 3: Least squares update

Figure 4 describes the least squares update step specialized to the case of a symmetric matrix. Our goal is to express this update step as an update step of the form $Y = AX + G$ so that we may apply our analysis of noisy subspace iteration. This syntactic transformation is explained in Section IV-A followed by a bound on the norm of the error term G in Section IV-B.

A. From alternating least squares to noisy subspace iteration

The optimizer Y satisfies a set of linear equations that we derive from the gradient of the objective function.

Lemma IV.1 (Optimality Condition)**.** *Let* $P_i \colon \mathbb{R}^n \to \mathbb{R}^n$ *be the linear projection onto the coordinates in* $\Omega_i = \{j \colon (i,j) \in \Omega\}$ *scaled by* $p^{-1} = n^2/(\mathbb{E}\,|\Omega|)$, *i.e.,* $P_i =$

$p^{-1}\sum_{j\in\Omega_i} e_j e_j^\top$. *Further, define the matrix* $B_i \in \mathbb{R}^{k\times k}$ *as* $B_i = X^\top P_i X$ *and assume that* B_i *is invertible. Then, for every* $i \in [n]$*, the i-th row of* Y *satisfies* $e_i^\top Y = e_i^\top A P_i X B_i^{-1}$.

The assumption that B_i is invertible is essentially without loss of generality. Indeed, we will later see that B_i is invertible (and in fact close to the identity matrix) with very high probability. We can now express the least squares update as $Y = AX + G$ where we derive some useful expression for G.

Lemma IV.2. *Let* $E = (I - XX^\top)U$*. We have* $Y = AX + G$ *where* $G = G^M + G^N$ *and the matrices* G^M *and* G^N *satisfy for each row* $i \in [n]$ *if* B_i *is invertible then*

$$e_i^\top G^M = e_i^\top U \Lambda_U E^\top P_i X B_i^{-1}$$
$$e_i^\top G^N = e_i^\top (N P_i X B_i^{-1} - NX)\,.$$

B. Deviation bounds for the least squares update

In this section we analyze the norm of the error term G from the previous section. More specifically, we prove a bound on the norm of each row of G. Our bound uses the fact that the matrix E appearing in the expression for the error term satisfies $\|E\| = \left\|V^\top X\right\|$. This gives us a bound in terms of the quantity $\left\|V^\top X\right\|$.

Lemma IV.3. *Let* $\delta \in (0,1)$*. Assume that each entry is included in* Ω *independently with probability*

$$p \gtrsim \frac{k\mu(X)\log n}{\delta^2 n}\,. \tag{5}$$

Then, for every $i \in [n]$*,* $\mathbb{P}\left\{\left\|e_i^\top G\right\| > \delta\cdot\left(\|e_i^\top M\|\cdot\left\|V^\top X\right\| + \|e_i^\top N\|\right)\right\} \leqslant \frac{1}{5}$.

C. Median least squares update

Given the previous error bound we can achieve a strong concentration bound by taking the component-wise median of multiple independent samples of the error term.

Lemma IV.4. *Let* $G_1,\ldots,G_t$ *be i.i.d. copies of* G*. Let* $\overline{G} = \mathrm{median}(G_1,\ldots,G_t)$ *be the component-wise median of* $G_1,\ldots,G_t$ *and assume* p *satisfies* (**??**)*. Then, for every* $i \in [n]$*,* $\mathbb{P}\left\{\left\|e_i^\top \overline{G}\right\| > \delta\left(\|e_i^\top M\|\cdot\left\|V^\top X\right\| + \|e_i^\top N\|\right)\right\} \leqslant \exp(-\Omega(t))$.

We can now conclude a strong concentration bound for the median of multiple independent solutions to the least squares minimization step. This way we can obtain the desired error bound for all rows simultaneously. This leads to the following extension of the least squares update rule.

Lemma IV.5. *Let* Ω *be a sample in which each entry is included independently with probability* $p \gtrsim \frac{k\mu(X)\log^2 n}{\delta^2 n}$ *. Let* $Y \leftarrow \textsc{MedianLS}(P_\Omega(A),\Omega,X,L,k)$*. Then, we have with probability* $1 - 1/n^3$ *that* $\overline{Y} = AX + \overline{G}$ *and* $\overline{G}$ *satisfies for every* $i \in [n]$ *the bound* $\left\|e_i^\top \overline{G}\right\| \leqslant \delta\left\|e_i^\top M\right\|\cdot\left\|V^\top X\right\| + \delta\left\|e_i^\top N\right\|$.

Input: Target dimension k, observed set of indices $\Omega \subseteq [n]\times[n]$ of an unknown symmetric matrix $A \in \mathbb{R}^{n\times n}$ with entries $P_\Omega(A)$, orthonormal matrix $X \in \mathbb{R}^{n\times k}$.
Algorithm $\textsc{MedianLS}(P_\Omega(A),\Omega,X,L,k)$:
1) $(\Omega_1,\ldots,\Omega_t) \leftarrow \textsc{Split}(\Omega,t)$ for $t = O(\log n)$.
2) $Y_i \leftarrow \mathrm{LS}(P_{\Omega_i}(A),\Omega_i,X,L,k)$
Output: Pair of matrices $(X, \mathrm{median}(Y_1,\ldots,Y_t))$

Figure 4: Median least squares update

V. Incoherence via smooth QR factorization

As part of our analysis of alternating minimization we need to show that the intermediate solutions X_ℓ have small coherence. For this purpose we propose an idea inspired by Smoothed Analysis of the QR factorization [16]. The problem with applying the QR factorization directly to Y_ℓ is that Y_ℓ might be ill-conditioned. This can lead to a matrix X_ℓ (via QR-factorization) that has large coordinates and whose coherence is therefore no longer as small as we desire. A naive bound on the condition number of Y_ℓ would lead to a large loss in sample complexity. What we show instead is that a small Gaussian perturbation to Y_ℓ leads to a sufficiently well-conditioned matrix $\widetilde{Y}_\ell = Y_\ell + H_\ell$. Orthonormalizing $\widetilde{Y}_\ell$ now leads to a matrix of small coherence. Intuitively, since the computation of Y_ℓ is already noisy the additional noise term has little effect so long as its norm is bounded by that of G_ℓ. Since we don't know the norm of G_ℓ, we have to search for the right noise parameter using a simple binary search. We call the resulting procedure SmoothQR and describe in in Figure 5.

Input: Matrix $Y \in \mathbb{R}^{n\times k}$, parameters $\mu,\varepsilon > 0$.
Algorithm $\textsc{SmoothQR}(Y,\varepsilon,\mu)$:
1) $X \leftarrow \mathrm{QR}(Y), H \leftarrow 0, \sigma \leftarrow \varepsilon\|Y\|/n$.
2) While $\mu(X) > \mu$ and $\sigma \leqslant \|Y\|$:
   a) $X \leftarrow \mathrm{GS}(Y + H)$ where $H \sim \mathrm{N}(0,\sigma^2/n)^{n\times k}$
   b) $\sigma \leftarrow 2\sigma$
Output: Pair of matrices (X,H)

Figure 5: Smooth Orthonormalization (SmoothQR)

To analyze the algorithm we begin with a lemma that analyzes the smallest singular value under a Gaussian perturbation. What makes the analysis easier is the fact that the matrices we're interested in are rectangular. The square case was considered in [16] and requires more involved arguments.

Lemma V.1. *Let* $G \in \mathbb{R}^{n\times k}$ *be any matrix with* $\|G\| \leqslant 1$ *and let* V *be a* $n-k$ *dimensional subspace with orthogonal projection* P_V*. Let* $H \sim \mathrm{N}(0,\tau^2/n)^{n\times k}$ *be a random Gaussian matrix. Assume* $k = o(n/\log n)$*. Then, with probability* $1 - \exp(-\Omega(n))$*, we have* $\sigma_k\left(P_V(G+H)\right) \geqslant \Omega(\tau)$.

The proof follows from standard concentration arguments and is contained in the full version. To use this lemma in our context we'll introduce a variant of μ-coherence that applies to matrices rather than subspaces.

Definition V.2 (ρ-coherence)**.** Given a matrix $G \in \mathbb{R}^{n\times k}$ we let $\rho(G) \stackrel{\text{def}}{=} \frac{n}{k}\max_{i\in[n]} \|e_i^\top G\|^2$.

The next lemma is our main technical tool in this section. It shows that adding a Gaussian noise term leads to a bound on the coherence after applying the QR-factorization.

Lemma V.3. *Let $k = o(n/\log n)$ and $\tau \in (0,1)$. Let $U \in \mathbb{R}^{n\times k}$ be an orthonormal matrix. Let $G \in \mathbb{R}^{n\times k}$ be a matrix such that $\|G\| \leqslant 1$. Let $H \sim \mathrm{N}(0,\tau^2/n)^{k\times n}$ be a random Gaussian matrix. Then, with probability $1-\exp(-\Omega(n)) - n^{-5}$, there is an orthonormal matrix $Q \in \mathbb{R}^{n\times 2k}$ such that:*

1) *$\mathcal{R}(Q) = \mathcal{R}([U \mid G+H])$ where $\mathcal{R}(Q)$ denotes the range of Q,*
2) *$\mu(Q) \leqslant O\left(\frac{1}{\tau^2}\cdot(\rho(G)+\mu(U)+\log n)\right)$.*

Proof: First note that $\mathcal{R}([U \mid G+H]) = \mathcal{R}([U \mid (I-UU^\top)(G+H)])$. Let $B = (I-UU^\top)(G+H)$. Applying the QR-factorization to $[U \mid B]$, we can find two orthonormal matrices $Q_1, Q_2 \in \mathbb{R}^{n\times k}$ such that have that $[Q_1 \mid Q_2] = [U \mid BR^{-1}]$ where $R \in \mathbb{R}^{k\times k}$. That is $Q_1 = U$ since U is already orthonormal. Moreover, the columns of B are orthogonal to U and therefore we can apply the QR-factorization to U and B independently. We can now apply Lemma V.1 to the $(n-k)$-dimensional subspace $U^\perp$ and the matrix $G+H$. It follows that with probability $1-\exp(-\Omega(n))$, we have $\sigma_k(B) \geqslant \Omega(\tau)$. Assume that this event occurs.

Also, observe that $\sigma_k(B) = \sigma_k(R)$. The second condition is now easy to verify $\frac{n}{k}\left\|e_i^\top Q\right\|^2 = \frac{n}{k}\left\|e_i^\top U\right\|^2 + \frac{n}{k}\left\|e_i^\top BR^{-1}\right\|^2 = \mu(U) + \frac{n}{k}\left\|e_i^\top BR^{-1}\right\|^2$. On the other hand, $\frac{n}{k}\left\|e_i^\top BR^{-1}\right\|^2 \leqslant \frac{n}{k}\left\|e_i^\top B\right\|^2\left\|R^{-1}\right\|^2 \leqslant O\left(\frac{n}{k\tau^2}\left\|e_i^\top B\right\|^2\right)$, where we used the fact that $\left\|R^{-1}\right\| = 1/\sigma_k(R) = O(1/\tau)$. Moreover, $\frac{n}{k}\left\|e_i^\top B\right\|^2 \leqslant 2\frac{n}{k}\left\|e_i^\top(I-UU^\top)G\right\|^2 + 2\rho((I-UU^\top)H) \leqslant 2\rho(G) + 2\rho(UU^\top G) + 2\rho((I-UU^\top)H)$. Finally, $\rho(UU^\top G) \leqslant \mu(U)\|U^\top G\|^2 \leqslant \mu(U)$ and, by Lemma V.4 below, we have $\rho((I-UU^\top)H) \leqslant O(\log n)$ with probability $1-1/n^5$. The lemma follows with a union bound over the failure probabilities. ∎

Lemma V.4. *Let P be the projection onto an $(n-k)$-dimensional subspace. Let $H \sim \mathrm{N}(0,1/n)^{n\times k}$. Then, $\rho(PH) \leqslant O(\log n)$ with probability $1-1/n^5$.*

The next lemma states that when SMOOTHQR is invoked on an input of the form $AX+G$ with suitable parameters, the algorithm outputs a matrix of the form $X' = \mathrm{QR}(AX+G+H)$ whose coherence is bounded in terms of $\mu(U)$ and $\rho(G)$ and moreover H satisfies a bound on its norm. The lemma also permits to trade-off the amount of additional noise introduced with the resulting coherence parameter.

Lemma V.5. *Let $\tau > 0$ and assume $k = o(n/\log n)$. There is an absolute constant $C_{V.5} > 0$ such that the following claim holds. Let $G \in \mathbb{R}^{n\times k}$. Let $X \in \mathbb{R}^{n\times k}$ be an orthonormal matrix such that $\nu \geqslant \max\{\|G\|, \|NX\|\}$. Assume that*

$$\mu \geqslant \frac{C_{V.5}}{\tau^2}\left(\mu(U) + \frac{\rho(G)+\rho(NX)}{\nu^2} + \log n\right).$$

Then, for every $\varepsilon \leqslant \tau\nu$ satisfying $\log(n/\varepsilon) \leqslant n$ and every $\mu \leqslant n$, we have with probability $1-O(n^{-4})$, the algorithm SMOOTHQR$(AX+G,\varepsilon,\mu)$ terminates in $O(\log(n/\varepsilon))$ steps and outputs (X',H) such that $\mu(X') \leqslant \mu$ and where H satisfies $\|H\| \leqslant \tau\nu$.

VI. Main theorem

The total sample complexity we achieve is the sum of two terms. The first one is used by the initialization step that we discuss in Section VII. The second term specifies the sample requirements for iterating the least squares algorithm. It therefore makes sense to define the following two quantities: $p_{\text{init}} = \frac{k^2\mu^*\|A\|_F^2\log n}{\gamma_k^2\sigma_k^2 n}$ and $p_{\text{LS}} = \frac{k\mu^*(\|M\|_F^2+\|N\|_F^2/\varepsilon^2)\log(n/\varepsilon)\log^2 n}{\gamma_k^5\sigma_k^2 n}$. While the first term has a quadratic dependence on k it does not depend on ε at all and it has single logarithmic factor. The second term features a linear dependence on k. Our main theorem shows that if the sampling probability is larger than the sum of these two terms, the algorithm converges rapidly to the true unknown matrix.

Theorem VI.1 (Main)**.** *Let $k, \varepsilon > 0$. Let $A = M+N$ be a symmetric $n\times n$ matrix where M is a matrix of rank k with the spectral decomposition $M = U\Lambda_U U^\top$ and $N = (I-UU^\top)A = V\Lambda_V V^\top$ satisfies (**??**). Let $\gamma_k = 1-\sigma_{k+1}/\sigma_k$ where σ_k is the smallest singular value of M and σ_{k+1} is the largest singular value of N.*

Then, there are parameters $\mu = \Theta(\gamma_k^{-2}k(\mu^*+\log n))$ and $L = \Theta(\gamma_k^{-1}\log(n/\varepsilon))$ such that the output (X,Y) of SALTLS$(P_\Omega(A),\Omega,k,L,\varepsilon,\mu)$ *satisfies $\|(I-UU^\top)X_L\| \leqslant \varepsilon$ with probability $9/10$.*

Before we prove the theorem in Section VI-A, we will state an immediate corollary that gives bounds on the reconstruction error in the Frobenius norm.

Corollary VI.2 (Reconstruction error)**.** *Under the assumptions of Theorem VI.1, we have that the output (X,Y) of* SALTLS *satisfies $\left\|M-XY^\top\right\|_F \leqslant \varepsilon\|A\|_F$ with probability $9/10$.*

A. Proof of Theorem VI.1

Proof: We first apply Theorem VII.1 (shown below) to conclude that with probability $19/20$, the initial matrix X_0 satisfies $\left\|V^\top X_0\right\| \leqslant 1/4$ and $\mu(X_0) \leqslant 32\mu(U)\log n$.

Assume that this event occurs. Our goal is now to apply Theorem III.8. Consider the sequence of matrices $\left\{(X_{\ell-1}, \widetilde{G}_\ell)\right\}_{\ell=1}^{L}$ obtained by the execution of SALTLS starting from X_0 and letting $\widetilde{G}_\ell = G_\ell + H_\ell$ where G_ℓ is the error term corresponding to the ℓ-step of MEDIANLS, and H_ℓ is the error term introduced by the application of SMOOTHQR at step ℓ. To apply Theorem III.8, we need to show that this sequence of matrices is $(\varepsilon/2)$-admissible for NSI with probability $19/20$. The theorem then directly gives that $\|V^\top X_L\| \leqslant \varepsilon$ and this would conclude our proof by summing up the error probabilities.

Let $\tau = \frac{\gamma_k}{128}$ and $\widehat{\mu} = \frac{C_{V.5}}{\tau^2}\left(20\mu^* + \log n\right)$. Let μ be any number satisfying $\mu \geqslant \widehat{\mu}$. Since $\widehat{\mu} = \Theta(\gamma_k^{-2} k(\mu^* + \log n))$, this satisfies the requirement in the theorem. We prove that with probability $9/20$, the following three claims hold:

1) $\{(X_{\ell-1}, G_\ell)\}_{\ell=1}^{L}$ is $(\varepsilon/4)$-admissible,
2) $\{(X_{\ell-1}, H_\ell)\}_{\ell=1}^{L}$ is $(\varepsilon/4)$-admissible,
3) for all $\ell \in \{0, \ldots, L-1\}$, we have $\mu(X_\ell) \leqslant \mu$.

This implies the claim that we want using a triangle inequality since $\widetilde{G}_\ell = G_\ell + H_\ell$.

The proof of these three claims is by mutual induction. For $\ell = 0$, we only need to check the third claim which follows form the fact that X_0 satisfies the coherence bound. Now assume that all three claims hold at step $\ell - 1$, we will argue that the with probability $1 - n/100$, all three claims hold at step ℓ. Since $L \leqslant n$, this is sufficient.

The first claim follows from Lemma IV.4 using the induction hypothesis that $\mu(X_{\ell-1}) \leqslant \widehat{\mu}$. Specifically, we apply the lemma with $\delta = c\min\{\gamma_k\sigma_k/\|M\|_F, \varepsilon\gamma_k\sigma_k/\|N\|_F\}$ for sufficiently small constant $c > 0$. The lemma requires the lower bound $p \gtrsim \frac{k\mu^* \log^2 n}{\delta^2 n}$. We can easily verify that the right hand side is a factor $L = \Theta(\gamma_k^{-1}\log(n/\varepsilon))$ smaller than what is provided by the assumption of the theorem. This is because new samples are used in each of the L steps so that we need to divide the given bound by L. Lemma IV.4 now gives with probability $1 - 1/n^3$ the upper bound $\|G_\ell\|_F \leqslant \frac{1}{4}\left(\frac{1}{32}\gamma_k\sigma_k \left\|V^\top X_{\ell-1}\right\| + \frac{\varepsilon}{32}\gamma_k\sigma_k\right)$. In particular, this satisfies the definition of $\varepsilon/4$-admissibility. We proceed assuming that this event occurs as the error probability is small enough to ignore.

The remaining two claims follow from Lemma V.5. We will apply the lemma to $AX_\ell + G_\ell$ with $\nu = \sigma_k(\left\|V^\top X_{\ell-1}\right\| + \varepsilon)$ and τ as above. Note that $\|NX_{\ell-1}\| \leqslant \sigma_k \left\|V^\top X_{\ell-1}\right\|$. Hence we have $\nu \geqslant \max\{\|G_\ell\|, \|NX_{\ell-1}\|\}$ as required by the lemma. The lemma also requires a lower bound μ. To satisfy the lower bound we invoke Lemma VI.3 showing that with probability $1 - 1/n^2$, we have $\frac{1}{\nu^2}\left(\rho(G) + \rho(NX)\right) \leqslant 10\mu^*$. We remark that this is the lemma that uses the assumption on N provided by (**??**). Again we assume this event occurs. In this case we have $\mu \geqslant \widehat{\mu} = \frac{C_{V.5}}{\tau^2}\left(20\mu^* + \log n\right)$ and so we see that μ satisfies the requirement of Lemma V.5. It follows that SMOOTHQR produces with probability $1 - 1/n^4$ a matrix H_ℓ such that $\|H_\ell\| \leqslant \tau\nu \leqslant \frac{\gamma_k\nu}{128} \leqslant \frac{1}{4}\left(\frac{1}{32}\gamma_k\sigma_k \left\|V^\top X_{\ell-1}\right\|_F + \frac{\varepsilon}{32}\gamma_k\sigma_k\right)$. In particular, H_ℓ satisfies the requirement of $(\varepsilon/4)$-admissibility. Moreover, the lemma gives that $\mu(X_\ell) \leqslant \mu$. This shows that also the second and third claim of our inductive claim continue to hold. All error probabilities we incurred were $o(1/n)$ and we can sum up the error probabilities over all $L \leqslant n$ steps to concludes the proof of the theorem. ∎

The following technical lemma was needed in the proof of Theorem VI.1.

Lemma VI.3. *Under the assumptions of Theorem VI.1, we have for every $\ell \in [L]$ and $\nu = \frac{\sigma_k}{32}(\left\|V^\top X_{\ell-1}\right\| + \varepsilon)$ with probability $1 - 1/n^2$, $\frac{1}{\nu^2}\left(\rho(G) + \rho(NX_{\ell-1})\right) \leqslant 3\mu^*$.*

We also needed a procedure $\textsc{Split}(\Omega, t)$ that takes a sample Ω and splits it into t independent samples that preserve the distributional assumption that we need. The next lemma is standard.

Lemma VI.4. *There is a procedure $\textsc{Split}(\Omega, t)$ such that if Ω is sampled by including each element independently with probability p, then $\textsc{Split}(\Omega, t)$ outputs independent random variables $\Omega_1, \ldots, \Omega_t$ such that each set Ω_i includes each element independently with probability $p_i \geqslant p/t$.*

VII. FINDING A GOOD STARTING POINT

Figure 6 describes an algorithm that computes the top k singular vectors of $P_\Omega(A)$ and truncates them in order to ensure incoherence. The algorithm serves as a fast initialization procedure for our main algorithm. This general approach is relatively standard in the literature. However, our truncation argument differs from previous approaches. Specifically, we use a random orthonormal transformation to spread out the entries of the singular vectors before truncation. This leads to a tighter bound on the coherence.

Input: Target dimension k, observed set of indices $\Omega \subseteq [n] \times [n]$ of an unknown symmetric matrix $A \in \mathbb{R}^{n \times n}$ with entries $P_\Omega(A)$, coherence parameter $\mu \in \mathbb{R}$.

Algorithm $\textsc{Initialize}(P_\Omega(A), \Omega, k, \mu)$:

1) Compute the first k singular vectors $W \in \mathbb{R}^{n \times k}$ of $P_\Omega(A)$.
2) $\widetilde{W} \leftarrow WO$ where $O \in \mathbb{R}^{k \times k}$ is a random orthonormal matrix.
3) $T \leftarrow \mathcal{T}_{\mu'}(\widetilde{W})$ with $\mu' = \sqrt{8\mu\log(n)/n}$ where $\mathcal{T}_c$ replaces each entry of its input with the nearest number in the interval $[-c, c]$.
4) $X \leftarrow QR(T)$

Output: Orthonormal matrix $X \in \mathbb{R}^{n \times k}$.

Figure 6: Initialization Procedure (INITIALIZE)

Theorem VII.1 (Initialization)**.** *Let $A = M + N$ be a symmetric $n \times n$ matrix where M is a matrix of rank k with the spectral decomposition $M = U\Lambda_U U^\top$ and $N = (I - UU^\top)A$ satisfies* **(??)***. Assume that each entry is included in Ω independently probability*

$$p \geqslant \frac{Ck(k\mu(U) + \mu_N)(\|A\|_F/\gamma_k\sigma_k)^2 \log n}{n} \qquad (6)$$

for a sufficiently large constant $C > 0$. Then, the algorithm INITIALIZE *returns an orthonormal matrix $X \in \mathbb{R}^{n\times k}$ such that with probability $9/10$, $\|V^\top X\|_F \leqslant 1/4$ and $\mu(X) \leqslant 32\mu(U)\log n$.*

ACKNOWLEDGMENTS

Thanks to David Gleich, Prateek Jain, Jonathan Kelner, Raghu Meka, Ankur Moitra, Nikhil Srivastava, and Mary Wootters for many helpful discussions. We thank the Simons Institute for Theoretical Computer Science at Berkeley, where some of this research was done.

REFERENCES

[1] R. M. Bell and Y. Koren, "Scalable collaborative filtering with jointly derived neighborhood interpolation weights," in *ICDM*. IEEE Computer Society, 2007, pp. 43–52.

[2] J. P. Haldar and D. Hernando, "Rank-constrained solutions to linear matrix equations using powerfactorization," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 584–587, 2009.

[3] Y. Koren, R. M. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[4] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computional Mathematics*, vol. 9, pp. 717–772, December 2009.

[5] B. Recht, "A simpler approach to matrix completion," *Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[6] E. J. Candès and T. Tao, "The power of convex relaxation: near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[7] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th ICML*. ACM, 2009, p. 58.

[8] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[9] M. Jaggi and M. Sulovský, "A simple algorithm for nuclear norm regularized problems," in *Proc. 27th ICML*. ACM, 2010, pp. 471–478.

[10] H. Avron, S. Kale, S. P. Kasiviswanathan, and V. Sindhwani, "Efficient and practical stochastic subgradient descent for nuclear norm regularization," in *Proc. 29th ICML*. ACM, 2012.

[11] C.-J. Hsieh and P. A. Olsen, "Nuclear norm minimization via active subspace selection," in *Proc. 31st ICML*. ACM, 2014.

[12] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Symposium on Theory of Computing (STOC)*. ACM, 2013, pp. 665–674.

[13] R. H. Keshavan, "Efficient algorithms for collaborative filtering," Ph.D. dissertation, Stanford University, 2012.

[14] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh, "Noisy matrix completion using alternating minimization," in *Proc. ECML PKDD*. Springer, 2013, pp. 194–209.

[15] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz, "Computational limits for matrix completion," *CoRR*, vol. abs/1402.2331, 2014.

[16] A. Sankar, D. A. Spielman, and S.-H. Teng, "Smoothed analysis of the condition numbers and growth factors of matrices," *SIAM J. Matrix Analysis Applications*, vol. 28, no. 2, pp. 446–476, 2006.

[17] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[18] ——, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 11, pp. 2057–2078, 2010.

[19] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Math. Program. Comput.*, vol. 5, no. 2, pp. 201–226, 2013.

[20] E. Hazan and S. Kale, "Projection-free online learning," in *ICML*. ACM, 2012.

[21] M. Hardt and A. Roth, "Beating randomized response on incoherent matrices," in *Proc. 44th Symposium on Theory of Computing (STOC)*. ACM, 2012, pp. 1255–1268.

[22] ——, "Beyond worst-case analysis in private singular vector computation," in *Proc. 45th Symposium on Theory of Computing (STOC)*. ACM, 2013.

[23] M. Hardt, "Robust subspace iteration and privacy-preserving spectral analysis," *arXiv*, vol. 1311:2495, 2013.

[24] P. Zhu and A. V. Knyazev, "Angles between subspaces and their tangents," *Arxiv preprint arXiv:1209.0523*, 2012.