

# Fast Matrix Completion Without the Condition Number

**Moritz Hardt**

*IBM Research Almaden*

MHARDT@US.IBM.COM

**Mary Wootters**

*University of Michigan*

WOOTTERS@UMICH.EDU

## Abstract

We give the first algorithm for Matrix Completion that achieves running time and sample complexity that is polynomial in the rank of the unknown target matrix, *linear* in the dimension of the matrix, and *logarithmic* in the condition number of the matrix. To the best of our knowledge, all previous algorithms either incurred a quadratic dependence on the condition number of the unknown matrix or a quadratic dependence on the dimension of the matrix. Our algorithm is based on a novel extension of Alternating Minimization which we show has theoretical guarantees under standard assumptions even in the presence of noise.

## 1. Introduction

Matrix Completion is the problem of recovering an unknown real-valued low-rank matrix from a possibly noisy subsample of its entries. The problem has received a tremendous amount of attention in signal processing and machine learning partly due to its wide applicability to recommender systems. A beautiful line of work showed that a particular convex program—known as nuclear norm minimization—achieves strong recovery guarantees under certain reasonable feasibility assumptions [Candès and Recht \(2009\)](#); [Candès and Tao \(2010\)](#); [Recht et al. \(2010\)](#); [Recht \(2011\)](#). Nuclear norm minimization boils down to solving a semidefinite program and therefore can be solved in polynomial time in the dimension of the matrix. Unfortunately, the approach is not immediately practical due to the large polynomial dependence on the dimension of the matrix. An ongoing research effort aims to design large-scale algorithms for nuclear norm minimization [Ji and Ye \(2009\)](#); [Mazumder et al. \(2010\)](#); [Jaggi and Suvovský \(2010\)](#); [Avron et al. \(2012\)](#); [Hsieh and Olsen \(2014\)](#). Such fast solvers, generally speaking, involve heuristics that improve empirical performance but may no longer preserve the strong theoretical guarantees of the nuclear norm approach.

A successful scalable algorithmic alternative to Nuclear Norm Minimization is based on Alternating Minimization [Bell and Koren \(2007\)](#); [Haldar and Hernando \(2009\)](#); [Koren et al. \(2009\)](#). Alternating Minimization aims to recover the unknown low-rank matrix by alternately optimizing over one of two factors in a purported low-rank decomposition. Each update is a simple least squares regression problem that can be solved very efficiently. As pointed out in [Hsieh and Olsen \(2014\)](#), even state of the art nuclear norm solvers often cannot compete with Alternating Minimization with regards to scalability. A shortcoming of Alternating Minimization is that formal guarantees are less developed than for Nuclear Norm Minimization. Only recently has there been progress in this direction [Keshavan \(2012\)](#); [Jain et al. \(2013\)](#); [Gunasekar et al. \(2013\)](#); [Hardt \(2013b\)](#).

Unfortunately, despite this recent progress all known convergence bounds for Alternating Minimization have at least a quadratic dependence on the *condition number* of the matrix. Here, the

condition number refers to the ratio of the first to the  $k$ -th singular value of the matrix, where  $k$  is the target rank of the decomposition. This dependence on the condition number can be a serious shortcoming. After all, Matrix Completion rests on the assumption that the unknown matrix is approximately low-rank and hence we should expect its singular values to decay rapidly. Indeed, strongly decaying singular values are a typical feature of large real-world matrices.

The dependence on the condition number in Alternating Minimization is not a mere artifact of the analysis. It arises naturally with the use of the Singular Value Decomposition (SVD). Alternating Minimization is typically initialized with a decomposition based on a truncated SVD of the partial input matrix. Such an approach must incur a polynomial dependence on the condition number. Many other approaches also crucially rely on the SVD as a sub-routine, e.g., Jain et al. (2010); Keshavan et al. (2010a,b), as well as most fast solvers for the nuclear norm. In fact, there appears to be a kind of dichotomy in the current literature on Matrix Completion: either the algorithm is *not fast* and has at least a quadratic dependence on the dimension of the matrix in its running time, or it is *not well-conditioned* and has at least a quadratic dependence on the condition number in the sample complexity. We emphasize that here we focus on formal guarantees rather than observed empirical performance, which may be better on certain instances.

**Main Problem:** Is there a sub-quadratic time algorithm for Matrix Completion with a sub-linear dependence on the condition number?

In fact, eliminating the polynomial dependence on the condition number for Alternating Minimization was posed as an open problem by Jain, Netrapalli and Sanghavi Jain et al. (2013).

In this work, we resolve the question in the affirmative. Specifically, we design a new variant of Alternating Minimization that achieves a *logarithmic* dependence on the condition number while retaining the fast running time of the standard Alternating Minimization framework. This is an exponential improvement in the condition number compared with all subquadratic time algorithms for Matrix Completion that we are aware of. Our algorithm works even in the noisy Matrix Completion setting and under standard assumptions—specifically, the same assumptions that support theoretical results for the nuclear norm. That is, we assume that the first  $k$  singular vector of the matrix span an incoherent subspace and that each entry of the matrix is revealed independently with a certain probability. While strong, these assumptions led to an interesting theory of Matrix Completion and have become a de facto standard when comparing theoretical guarantees.

## 1.1. Our Results

For the sake of exposition we begin by explaining our results in the *exact* Matrix Completion setting, even though our results here are a direct consequence of our theorem for the noisy case. In the exact problem the goal is to recover an unknown rank  $k$  matrix  $M$  from a subsample  $\Omega \subset [n] \times [n]$  of its entries where each entry is included independently with probability  $p$ . We assume that the unknown matrix  $M = U\Lambda U^T$  is a symmetric  $n \times n$  matrix with nonzero singular values  $\sigma_1 \geq \dots \geq \sigma_k > 0$ . Following Hardt (2013b), our result generalizes straightforwardly to rectangular matrices. To state our result we need to define the *coherence* of the subspace spanned by  $U$ . Intuitively, the coherence controls how large the projection is of any standard basis vector onto the space spanned by  $U$ . Formally, for a  $n \times k$  matrix  $U$  with orthonormal columns, the coherence of  $U$  is  $\mu(U) = \max_{i \in [n]} \frac{n}{k} \|e_i^T U\|_2^2$ , where  $e_1, \dots, e_n$  is the standard basis of  $\mathbb{R}^n$ .

We show that our algorithm outputs a low-rank factorization  $XY^T$  such that with high probability  $\|M - XY^T\|^2 \leq \varepsilon \|M\|$  provided that the expected size of  $\Omega$  satisfies

$$pn^2 = O\left(nk^c \mu(U)^2 \log\left(\frac{\sigma_1}{\sigma_k}\right) \log^2\left(\frac{n}{\varepsilon}\right)\right). \quad (1)$$

Here, the exponent  $c > 0$  is bounded by an absolute constant. While we did not focus on minimizing the exponent, our results imply that the value of  $c$  can be chosen smaller if the singular values of  $M$  are well-separated. The formal statement follows from Theorem 1. Note that the dependence on the error  $\varepsilon$  is only poly-logarithmic. This linear convergence rate makes near exact recovery feasible with a small number of steps. We also show that the running time of our algorithm is bounded by  $\tilde{O}(\text{poly}(k)pn^2)$ , which is nearly linear in the number of revealed entries except for a polynomial overhead in  $k$ . For small values of  $k$  and  $\mu(U)$ , the total running time is nearly linear in  $n$ .

**Noisy Matrix Completion.** We now discuss our more general result that applies to the noisy or robust Matrix Completion problem. Here, the unknown matrix is only close to low-rank, typically in Frobenius norm. Our results apply to any matrix of the form

$$A = M + N = U\Lambda U^T + N, \quad (2)$$

where  $M = U\Lambda U^T$  is a matrix of rank  $k$  as before and  $N = (I - UU^T)A$  is the part of  $A$  not captured by the dominant singular vectors. We note that  $N$  can be an arbitrary deterministic matrix. The assumption that we will make is that  $N$  satisfies the following incoherence conditions:

$$\max_{i \in [n]} \|e_i^T N\|_2^2 \leq \frac{\mu_N}{n} \cdot \min\{\|N\|_F^2, \sigma_k^2\} \quad \text{and} \quad \max_{i,j \in N} |N_{ij}| \leq \frac{\mu_N \|N\|_F}{n}. \quad (3)$$

Recall that  $e_i$  denotes the  $i$ -th standard basis vector so that  $\|e_i^T N\|_2$  is the Euclidean norm of the  $i$ -th row of  $N$ . The conditions state no entry of  $N$  should be too large compared to the norm of the corresponding row in  $N$ , and no row of  $N$  should be too large compared to  $\sigma_k$ . Our bounds will be in terms of a combined coherence parameter  $\mu^*$  satisfying

$$\mu^* \geq \max\{\mu(U), \mu_N\}. \quad (4)$$

We show that our algorithm outputs a rank  $k$  factorization  $XY^T$  such that with high probability  $\|A - XY^T\| \leq \varepsilon \|M\| + (1 + o(1))\|N\|$ , where  $\|\cdot\|$  denotes the spectral norm. It follows from our argument that we can have the same guarantee in Frobenius norm as well. To achieve the above bound we show in Theorem 1 that it is sufficient to have an expected sample size

$$pn^2 = O\left(n \cdot \text{poly}(k/\gamma_k)(\mu^*)^2 \log\left(\frac{\sigma_1}{\sigma_k}\right) \left(\log^2\left(\frac{n}{\varepsilon}\right) + \left(\frac{\|N\|_F}{\varepsilon \|M\|_F}\right)^2\right)\right). \quad (5)$$

Here,  $\gamma_k = 1 - \sigma_{k+1}/\sigma_k$  indicates the separation between the singular values  $\sigma_k$  and  $\sigma_{k+1}$ . The theorem is a strict generalization of the noise-free case, which we recover by setting  $N = 0$  and hence  $\gamma_k = 1$ . Compared to our noise-free bound above, there are two new parameters that enter the sample complexity:  $\gamma_k$  and  $\|N\|_F/\varepsilon\|M\|_F$ . To interpret this difference, suppose that  $A$  has a good low-rank approximation in Frobenius norm and that  $\sigma_k$  and  $\sigma_{k+1}$  are well-separated: formally,  $\|N\|_F \leq \varepsilon\|A\|_F$  for  $\varepsilon \leq 1/2$  and  $\gamma_k = \Omega(1)$ . Then the bound above implies that our algorithm then finds a good rank  $k$  approximation with at most  $O(\text{poly}(k) \log(\sigma_1/\sigma_k)(\mu^*)^2 n)$  samples, recovering the noise-free bound up to a constant factor.

For an extended discussion of related work see Section 2.1. We proceed in the next section with a detailed proof overview and a description of our notation.

## 2. Overview

As the proof of our main theorem is somewhat complex we will begin with an extensive informal overview of the argument. In order to understand our main algorithm, it is necessary to understand the basic Alternating Minimization algorithm first.

**Alternating Least Squares.** Given a subsample  $\Omega$  of entries drawn from an unknown matrix  $A$ , Alternating Minimization starts from a poor approximation  $X_0 Y_0^T$  to the target matrix and iteratively refines the approximation 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  $\ell$  we solve the optimization problem

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

This optimization step is then repeated with  $X_\ell$  fixed in order to determine  $Y_\ell$ . Since we assume without loss of generality that  $A$  is symmetric these two steps can be combined. Previous work exploited that Alternating Least Squares update can be interpreted as a noisy power method update step. That is,  $Y_\ell = AX_{\ell-1} + G_\ell$  for a noise matrix  $G_\ell$ . In this view, the convergence of the algorithm can be controlled by  $\|G_\ell\|$ , the spectral norm of the noise matrix. To a rough approximation, this spectral norm initially behaves like  $O(\sigma_1/\sqrt{pn})$ , ignoring factors of  $k$  and  $\mu(U)$ . Since we would like to discover singular vectors corresponding to singular values of magnitude  $\sigma_k$ , we need that the error term satisfies  $\|G_\ell\| \ll \sigma_k$ : otherwise we cannot rule out that the noise term wipes out any correlation between  $X$  and the  $k$ -th singular vector. In order to achieve this, we would need to set  $pn = O((\sigma_1/\sigma_k)^2)$  and this is where a quadratic dependence on the condition number arises. This is not the only reason for this dependence: Alternating Minimization seems to exhibit a linear convergence rate only once  $X_\ell$  is already “somewhat close” to the desired subspace  $U$ . This is why typically the algorithm is initialized with a truncated SVD of the matrix  $P_\Omega(A)$  where  $P_\Omega$  is the projection onto the subsample  $\Omega$ . We again face the issue that  $\|A - P_\Omega(A)\|$  behaves roughly like  $O(\sigma_1/\sqrt{pn})$  and so we run into the same problem here as well.

A natural idea to fix these problems is a *deflation* approach. If it so happens that  $\sigma_1 \gg \sigma_k$ , then there must be an  $r < k$  such that  $\sigma_1 \approx \sigma_r \gg \sigma_k$ . In this case, we can try to first run Alternating Minimization with  $r$  vectors instead of  $k$  vectors. This results in a rank  $r$  factorization  $XY^T$ . We then subtract this matrix off of the original matrix and continue with  $A' = A - XY^T$ . This approach was in particular suggested by Jain et al. [Jain et al. \(2013\)](#) to eliminate the condition number dependence. Unfortunately, as we will see next, this approach runs into serious issues.

**Why standard deflation does not work.** Given any algorithm NOISYMC for noisy matrix completion, whose performance depends on the condition number of  $A$ , we may hope to use NOISYMC in a black-box way to obtain a deflation-based algorithm which does not depend on the condition number, as follows. Suppose that we know that the spectrum of  $A$  comes in blocks,

$$\sigma_1 = \sigma_2 = \dots = \sigma_{r_1} \gg \sigma_{r_1+1} = \sigma_{r_1+2} = \dots = \sigma_{r_2} \gg \sigma_{r_2+1} = \dots$$

and so on. We could imagine running NOISYMC on  $P_\Omega(A)$  with target rank  $r_1$ , to obtain an estimate  $M^{(1)}$ . Then we may run NOISYMC again on  $P_\Omega(A - M^{(1)}) = P_\Omega(A) - P_\Omega(M^{(1)})$  with target rank  $r_2 - r_1$ , to obtain  $M^{(2)}$ , and so on. At the end of the day, we would hope to approximate  $A \approx M^{(1)} + M^{(2)} + \dots$ . Because we are focusing only on a given “flat” part of the spectrum at a

time, the dependence of NOISYMC on the condition number should not matter. A major problem with this approach is that the error builds up rather quickly. More precisely, any matrix completion algorithm run on  $A$  with target rank  $r_1$  must have error on the order of  $\sigma_{r_1+1}$  since this is the spectral norm of the “noise part” that prevents the algorithm from converging further. Therefore, the matrix  $A - M^{(1)}$  might now have  $2r_1$  problematic singular vectors corresponding to relatively large singular values, namely those vectors arising from the residuals of the first  $r_1$  singular vectors, as well as those arising from the approximation error. This multiplicative blow-up makes it difficult to ensure convergence.

**Soft deflation.** The above intuition may make a “deflation”-based argument seem hopeless. We instead use an approach that looks similar to deflation but makes an important departure from it. Intuitively, our algorithm is a single execution of Alternating Minimization. However, we dynamically grow the number of vectors that Alternating Minimization maintains until we’ve reached  $k$  vectors. At that point we let the algorithm run to convergence. More precisely, the algorithm proceeds in at most  $k$  epochs. Each epoch roughly proceeds as follows:

**Inductive Hypothesis:** At the beginning of epoch  $t$ , the algorithm has a rank  $r_{t-1}$  factorization  $X_{t-1}Y_{t-1}^T$  that has converged to within error  $\sigma_{r_{t-1}+1}/100$ . At this point, the  $(r_{t-1} + 1)$ -th singular vector prevents further convergence.

**Gap finding:** What can we say about the matrix  $A_t = A - X_{t-1}Y_{t-1}^T$  at this point? We know that the first  $r_{t-1}$  singular vectors of  $A$  are removed from the top of the spectrum of  $A_t$ . Moreover, each of the remaining singular vectors in  $A$  is preserved so long as the corresponding singular value is greater than  $\sigma_{r_{t-1}+1}/10$ . This follows from perturbation bounds and we ignore a polynomial loss in  $k$  at this point. Importantly, the top of the spectrum of  $A_t$  corresponds is correlated with the next block of singular vectors in  $A$ . This motivates the next step in epoch  $t$ , which is to compute the top  $k - r_{t-1}$  singular vectors of  $A_t$  up to an approximation error of  $\sigma_{r_{t-1}+1}/10$ . Among these singular vectors we now identify a gap in singular values, that is we look for a number  $d_t$  such that  $\sigma_{r_{t-1}+d_t} \leq \sigma_{r_{t-1}+1}/2$ .

**Alternating Least Squares:** At this point we have identified a new block of  $d_t$  singular vectors and we arrange them into an orthogonal matrix  $P_t \in \mathbb{R}^{n \times d_t}$ . We can now argue that the matrix  $W = [X_{t-1} | P_t]$  is close (in principal angle) to the first  $r_t = r_{t-1} + d_t$  singular vectors of  $A$ . What this means is that  $W$  is a good initializer for the Alternating Minimization algorithm which we now run on  $W$  until it converges to a rank  $r_t$  factorization  $X_t Y_t^T$  that satisfies the induction hypothesis of the next epoch.

We call this algorithm SOFTDEFLATE. The crucial difference to the deflation approach is that we always run Alternating Minimization on a subsampling  $P_\Omega(A)$  of the original matrix  $A$ . We only ever compute a deflated matrix  $P_\Omega(A - XY^T)$  for the purpose of initializing the next epoch of the algorithm. This prevents the error accumulation present in the basic deflation approach.

This simple description glosses over many details and there are a few challenges to be overcome in order to make the idea work. For example, we have not said how to determine the appropriate “gaps”  $d_t$ . This requires a little bit of care. Indeed, these gaps might be quite small: if the (additive) gap between  $\sigma_r$  and  $\sigma_{r+1}$  is on the order of, say,  $\frac{\log^2(k)}{k} \sigma_r$ , for all  $r \leq k$ , then the condition number of the matrix may be super-polynomial in  $k$ , a price we are not willing to pay. Thus, we need to be able to identify gaps between  $\sigma_r$  and  $\sigma_{r+1}$  which are on the order of  $\sigma_r/k$ . To do this, we must make sure that our estimates of the singular values of  $A - X_{t-1}Y_{t-1}^T$  are sufficiently precise.

**Ensuring Coherence.** Another major issue that such an algorithm faces is that of coherence. As mentioned above, incoherence is a standard (and necessary) requirement of matrix completion algorithms, and so in order to pursue the strategy outlined above, we need to be sure that the estimates  $X_{t-1}$  stay incoherent. For our first “rough estimation” step, our algorithm carefully truncates (entrywise) its estimates, in order to preserve the incoherence conditions, without introducing too much error. We note that we cannot reuse the truncation analysis of Jain et al. [Jain et al. \(2013\)](#) for this step, as it incurred a dependence on the condition number. Coherence in the Alternating Minimization step is handled by the algorithm and analysis of [Hardt \(2013b\)](#), upon which we build.

## 2.1. Further Discussion of Related Work

Our work is most closely related to recent works on convergence bound for Alternating Minimization [Keshavan \(2012\)](#); [Jain et al. \(2013\)](#); [Gunasekar et al. \(2013\)](#); [Hardt \(2013a\)](#). Our bounds are in general incomparable. We achieve an exponential improvement in the condition number compared to all previous works, while losing polynomial factors in  $k$ . Our algorithm and analysis crucially builds on [Hardt \(2013b\)](#). In particular we use the version and analysis of Alternating Minimization derived in that work more or less as a black box. We note that the analyses of Alternating Minimization in other previous works would not be sufficiently strong to be used in our algorithm. In particular, the use of noise addition to ensure coherence already gets rid of one source of the condition number that all previous papers incur.

We are not aware of any fast nuclear norm solver with theoretical guarantees that do not depend polynomially on the condition number. The work of Keshavan et al. [Keshavan et al. \(2010a,b\)](#) gives another alternative to nuclear norm minimization that has theoretical guarantees. However, these bounds have a quartic dependence on the condition number. There are a number of fast algorithms for matrix completion based on either (Stochastic) Gradient Descent [Recht and Ré \(2013\)](#) or (On-line) Frank-Wolfe [Jaggi and Sulovský \(2010\)](#); [Hazan and Kale \(2012\)](#). However, the theoretical guarantees for these algorithms are typically in terms of the error on the observed entries, rather than on the error between the recovered matrix and the unknown matrix itself. Further, these algorithms typically have polynomial, rather than logarithmic, dependence on the accuracy parameter  $\varepsilon$ . Since setting  $\varepsilon \approx \sigma_k/\sigma_1$  is required in order to accurately recover the first  $k$  singular vectors of  $A$ , a polynomial dependence in  $\varepsilon$  implies a polynomial dependence on the condition number.

## 2.2. Notation

For a matrix  $A$ ,  $\|A\|$  denotes the spectral norm, and  $\|A\|_F$  the Frobenius norm. We will also use  $\|A\|_\infty = \max_{i,j} |A_{i,j}|$  to mean the entry-wise  $\ell_\infty$  norm. For a vector  $v$ ,  $\|v\|_2$  denotes the  $\ell_2$  norm. Throughout,  $C, C_0, C_1, C_2, \dots$  will denote absolute constants, and  $C$  may change from instance to instance. We also use standard asymptotic notation  $O(\cdot)$  and  $\Omega(\cdot)$ , and we occasionally use  $f \lesssim g$  (resp.  $\gtrsim$ ) to mean  $f = O(g)$  (resp.  $f = \Omega(g)$ ) to remove notational clutter. Here, the asymptotics are taken as  $k, n \rightarrow \infty$ . For a matrix  $X \in \mathbb{R}^{n \times k}$ ,  $\mathcal{R}(X)$  denotes the span of the columns of  $X$ , and  $\Pi_X$  denotes the orthogonal projection onto  $\mathcal{R}(X)$ . Similarly,  $\Pi_{X^\perp}$  denotes the projection onto  $\mathcal{R}(X)^\perp$ . For a set random  $\Omega \subset [n] \times [n]$  and a matrix  $A \in \mathbb{R}^{n \times n}$ , we define the (normalized) projection operator  $P_\Omega$  as

$$P_\Omega(A) := \frac{n^2}{\mathbb{E}|\Omega|} \sum_{(i,j) \in \Omega} A_{i,j} e_i e_j^T$$

to the be matrix  $A$ , restricted to the entries indexed by  $\Omega$  and renormalized. Our algorithm, and its proof, will involve choosing a sequence of integers  $r_1 < \dots < r_t \leq k$ , which will mark the significant “gaps” in the spectrum of  $A$ . Given such a sequence, we will decompose  $A$  as

$$A = M^{(\leq t)} + N_t = M^{(1)} + M^{(2)} + \dots + M^{(t)} + N_t, \quad (6)$$

where  $M^{(\leq t)}$  has the spectral decomposition  $M^{(\leq t)} = U^{(\leq t)} \Lambda_{(\leq t)} (U^{(\leq t)})^T$  and  $\Lambda_{(\leq t)}$  contains the eigenvalues corresponding to singular values  $\sigma_1 \geq \dots \geq \sigma_{r_t}$ . We may decompose  $M^{(\leq t)}$  as the sum of  $M^{(j)}$  for  $j = 1 \dots t$ , where each  $M^{(j)}$  has the spectral decomposition  $M^{(j)} = U^{(j)} \Lambda_j (U^{(j)})^T$  corresponding to the singular values  $\sigma_{r_{j-1}+1}, \dots, \sigma_{r_j}$ . Similarly, the matrix  $N_t$  may be written as  $N_t = (V_t) \Lambda_{(>t)} (V_t)^T$ , and contains the singular values  $\sigma_{r_t+1}, \dots, \sigma_n$ . Eventually, our algorithm will stop at some maximum  $t = T$ , for which  $r_t = k$ , and we will have  $A = M + N = M^{(\leq T)} + N_T$  as in (2). We will use the notation  $U^{(\leq j)}$  to denote the concatenation  $U^{(\leq j)} = [U^{(1)} | U^{(2)} | \dots | U^{(j)}]$ . Observe that this is consistent with the definition of  $U^{(\leq t)}$  above. Additionally, for a matrix  $X \in \mathbb{R}^{n \times r_t}$ , we will write  $X = [X^{(1)} | X^{(2)} | \dots | X^{(t)}]$ , where  $X^{(j)}$  contains the  $r_{j-1} + 1, \dots, r_j$  columns of  $X$ , and we will write  $X^{(\leq j)} = [X^{(1)} | X^{(2)} | \dots | X^{(j)}]$ . Occasionally, we will wish to use notation like  $U^{(\leq r)}$  to denote the first  $r$  columns (rather than the first  $r_r$  columns). This will be pointed out when it occurs. For an index  $r \leq n$ , we quantify the gap between  $\sigma_r$  and  $\sigma_{r+1}$  by

$$\gamma_r := 1 - \frac{\sigma_{r+1}}{\sigma_r}. \quad (7)$$

and we will define

$$\gamma := \min \left\{ \gamma_r : r \in [n], \gamma_r \geq \frac{1}{4k} \right\}. \quad (8)$$

By definition, we always have  $\gamma \geq 1/4k$ ; for some matrices  $A$ , it may be much larger, and this will lead to improved bounds. Our analysis will also depend on the “final” gap quantified by  $\gamma_k$ , whether or not it is larger than  $1/4k$ . To this end, we define

$$\gamma^* := \min \{ \gamma, \gamma_k \}. \quad (9)$$

### 3. Algorithms and Results

In Algorithm 1 we present our main algorithm SOFTDEFLATE. It uses several subroutines that are presented in Section 3.1.

#### 3.1. Overview of Subroutines

SOFTDEFLATE uses a number of subroutines that we outline here before explicitly presenting them:

- S-M-ALTLS (Algorithm 2) is the main Alternating Least Squares procedure that was given and analyzed in [Hardt \(2013b\)](#). We use this algorithm and its analysis. S-M-ALTLS by itself has a quadratic dependence on the condition number which is why we can only use it as a subroutine.
- SMOOTHQR (Algorithm 3) is a subroutine of S-M-ALTLS which is used to control the coherence of intermediate solutions arising in S-M-ALTLS. Again, we reuse the analysis of SMOOTHQR from [Hardt \(2013b\)](#). SMOOTHQR orthonormalizes its input matrix after adding

**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  $P_\Omega(A)$ ; Accuracy parameter  $\varepsilon$ ; Noise parameter  $\Delta$  with  $\|A - A_k\|_F \leq \Delta$ ; Coherence parameter  $\mu^*$ , satisfying (4), and a parameter  $\mu_0$ ; Probabilities  $p_0$  and  $p_t, p'_t$  for  $t = 1, \dots, k$ ; Number of iterations  $L_t \in \mathbb{N}$ , for  $t = 1, \dots, k$  runs of S-M-ALTLS, and a parameter  $s_{\max} \in \mathbb{N}$  for S-M-ALTLS, and a number of iterations  $L$  for runs of SUBSIT.

- 1 Let  $p = \sum_t (p_t + p'_t)$ .
- 2 Break  $\Omega$  randomly into  $2k + 1$  sets,  $\Omega_0$  and  $\Omega_1, \Omega'_1, \dots, \Omega_k, \Omega'_k$ , so that  $\mathbb{E}|\Omega_t| = \frac{p_t}{p}|\Omega|$  and  $\mathbb{E}|\Omega'_t| = \frac{p'_t}{p}|\Omega|$  (See Section D).
- 3  $s_0 \leftarrow \|P_{\Omega_0}(A)\|$  // Estimate  $\sigma_1(A)$
- 4 Initialize  $X_0 = Y_0 = 0, r_0 = 0$
- 5 **for**  $t = 1 \dots k$  **do**
- 6      $\tau_t \leftarrow \frac{\mu^*}{np_t} (2ks_{t-1} + \Delta)$
- 7      $T_t \leftarrow \text{TRUNCATE}(P_{\Omega_t}(A) - P_{\Omega_t}(X_{t-1}Y_{t-1}^T), \tau_t)$  // TRUNCATE( $M, c$ ) truncates  $M$  so that  $|\tilde{M}_{ij}| \leq c$
- 8      $\tilde{U}_t, \tilde{\sigma} \leftarrow \text{SUBSIT}(T_t, k - r_{t-1}, L)$  // Estimate the top  $k - r_{t-1}$  spectrum of  $T_t$ .
- 9     **If**  $\tilde{\sigma}_1 < 10\varepsilon s_0$  **then return**  $X_{t-1}, Y_{t-1}$
- 10      $d_t \leftarrow \min \left\{ i \leq k - r_{t-1} : \sigma_{i+1}(\tilde{T}_t) \leq \left(1 - \frac{1}{4k}\right) \sigma_i(\tilde{T}_t) \right\} \cup \{k - r_{t-1}\}$
- 11      $r_t \leftarrow r_{t-1} + d_t$  //  $r_t$  is an estimate of the next ‘‘gap’’ in the spectrum of  $A$
- 12      $s_t \leftarrow \tilde{\sigma}_{d_t}$  //  $s_t$  is an estimate of  $\sigma_{r_t}(A)$
- 13      $\tilde{Q}_t \leftarrow (\tilde{U}_t)^{(\leq d_t)}$  // Keep the first  $d_t$  columns of  $\tilde{U}_t$
- 14      $\bar{Q}_t \leftarrow \text{TRUNCATE}\left(\tilde{Q}_t B, 8\sqrt{\frac{\mu^* \log(n)}{n}}\right)$  // where  $B \in \mathbb{R}^{n \times n}$  is a random orthonormal matrix.
- 15      $W_t \leftarrow \text{QR}([X_{t-1} | \bar{Q}_t])$  //  $W_t$  is a rough estimate of  $U^{(\leq t)}$
- 16      $\mu_t \leftarrow (\sqrt{\mu_0} + (t-1)\sqrt{\mu^* k})^2$
- 17      $(X_t, Y_t) \leftarrow \text{S-M-ALTLS}(A, \Omega'_t, R_0 = W_t, L = L_t, s_{\max} = s_{\max}, k = r_t, \zeta = \varepsilon s_0 k^{-5}, \mu = \mu_t)$  //  $X_t$  is a good estimate of  $U^{(\leq t)}$
- 18     **If**  $r_t \geq k$  **then return**  $(X_t, Y_t)$
- 19 **end**

**Output:** Pair of matrices  $(X, Y)$ .

**Algorithm 1:** SOFTDEFLATE: Approximates an approximately low-rank matrix from a few entries.



a Gaussian noise matrix. This step allows tight control of the coherence of the resulting matrix. We defer the description of SMOOTHQR to Section B where we need it for the first time.

- SUBSIT is a standard textbook version of the Subspace Iteration algorithm (Power Method). We use this algorithm as a fast way to approximate the top singular vectors of a matrix arising in SOFTDEFLATE. We use only standard properties of SUBSIT in our analysis. For this reason we defer the description and analysis of SUBSIT to Section E.3.

**Input:** Number of iterations  $L \in \mathbb{N}$ , parameter  $s_{\max} \in \mathbb{N}$ , 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  $P_{\Omega}(A)$ , initial orthonormal matrix  $R_0 \in \mathbb{R}^{n \times k}$ , and parameters  $\zeta, \mu$

20 Break  $\Omega$  randomly into sets  $\Omega_1, \dots, \Omega_L$  with equal expected sizes. (See Section D).

21 **for**  $\ell = 1$  to  $L$  **do**

22     Break  $\Omega_{\ell}$  randomly into subsets  $\Omega_{\ell}^{(1)}, \dots, \Omega_{\ell}^{(T)}$  with equal expected sizes.

23     **for**  $s = 1$  to  $s_{\max}$  **do**

24          $S_{\ell}^{(s)} \leftarrow \arg \min_{S \in \mathbb{R}^{n \times k}} \|P_{\Omega_{\ell}}(A - R_{\ell-1}S^T)\|_F^2$

25     **end**

26      $S_{\ell} \leftarrow \text{median}_s(S_{\ell}^{(s)})$  // The median is applied entry-wise.

27      $R_{\ell} \leftarrow \text{SMOOTHQR}(S_{\ell}, \zeta, \mu)$

28 **end**

**Output:** Pair of matrices  $(R_{L-1}, S_L)$

**Algorithm 2:** S-M-ALTLS( $P_{\Omega}(A), \Omega, R_0, L, s_{\max}, k, \zeta, \mu$ ) (Smoothed-Median-Alternating Least Squares)

### 3.2. Statement of the main theorem

Our main theorem is that, when the number of samples is  $\text{poly}(k)n$ , SOFTDEFLATE returns a good estimate of  $A$ , with at most logarithmic dependence on the condition number.

**Theorem 1** *There is a constant  $C$  so that the following holds. Let  $A \in \mathbb{R}^{n \times n}$ ,  $k \leq n$ , and write  $A = M + N$ , where  $M$  is the best rank- $k$  approximation to  $A$ . Let  $\gamma, \gamma^*$  be as in (8), (9). Choose parameters for Algorithm 1 so that  $\varepsilon > 0$ ;  $\mu^*$  satisfies (4);  $\mu_0 \geq \frac{C}{(\gamma^*)^2} \left( \mu^* \left( k + \left( \frac{k^4 \Delta}{\varepsilon \sigma_1} \right)^2 \right) + \log(n) \right)$ ;  $\Delta \geq \|N\|_F$ ;  $L_t \geq \frac{C}{\gamma^*} \log \left( \frac{k \sigma_{r_t}}{\sigma_{r_t+1} + \varepsilon \sigma_1} \right)$ , and  $L \geq C k^{7/2} \log(n)$ ; and  $s_{\max} \geq C \log(n)$ . There is a choice of  $p_t, p'_t$  (given in the proof below) so that*

$$p = \sum p_t + \sum p'_t \leq C \frac{k^9}{(\gamma^*)^3 n} \log \left( k \cdot \frac{\sigma_1}{\sigma_k + \varepsilon \sigma_1} \right) \left( 1 + \left( \frac{\Delta}{\varepsilon \|M\|} \right)^2 \right) (\mu_0 + \mu^* k \log(n)) \log^2(n)$$

so that the following holds. Suppose that each element of  $[n] \times [n]$  is included in  $\Omega$  independently with probability  $p$ . Then the matrices  $X, Y$  returned by SOFTDEFLATE satisfy with probability at least  $1 - 1/n$ ,  $\|A - XY^T\| \leq (1 + o(1)) \|N\| + \varepsilon \|M\|$ .

**Remark 2 (Error guarantee)** *The guarantee of  $\|A - XY^T\| \leq (1 + o(1)) \|N\| + \varepsilon \|M\|$  is what naturally falls out of our analysis: the natural value for the  $o(1)$  term is polynomially small in  $k$ . It is not hard to see in the proof that we may make this term as small as we like, say,  $(1 + \alpha) \|N\|$ , by paying a logarithmic penalty  $\log(1/\alpha)$  in the choice of  $p$ . It is also not hard to see that we may have a similar conclusion for the Frobenius norm.*

**Remark 3 (Obtaining the parameters)** *As written, then algorithm requires the user to know several parameters which depend on the unknown matrix  $A$ . For some parameters, these requirements are innocuous. For example, to obtain  $p'_t$  or  $L_t$ , the user is required to have a bound on  $\log(\sigma_{r_t}/\sigma_{r_t+1})$ . Clearly, a bound on the condition number of  $A$  will suffice, but more importantly, the estimates  $s_t$  which appear in Algorithm 1 may be used as proxies for  $\sigma_{r_t}$ , and so the parameters  $p'_t$  can actually be determined relatively precisely on the fly. For other parameters, like  $\mu^*$  or  $k$ , we assume that the user has a good estimate from other sources. While this is standard in the Matrix Completion literature, we acknowledge that these values may be difficult to come by.*

### 3.3. Running Time

The running time of SOFTDEFLATE is linear in  $n$ , polynomial in  $k$ , and logarithmic in the condition number  $\sigma_1/\sigma_k$  of  $A$ . Indeed, the outer loop performs at most  $k$  epochs, and the nontrivial operations in each epoch are S-M-ALTLS, QR, and SUBSIT. All of the other operations (truncation, concatenation) are done on matrices which are either  $n \times k$  (requiring at most  $nk$  operations) or on the subsampled matrices  $P_{\Omega_t}(A)$ , requiring on the order of  $pn^2$  operations.

Running SUBSIT requires  $L = O(k^{7/2} \log(n))$  iterations; each iteration includes multiplication by a sparse matrix, followed by QR. The matrix multiplication takes time on the order of

$$p_t n^2 = n \operatorname{poly}(k) \log(n) \left(1 + \frac{\Delta}{\varepsilon \sigma_1}\right),$$

the number of nonzero entries of  $A$ , and QR takes time  $O(k^2 n)$ . Each time S-M-ALTLS is run, it takes  $L_t$  iterations, and we will show (following the analysis of [Hardt \(2013b\)](#)) that it requires  $\operatorname{poly}(k)n \log(n) \log(n/\varepsilon)$  operations per iteration. Thus, given the choice of  $L_t$  in Theorem 1, the total running time of SOFTDEFLATE on the order of

$$\tilde{O} \left( n \cdot \operatorname{poly}(k) \cdot \left(1 + \frac{\Delta}{\varepsilon \sigma_1}\right) \cdot \log \left( \frac{\sigma_1}{\sigma_k + \varepsilon \sigma_1} \right) \right),$$

where the  $\tilde{O}$  hides logarithmic factors in  $n$ .

## 4. Proof of Main Theorem

In this section, we prove Theorem 1. The proof proceeds by maintaining a few inductive hypotheses, given below, at each epoch. When the algorithm terminates, we will show that the fact that these hypotheses still hold imply the desired results. Suppose that at the beginning of step  $t$  of Algorithm 1, we have identified some indices  $r_1, \dots, r_{t-1}$ , and recovered estimates  $X_{t-1}, Y_{t-1}$  which capture the singular values  $\sigma_1, \dots, \sigma_{r_{t-1}}$  and the corresponding singular vectors. The goals of the current step of Algorithm 1 are then to (a) identify the next index  $r_t$  which exhibits a large “gap” in the spectrum,

and (b) estimate the singular values  $\sigma_{r_{t-1}+1}, \dots, \sigma_{r_t}$  and the corresponding singular vectors. Letting  $r_t$  be the index obtained by Algorithm 1, we will decompose  $A = M^{(<t)} + N_{t-1} = M^{(\leq t)} + N_t$  as in (6). As in Section D, we treat the  $\Omega_t$  and  $\Omega'_t$  as independent random sets, with each entry included with probability  $p_t$  or  $p'_t$ , respectively. We will keep track of the *principal angles* between the subspaces  $\mathcal{R}(\binom{\leq j}{X_{t-1}})$  and  $\mathcal{R}(\binom{\leq j}{U})$ . More precisely, for matrices  $A, B \in \mathbb{R}^{n \times r_j}$  with orthogonal columns, we define  $\sin \theta(A, B) := \|(A_\perp)^T B\|$ .

We will maintain the following inductive hypotheses. At the beginning of epoch  $t$  of SOFTDEFLATE, we assert

$$\sigma_{r_j} \sin \theta(X_{t-1}^{\binom{\leq j}{}}, U^{\binom{\leq j}{}}) \leq \frac{1}{k^4} (\sigma_{r_{t-1}+1} + \varepsilon \|M\|) \quad \forall j \leq t-1 \quad (\text{H1})$$

and

$$\|M^{(<t)} - X_{t-1} Y_{t-1}^T\| \leq \frac{\sigma_{r_{t-1}+1} + \varepsilon \|M\|}{C_0 k^3} \quad (\text{H2})$$

for some sufficiently large constant  $C_0$  determined by the proof. We also maintain that the current estimate  $X_{t-1}$  is incoherent:

$$\max_{i \in [n]} \|e_i^T X_{t-1}\|_2 \leq \sqrt{\frac{k}{n}} \left( \sqrt{\mu_0} (1 + C_5/k)^{t-1} + (t-1) 16 \sqrt{\mu^* \log(n)} \right) =: \sqrt{\frac{k \mu_{t-1}}{n}} \quad (\text{H3})$$

for a constant  $C_5$ . Above, equation (H3) defines  $\mu_{t-1}$ . Observe that when  $t = 1$ , everything in sight is zero and the hypotheses (H1), (H2), (H3) are satisfied. Finally, we assume that the estimate  $s_{t-1}$  of  $\sigma_{r_{t-1}+1}$  is good.

$$\frac{1}{2} \sigma_{r_{t-1}+1} \leq s_{t-1} \leq 2 \sigma_{r_{t-1}+1} \quad (\text{H4})$$

The base case for (H4) is handled by the choice of  $s_0$  in Algorithm 1. Indeed, Lemma 21 in the appendix implies that as long as

$$p_0 \gtrsim \frac{\mu^* \log(n) \left( \sqrt{k} + \frac{\Delta}{\sigma_1} \right)^2}{n}, \quad (\text{10})$$

then with probability  $1 - 1/\text{poly}(n)$ ,

$$\frac{1}{2} \sigma_1 \leq \|P_{\Omega_0}(A)\| \leq 2 \sigma_1.$$

and so (H4) is satisfied. Now, suppose that the inductive hypotheses (H1), (H2), (H3), and (H4) hold. We break up the inner loop of SOFTDEFLATE into two main steps. In the first step, lines 6 to 15 in Algorithm 1, the goal is to obtain an estimate  $r_t$  of the next ‘‘gap,’’ as well as an estimate  $W_t$  of the subspace  $U^{(\leq t)}$ . We analyze this step in Lemma 4 below.

**Lemma 4** *There exists a constants  $C, C_1$  so that the following holds. Suppose that*

$$p_t \geq \frac{C(\mu^*)^2 \log(n) \left( k^2 + \left( \frac{\Delta}{\varepsilon \|M\|} \right)^2 \right)}{n \varepsilon_0^2},$$

where  $\varepsilon_0 \leq \frac{1}{4C_1 k^{5/2}}$ . Further assume that the inductive hypotheses (H1), (H2), (H3), and (H4) hold. Then with probability at least  $1 - 1/n^2$  over the choice of  $\Omega_t$  and the randomness in SUBSIT, one of the following statements must hold:

- (a) Algorithm 1 terminates at line 9, and returns  $X_{t-1}, Y_{t-1}$  so that  $\|A - X_{t-1}Y_{t-1}^T\| \leq C\varepsilon \|M\|$
- (b) Algorithm 1 does not terminate at line 9, and the following conditions hold. First, The error level  $\varepsilon$  has not yet been reached:

$$\varepsilon \|M\| \leq \sigma_{r_{t-1}+1}. \quad (11)$$

Second, the index  $r_t$  recovered obeys

$$\frac{\sigma_{r_t+1}}{\sigma_{r_t}} \leq 1 - \gamma \quad \text{and} \quad \frac{\sigma_{r_{t-1}+1}}{\sigma_{r_t}} \leq e. \quad (12)$$

Third, the matrix  $W_t$  has orthonormal columns, and satisfies

$$\sin \theta(W_t, U^{(\leq t)}) \leq \frac{1}{k} \quad \text{and} \quad \max_i \|e_i^T W_t\|_2 \leq \sqrt{\frac{k\mu_t}{n}}, \quad (13)$$

where  $\mu_t$  is defined as in (H3). Fourth, the estimate  $s_t$  satisfies (H4).

The proof of Lemma 4 is given in Section A. In the second part of SOFTDEFLATE, lines 16 to 17 in Algorithm 1, we run S-M-ALTLS, initialized with the subspace  $W_t$  returned by the first part of the algorithm. Lemma 5 below shows that S-M-ALTLS improves the estimate  $W_t$  to the desired accuracy, so that we may move on to the next iteration of SOFTDEFLATE.

**Lemma 5** *Assume that the conclusion (b) of Lemma 4 holds, as well as the inductive hypotheses (H1), (H2), (H3), and (H4). There is a constant  $C$  so that the following holds. Let  $\gamma^*$  be as in (9). Suppose that*

$$\mu_t \geq \frac{C}{(\gamma^*)^2} \left( \mu^* \left( k + \left( \frac{k^4 \Delta}{\varepsilon \sigma_1} \right)^2 \right) + \log(n) \right) \quad \text{and} \quad p'_t \geq \frac{CL_t s_{\max} \cdot k^9 \mu_t \log(n) \left( k + \left( \frac{\Delta}{\varepsilon \|M\|} \right)^2 \right)}{(\gamma^*)^2 n}$$

where  $L_t \geq \frac{C}{\gamma^*} \log \left( \frac{k\sigma_{r_t}}{\sigma_{r_t+1} + \varepsilon \|M\|} \right)$  and  $s_{\max} \geq C \log(n)$ . Then after  $L_t$  steps of S-M-ALTLS with the initial matrix  $W_t$ , and parameters  $\mu_t, \varepsilon$ , with probability at least  $1 - 1/n^2$ , over the choice of  $\Omega'_t$ , the inductive hypotheses (H1), (H2), and (H3) hold for  $t$ .

The proof of Lemma 5 is addressed in Section B. Theorem 1 now follows using 4 and 5. First, we choose  $\mu_0$  as in the statement of Theorem 1. Because  $\mu_t \geq \mu_0$  for all  $t = 1, \dots, T$ , this implies that  $\mu_t$  satisfies the requirements of Lemma 5. Then, the hypotheses of Lemma 5 are implied by the conclusions of the favorable case of Lemma 4. Now, a union bound over at most  $k$  epochs of SOFTDEFLATE ensures that with probability at least  $1 - 2k/n^2 \geq 1 - 1/n$ , the conclusions of both lemmas hold every round that their hypotheses hold. If SOFTDEFLATE terminates with the guarantees (a) of Lemma 4, then  $\|A - X_T Y_T^T\| \leq C\varepsilon \|M\|$ . On the other hand, if (b) holds, then Lemma 4 implies (H4) and the hypotheses of Lemma 5, and then Lemma 5 implies that with probability  $1 - 1/n^2$ , the remaining inductive hypotheses (H1), (H2), and (H3) for the next round. Thus, if the situation (a) above never occurs, then the hypotheses (and the conclusions) of Lemma 5 hold until SOFTDEFLATE terminates because  $r_t = k$ . In either case,  $\|A - X_t Y_t^T\| \leq \|N\| \left( 1 + \frac{1}{C_0 k^3} \right) + C\varepsilon \|M\|$ . Finally, the bound on  $p$  follows by adding the bound on  $p_0$  in (10) with the bounds on  $p_t$  and  $p'_t$  from Lemmas 4 and 5. This completes the proof of Theorem 1.

## Acknowledgments

We thank the Simons Institute for Theoretical Computer Science at Berkeley, where part of this work was done.

## References

- Haim Avron, Satyen Kale, Shiva Prasad Kasiviswanathan, and Vikas Sindhwani. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *Proc. 29th ICML*. ACM, 2012.
- Robert M. Bell and Yehuda Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *ICDM*, pages 43–52. IEEE Computer Society, 2007.
- Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, December 2009.
- 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.
- Suriya Gunasekar, Ayan Acharya, Neeraj Gaur, and Joydeep Ghosh. Noisy matrix completion using alternating minimization. In *Proc. ECML PKDD*, pages 194–209. Springer, 2013.
- Justin P. Haldar and Diego Hernando. Rank-constrained solutions to linear matrix equations using powerfactorization. *IEEE Signal Process. Lett.*, 16(7):584–587, 2009.
- Moritz Hardt. Robust subspace iteration and privacy-preserving spectral analysis. *arXiv*, 1311:2495, 2013a.
- Moritz Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv*, 1312.0925, 2013b.
- Elad Hazan and Satyen Kale. Projection-free online learning. In *Proc. 29th ICML*. ACM, 2012.
- Cho-Jui Hsieh and Peder A. Olsen. Nuclear norm minimization via active subspace selection. In *Proc. 31st ICML*. ACM, 2014.
- Martin Jaggi and Marek Sulovský. A simple algorithm for nuclear norm regularized problems. In *Proc. 27th ICML*, pages 471–478. ACM, 2010.
- Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In *Proc. 24th Neural Information Processing Systems (NIPS)*, pages 937–945, 2010.
- Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proc. 45th Symposium on Theory of Computing (STOC)*, pages 665–674. ACM, 2013.
- Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proc. 26th ICML*, page 58. ACM, 2009.

- Raghunandan H. Keshavan. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.
- Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010a.
- Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010b.
- Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.
- Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.
- Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.
- Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.
- Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- Gilbert W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press London, 1990.
- G.W. Stewart. *Matrix Algorithms. Volume II: Eigensystems*. Society for Industrial and Applied Mathematics, 2001.
- Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

## Appendix A. Proof of Lemma 4

In this section, we prove Lemma 4, which shows that either Algorithm 1 hits the precision parameter  $\varepsilon$  and returns, or else produces an estimate  $W_t$  for  $U^{(\leq t)}$  that is close enough to run S-M-ALTLS on. There are several rounds of approximations between the beginning of iteration  $t$  and the output  $W_t$ . For the reader’s convenience, we include an informal synopsis of the notation in Figure 1. We will first argue that the matrix  $N_{t-1}$  is close to the truncated, subsampled, noisy estimate  $T_t$ .

**Lemma 6** *Let  $T_t$  be as in Algorithm 1, and choose any constant  $C_1 > 0$ . Suppose that the inductive hypotheses (H2) and (H4) hold. Suppose that  $p_t$  is as in the statement of Lemma 4. Then, for a sufficiently large choice of  $C_0$  in the hypothesis (H2) (depending only on  $C_1$ ), with probability at least  $1 - 1/n^2$ ,*

$$\|T_t - N_{t-1}\| \leq \frac{\sigma_{r_{t-1}+1} + \varepsilon \|M\|}{2C_1 k^{5/2}}.$$

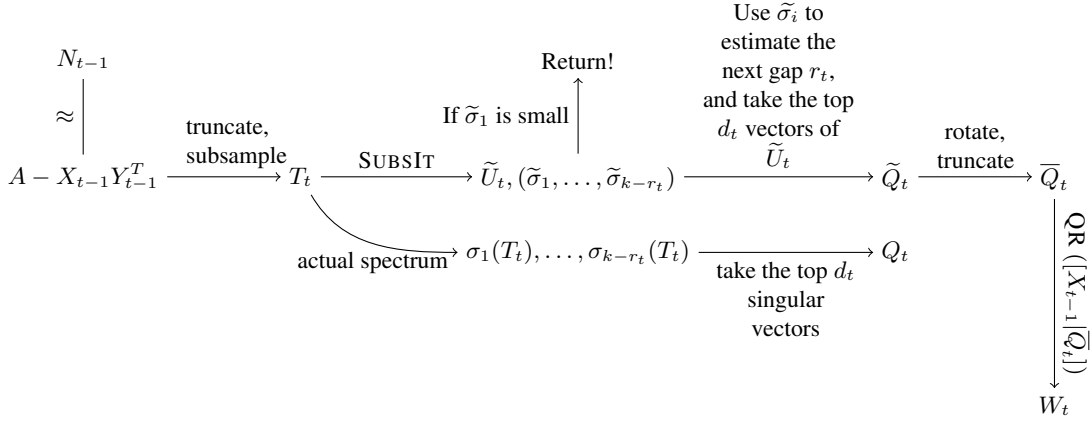


Figure 1: Schematic of the first part of SOFTDEFLATE. The top line indicates how  $W_t$  is formed from the matrix  $T_t$ . We will show that  $\bar{Q}_t$  approximates  $U^{(t)}$ , the next chunk of singular vectors in  $N_{t-1}$ , and this will imply by induction that  $W_t$  approximates  $U^{(\leq t)}$ . The second line in the figure indicates some notation which will be useful for our analysis, but which is not used by the algorithm.

**Proof** Write

$$A - X_{t-1}Y_{t-1}^T = N_{t-1} + \left( M^{(<t)} - X_{t-1}Y_{t-1}^T \right) =: N_{t-1} + E_{t-1} =: \widetilde{N}_{t-1}.$$

Let  $\mathcal{T}$  denote the TRUNCATE operator. As in Algorithm 1, consider

$$T_t = \mathcal{T}(P_{\Omega_t}(\widetilde{N}_{t-1}), \tau_t) = P_{\Omega_t} \left( \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) \right),$$

where as in Line 6,  $\tau_t = \frac{\mu^*}{np_t} (2ks_{t-1} + \Delta)$ . Above, we used that the sampling operation  $P_{\Omega_t}$  and the truncate operator  $\mathcal{T}$  commute after adjusting for the normalization factor  $p_t^{-1}$  in the definition of  $P_{\Omega_t}$ . Because  $N_{t-1}$  is incoherent, each of its entries is small. More precisely, by the incoherence implication (37) along with the guarantee (H4) on  $s_{t-1}$ , we have

$$\|N_{t-1}\|_{\infty} \leq \frac{\mu^*}{n} (k\sigma_{r_{t-1}+1} + \Delta) \leq \frac{\mu^*}{n} (2ks_{t-1} + \Delta) = p_t \tau_t.$$

Thus, each entry of  $\widetilde{N}_{t-1} = N_{t-1} + E_{t-1}$  is the sum of something smaller than  $p_t \tau_t$  from  $N_{t-1}$ , and an error term from  $E_{t-1}$ , and so truncating entrywise to  $p_t \tau_t$  can only remove mass from the contribution of  $E_{t-1}$ . This implies that for all  $i, j$ ,

$$\left| \widetilde{N}_{t-1} - \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) \right|_{i,j} \leq |E_{t-1}|_{i,j},$$

and so using (H2),

$$\left\| \widetilde{N}_{t-1} - \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) \right\|_F \leq \|E_{t-1}\|_F \leq \sqrt{2k} \frac{(\sigma_{r_{t-1}+1} + \varepsilon \|M\|)}{C_0 k^3} = \frac{\sqrt{2}(\sigma_{r_{t-1}+1} + \varepsilon \|M\|)}{C_0 k^{5/2}}. \quad (14)$$

Above, we used the fact that  $E_{t-1} = M^{(<t)} - X_{t-1}Y_{t-1}^T$  has rank at most  $2k$ , and hence  $\|E_{t-1}\|_F \leq \sqrt{2k} \|E_{t-1}\|$ . Next, we bound the difference between  $T_t$  and  $\mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t)$ . Lemma 21 in the appendix bounds the effect of subsampling in operator norm. It implies that with probability  $1 - 1/\text{poly}(n)$  over the choice of  $\Omega_t$ , we have

$$\begin{aligned} \left\| \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) - T_t \right\| &= \left\| \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) - P_\Omega(\mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t)) \right\| \\ &\lesssim \sqrt{\frac{\max_i \left\| e_i^T \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) \right\|_2^2 \log(n)}{p_t}} + \frac{\left\| \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) \right\|_\infty \log(n)}{p_t} \\ &\leq \sqrt{\frac{n(p_t \tau_t)^2 \log(n)}{p_t}} + \frac{(p_t \tau_t) \log(n)}{p_t} \\ &\leq \left( \sqrt{\frac{\log(n)}{p_t n}} + \frac{\log(n)}{p_t n} \right) (\mu^* (4k\sigma_{r_{t-1}+1} + \Delta)), \end{aligned}$$

using the fact that

$$p_t \tau_t = \frac{\mu^*}{n} (2k s_{t-1} + \Delta) \leq \frac{\mu^*}{n} (4k\sigma_{r_{t-1}+1} + \Delta)$$

by (H4). Thus, our choice of  $p_t$  implies that

$$\left\| \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) - T_t \right\| \leq \varepsilon_0 (\sigma_{r_{t-1}+1} + \varepsilon \|M\|). \quad (15)$$

Together with (14) we conclude that

$$\begin{aligned} \|N_{t-1} - T_t\| &\leq \|N_{t-1} - \widetilde{N}_{t-1}\| + \|\widetilde{N}_{t-1} - \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t)\| + \left\| \mathcal{T}(\widetilde{N}_{t-1}, p_t \tau_t) - T_t \right\| \\ &\leq \varepsilon_0 (\sigma_{r_{t-1}+1} + \varepsilon \|M\|) + \frac{2\sqrt{2}(\sigma_{r_{t-1}+1} + \varepsilon \|M\|)}{C_0 k^{5/2}}. \end{aligned}$$

The choice of  $\varepsilon_0$  and a sufficient choice of  $C_0$  (depending only on  $C_1$ ) completes the proof.  $\blacksquare$

Suppose for the rest of the proof that the conclusion of Lemma 6 holds. The first thing SOFT-DEFLATE does after computing  $T_t$  is to obtain estimates  $\widetilde{U}_t$  and  $\widetilde{\sigma}_1, \dots, \widetilde{\sigma}_{k-r_t}$  for the top singular values and vectors of  $T_t$ . These estimates are recovered by SUBSIT in Line 8 of Algorithm 1. We first wish to show that the estimated singular values are close to the actual singular values of  $T_t$ . For this, we will invoke Theorem 19 in the appendix, which implies that as long as the number of iterations  $L$  of SUBSIT satisfies

$$L \geq Ck^{7/2} \log(n),$$

for a sufficiently large constant  $C$ , then with probability  $1 - 1/\text{poly}(n)$ , we have

$$|\sigma_j(T_t) - \widetilde{\sigma}_j| \leq \frac{\|T_t\|}{2C_1 k^{5/2}} \quad \text{for all } j. \quad (16)$$

Above, we took a union bound over all  $j$ . Again, we condition on this event occurring. Thus, with our choice of  $L$ , the estimates  $\widetilde{\sigma}_j$  are indeed close to the singular values  $\sigma_j(T_t)$ , which by Lemma 6 are with high probability close to the singular values  $\sigma_{r_{t-1}+j}$  of  $N_{t-1}$  itself.



Before we consider the next step (to  $\tilde{Q}_t$ ) in Figure 1, consider the case when Algorithm 1 returns at line 9. Then  $\tilde{\sigma}_1 \leq 10\varepsilon s_0 \leq 20\varepsilon\sigma_1$ , and so using (16) above we find that  $\|T_t\| \leq 21\varepsilon\sigma_1$ . Then by Lemma 6,

$$\sigma_{r_{t-1}+1} = \|N_{t-1}\| \leq \|T_t\| + \|N_{t-1} - T_t\| \leq 21\varepsilon\sigma_1 + \frac{\sigma_{r_{t-1}+1} + \varepsilon\sigma_1}{2C_1k^{5/2}}.$$

Thus, for sufficiently large  $C_1$ , we conclude  $\sigma_{r_{t-1}+1} \leq 22\varepsilon\sigma_1$ . In this case, we are done:

$$\begin{aligned} \|A - X_{t-1}Y_{t-1}^T\| &\leq \|M^{(<t)} - X_{t-1}Y_{t-1}^T\| + \|N_{t-1}\| \\ &\leq \frac{\sigma_{r_{t-1}+1} + \varepsilon\|M\|}{C_0k^3} + \sigma_{r_{t-1}+1} \\ &\leq 23\varepsilon\sigma_1. \end{aligned}$$

and case (a) of the conclusion holds, as long as Lemma 6 does.

On the other hand, suppose that Algorithm 1 does not return at line 9 (and continue to assume that Lemma 6 holds). As above, (16) implies that since  $\tilde{\sigma}_1 \geq 10\varepsilon$ , we must have

$$\|T_t\| \geq \frac{5\varepsilon\sigma_1}{1 - \frac{1}{2C_1k^{5/2}}}.$$

Then by Lemma 6,

$$\sigma_{r_{t-1}+1} \geq \frac{5\varepsilon\sigma_1}{1 - \frac{1}{2C_1k^{5/2}}} - \frac{\sigma_{r_{t-1}+1} + \varepsilon\sigma_1}{2C_1k^{5/2}},$$

which implies that

$$\varepsilon\sigma_1 < \sigma_{r_{t-1}+1}. \quad (17)$$

This establishes the conclusion (11). With (17), Lemma 6 and (16) together imply that

$$\forall r \leq n, \quad |\sigma_r - \tilde{\sigma}_{r-r_{t-1}}| \leq \|N_{t-1} - T_t\| + |\sigma_{r-r_{t-1}}(T_t) - \tilde{\sigma}_{r-r_{t-1}}| \leq \frac{\sigma_{r_{t-1}+1}}{C_1k^{5/2}}. \quad (18)$$

Above, we use Lemma 16 in the appendix in the first inequality.

We now show that the choice of  $d_t$  in Line 10 of Algorithm 1 accurately identifies a ‘‘gap’’ in the spectrum.

**Lemma 7** *Suppose that the hypotheses and conclusions of Lemma 6 hold, and in particular that (18) holds. Then the value  $r_t = r_{t-1} + d_t$  obtained in Line 10 of Algorithm 1 satisfies:*

$$\frac{\sigma_{r_t+1}}{\sigma_{r_t}} \leq 1 - \gamma \quad \text{and} \quad \frac{\sigma_{r_{t-1}+1}}{\sigma_{r_t}} \leq e.$$

**Proof** Let  $d_t^*$  be the ‘‘correct’’ choice of  $d_t$ ; that is,  $d_t^*$  be the smallest positive integer  $d \leq k - r_{t-1}$  so that

$$1 - \frac{\sigma_{r_{t-1}+d+1}}{\sigma_{r_{t-1}+d}} \geq 1 - \frac{1}{k},$$

or let  $d_t^* = d - r_{t-1}$  if such an index does not exist. Write  $r_t^* = r_{t-1} + d_t^*$ . By definition, because  $d_t^*$  is the smallest such  $d$  (or smaller than any such  $d$  in the case that  $r_t^* = k$ ), we have

$$\frac{\sigma_{r_{t-1}+1}}{\sigma_{r_t^*}} \leq \left(1 + \frac{1}{k}\right)^{d_t^*} \leq e. \quad (19)$$

Thus, (18) reads

$$|\tilde{\sigma}_j - \sigma_{r_{t-1}+j}| \leq \frac{\sigma_{r_{t-1}+1}}{C_1 k^{5/2}} \leq \frac{e\sigma_{r_t^*}}{C_1 k^{5/2}}. \quad (20)$$

Suppose that, for some  $j \leq d_t^*$ , we have

$$\frac{\sigma_{r_{t-1}+j+1}}{\sigma_{r_{t-1}+j}} \geq 1 - \frac{1}{4k}.$$

Then, using (20),

$$\frac{\tilde{\sigma}_{j+1}}{\tilde{\sigma}_j} \geq \frac{\sigma_{r_{t-1}+j+1} - \frac{e\sigma_{r_t^*}}{C_1 k^{5/2}}}{\sigma_{r_{t-1}+j} + \frac{e\sigma_{r_t^*}}{C_1 k^{5/2}}} \geq \frac{\sigma_{r_{t-1}+j+1} \left(1 - \frac{e}{C_1 k^{5/2}}\right)}{\sigma_{r_{t-1}+j} \left(1 + \frac{e}{C_1 k^{5/2}}\right)} \geq 1 - \frac{1}{2k},$$

assuming  $C_1$  is sufficiently large. In Algorithm 1, we choose  $d_t$  in Line 10 so that there is no  $j < d_t$  with

$$\frac{\tilde{\sigma}_{j+1}}{\tilde{\sigma}_j} \leq 1 - \frac{1}{2k}.$$

Thus, if there were a big gap, the algorithm would have found it: more precisely, using the definition of  $\gamma$ , we have

$$\frac{\sigma_{r_t+1}}{\sigma_{r_t}} < 1 - \frac{1}{4k} \leq 1 - \gamma.$$

This establishes the first conclusion of the lemma. Now, a similar analysis as above shows that if for any  $j \leq d_t^*$  we have

$$\frac{\sigma_{r_{t-1}+j+1}}{\sigma_{r_{t-1}+j}} \leq 1 - \frac{1}{k},$$

then

$$\frac{\tilde{\sigma}_{j+1}}{\tilde{\sigma}_j} \leq 1 - \frac{1}{2k},$$

assuming  $C_1$  is sufficiently large. That is, our algorithm will always find a small gap, if it exists. In particular, if  $r_t^* < k$ , we have

$$\frac{\sigma_{r_t^*+1}}{\sigma_{r_t^*}} \leq 1 - \frac{1}{k}$$

and hence  $d_t \leq d_t^*$ . On the other hand, if  $r_t^* = k$ , then we must have  $d_t = d_t^*$ . In either case,  $d_t \leq d_t^*$ , and so

$$\frac{\sigma_{r_{t-1}+1}}{\sigma_{r_t}} \leq \frac{\sigma_{r_{t-1}+1}}{\sigma_{r_t^*}} \leq \left(1 + \frac{1}{k}\right)^{d_t^*} \leq e.$$

This completes the proof of Lemma 7. ■

Now, we are in a position to verify the inductive hypothesis (H4) for the next round, in the favorable case that Lemma 6 holds. By definition, we have  $s_t = \tilde{\sigma}_{d_t}$ , and (18), followed by Lemma 7 implies that

$$|\sigma_{r_t} - s_t| \leq \frac{\sigma_{r_{t-1}+1}}{C_1 k^2} \leq \frac{e\sigma_{r_t}}{C_1 k^{5/2}}.$$

In particular,

$$\left(1 - \frac{2e}{C_1 k^{5/2}}\right) \sigma_{r_t} \leq s_t \leq \left(1 + \frac{2e}{C_1 k^{5/2}}\right) \sigma_{r_t},$$

establishing (H4) for  $s_t$ .

Now that we know that the “gap” structure of the singular values of  $N_{t-1}$  is reflected by the estimates  $\tilde{\sigma}_j$ , we will show that the top singular vectors are also well-approximated by the estimates  $\tilde{Q}_t$ . Recall from Algorithm 1 that  $\tilde{Q}_t \in \mathbb{R}^{n \times d_t}$  denotes the first  $d_t$  columns of  $\tilde{U}_t$ , which are estimates of the top singular vectors of  $T_t$ . Let  $Q_t$  denote the (actual) top  $d_t$  singular vectors of  $T_t$ . We will first show that  $Q_t$  is close to  $U^{(t)}$ , and then that  $Q_t$  is also close to  $\tilde{Q}_t$ .

**Lemma 8** *Suppose that the conclusions of Lemma 6 and Lemma 7 hold, and that (17) holds. Then*

$$\sin \theta(U^{(t)}, Q_t) \leq \frac{4e}{C_1 k^{3/2}}.$$

**Proof** We will use a  $\sin \theta$  theorem (Theorem 17, due to Wedin, in the appendix) to control the perturbation of the subspaces. Theorem 17 implies

$$\begin{aligned} \sin \theta(U^{(t)}, Q_t) &\leq \frac{\|T_t - N_{t-1}\|}{|\sigma_{d_t}(T_t) - \sigma_{r_{t+1}}|} && \text{Theorem 17} \\ &\leq \frac{2e\sigma_{r_t}}{C_1 k^{5/2} |\sigma_{d_t}(T_t) - \sigma_{r_{t+1}}|} && \text{By Lemmas 6 and 7, and (17)} \\ &\leq \frac{2e\sigma_{r_t}}{C_1 k^{5/2} \left( \sigma_{r_t} \left( 1 - \frac{2e}{C_1 k^{5/2}} \right) - \sigma_{r_{t+1}} \right)} && \text{By (18) and Lemma 7} \\ &\leq \frac{e\sigma_{r_t}}{C_1 k^{5/2} \left( \sigma_{r_t} \left( 1 - \frac{2e}{C_1 k^{5/2}} \right) - \sigma_{r_t}(1 - \gamma) \right)} && \text{By Lemma 7} \\ &\leq \frac{4e}{C_1 k^{3/2}}. \end{aligned}$$

■

Now, we show that  $Q_t$  is close to  $\tilde{Q}_t$ .

**Lemma 9** *Suppose that the conclusions of Lemma 6 and Lemma 7 hold, and that (17) holds. Then with probability  $1 - 1/n^2$ ,*

$$\sin \theta(Q_t, \tilde{Q}_t) \leq \frac{1}{\text{poly}(n)}.$$

**Proof** By (16), Lemma 6, and Lemma 7, a similar computation as in the proof of Lemma 8 shows that

$$\begin{aligned} \frac{\sigma_{d_{t+1}}(T_t)}{\sigma_{d_t}(T_t)} &\leq \left( \frac{\tilde{\sigma}_{d_{t+1}}}{\tilde{\sigma}_{d_t}} \right) \left( 1 + \frac{8e}{C_1 k^{5/2}} \right) \\ &\leq \left( 1 - \frac{1}{4k} \right) \left( 1 + \frac{8e}{C_1 k^{5/2}} \right) \\ &\leq 1 - \frac{1}{k} \end{aligned}$$

using the choice of  $d_t$  in the second-to-last line. Thus, by Theorem 19 in the appendix, and the choice of  $L \gtrsim k \log(n)$  in SUBSIT, we have with probability  $1 - 1/\text{poly}(n)$  that

$$\sin \theta(Q_t, \tilde{Q}_t) \leq \text{poly}(n) \left( 1 - \frac{1}{2k} \right)^L \leq \frac{1}{\text{poly}(n)}.$$

■

Together, Lemmas 8 and 9 imply that, when Lemma 6 and the favorable case for SUBSIT hold,

$$\sin \theta(U^{(t)}, \tilde{Q}_t) \leq \frac{8e}{C_1 k^{3/2}}.$$

Finally, this implies, via Lemma 18 in the appendix, that there is some unitary matrix  $O \in \mathbb{R}^{k \times k}$  so that

$$\|U^{(t)}O - \tilde{Q}_t\| \leq \frac{16e}{C_1 k^{3/2}},$$

and using the fact that  $U^{(t)}$  and  $\tilde{Q}_t$  have rank at most  $k$ , we have that

$$\|U^{(t)}O - \tilde{Q}_t\|_F \leq \frac{16\sqrt{2}e}{C_1 k}. \quad (21)$$

As in Algorithm 1, let  $B$  be a random orthogonal matrix, and let  $\bar{Q}_t$  be the truncation

$$\bar{Q}_t = \text{TRUNCATE} \left( \tilde{Q}_t B, 8\sqrt{\frac{\mu^* \log(n)}{n}} \right).$$

The reason for the random rotation is that while  $U^{(t)}O$  is reasonably incoherent (because  $U^{(t)}$  is),  $U^{(t)}OB$  is, with high probability, even more incoherent. More precisely, as in [Hardt \(2013b\)](#), we have

$$\mathbb{P} \left\{ \|U^{(t)}OB\|_\infty > 8\sqrt{\frac{\mu^* \log(n)}{n}} \right\} \leq \frac{1}{n^2}, \quad (22)$$

where the probability is over the choice of  $B$ . Suppose that the favorable case in (22) occurs, so that  $\|U^{(t)}OB\|_\infty \leq 8\sqrt{\mu^* \log(n)/n}$ . In the Frobenius norm,  $\bar{Q}_t$  is the projection of  $\tilde{Q}_t$  onto the (entrywise)  $\ell_\infty$ -ball of radius  $8\sqrt{\mu^* \log(n)/n}$  in  $\mathbb{R}^{n \times d_t}$ . Thus,

$$\|\bar{Q}_t - \tilde{Q}_t B\|_F \leq \|X - \tilde{Q}_t B\|_F$$

for any  $X$  in this scaled  $\ell_\infty$ -ball, and in particular

$$\|\bar{Q}_t - \tilde{Q}_t B\|_F \leq \|U^{(t)}OB - \tilde{Q}_t B\|_F.$$

Thus, (21) implies that

$$\begin{aligned} \|U^{(t)}OB - \bar{Q}_t\|_F &\leq \|U^{(t)}OB - \tilde{Q}_t B\|_F + \|\tilde{Q}_t B - \bar{Q}_t\|_F \\ &\leq 2 \|U^{(t)}OB - \tilde{Q}_t B\|_F = 2 \|U^{(t)}O - \tilde{Q}_t\|_F \leq \frac{32\sqrt{2}e}{C_1 k}. \end{aligned} \quad (23)$$

Next, we consider the matrix  $W_t = \text{QR}([X_{t-1} | \bar{Q}_t])$ . Because  $X_{t-1}$  has orthonormal columns, this matrix has the form  $W_t = [X_{t-1} | P_t]$ , where  $P_t \in \mathbb{R}^{n \times d_t}$  has orthonormal columns,  $P_t \perp X_{t-1}$ , and

$$\mathcal{R}(P_t) = \mathcal{R}((I - X_{t-1}X_{t-1}^T)\bar{Q}_t) = \mathcal{R}(Z_t),$$

where we define  $Z_t := (I - X_{t-1}X_{t-1}^T)\bar{Q}_t$  to be the projection of  $\bar{Q}_t$  onto  $\mathcal{R}(X_{t-1})^\perp$ . Because  $\bar{Q}_t$  is close to  $U^{(t)}OB$ , and  $X_{t-1}$  is close to  $U^{(<t)}$ ,  $Z_t$  is close to  $U^{(t)}OB$ . More precisely,

$$\begin{aligned}
 \|Z_t - U^{(t)}OB\| &\leq \left\| (I - X_{t-1}X_{t-1}^T)(\bar{Q}_t - U^{(t)}OB) \right\| + \left\| X_{t-1}X_{t-1}^T U^{(t)}OB \right\| && \text{by the triangle inequality} \\
 &\leq \left\| \bar{Q}_t - U^{(t)}OB \right\|_F + \sin \theta(X_{t-1}, U^{(<t)}) \\
 &\leq \frac{32\sqrt{2}e}{C_1k} + \frac{1}{k^4} \left( \frac{\sigma_{r_{t-1}+1} + \varepsilon \|M\|}{\sigma_{r_{t-1}}} \right) && \text{by (23) and (H2)} \\
 &\leq \frac{32\sqrt{2}e}{C_1k} + \frac{1}{k^4} \left( \frac{2\sigma_{r_{t-1}+1}}{\sigma_{r_{t-1}}} \right) && \text{by (17)} \\
 &\leq \frac{64\sqrt{2}e}{C_1k} && \text{for sufficiently large } k.
 \end{aligned}$$

Further, the Gram-Schmidt process gives a decomposition

$$P_t R = Z_t,$$

where the triangular matrix  $R$  has the same spectrum as  $Z_t$ . In particular,

$$\|R^{-1}\| = \frac{1}{\sigma_{\min}(Z_t)} \leq \frac{1}{\|U^{(t)}\| - \frac{64\sqrt{2}e}{C_1k}} \leq 2$$

for sufficiently large  $C_1$ . Thus,

$$\begin{aligned}
 \sin \theta(U^{(\leq t)}, P_t) &= \left\| (U_{\perp}^{(\leq t)})^T P_t \right\| \\
 &= \left\| (U_{\perp}^{(\leq t)})^T Z_t R^{-1} \right\| \\
 &\leq 2 \left\| (U_{\perp}^{(\leq t)})^T Z_t \right\| \\
 &\leq 2 \left\| (U_{\perp}^{(\leq t)})^T U^{(t)}OB \right\| + 2 \left\| (U_{\perp}^{(\leq t)})^T (Z_t - U^{(t)}OB) \right\| \\
 &= 2 \left\| (U_{\perp}^{(\leq t)})^T (Z_t - U^{(t)}OB) \right\| \\
 &\leq \frac{128\sqrt{2}e}{C_1k}, \tag{24}
 \end{aligned}$$

where above we used that  $(U_{\perp}^{(\leq t)})^T U^{(t)} = 0$ . Next,

$$\begin{aligned}
 \max_i \|e_i^T P_t\|_2 &\leq \max_i \|e_i^T Z_t\|_2 \|R^{-1}\| \\
 &\leq 2 \left( \max_i \|e_i^T \bar{Q}_t\|_2 + \max_i \|e_i^T X_{t-1}X_{t-1}^T \bar{Q}_t\|_2 \right) \\
 &\leq 2 \left( \max_i \|e_i^T \bar{Q}_t\|_2 + \max_i \|e_i^T X_{t-1}\|_2 \left( \left\| X_{t-1}^T U^{(t)}OB \right\| + \left\| X_{t-1}^T (U^{(t)}OB - \bar{Q}_t) \right\| \right) \right) \\
 &\leq 2 \left( \max_i \|e_i^T \bar{Q}_t\|_2 + \max_i \|e_i^T X_{t-1}\|_2 \left( \left\| X_{t-1}^T U^{(t)} \right\| + \left\| U^{(t)}OB - \bar{Q}_t \right\| \right) \right) \\
 &\leq 16\sqrt{\frac{k\mu^* \log(n)}{n}} + 2\sqrt{\frac{k\mu_{t-1}}{n}} \left( \frac{2}{k^4} + \frac{32\sqrt{2}e}{C_1k} \right),
 \end{aligned}$$

where we have used the definition of  $\overline{Q}_t$ , the incoherence of  $X_{t-1}$ , and the computations above in the final line. Thus,

$$\max_i \|e_i^T P_t\|_2 \leq \sqrt{\frac{k}{n}} \left( 16\sqrt{\mu^* \log(n)} + \frac{C_5 \sqrt{\mu_{t-1}}}{k} \right) \quad (25)$$

for some constant  $C_5$ . Thus, when the conclusions of Lemma 6 hold,  $P_t$  is both close to  $U^{(t)}$  and incoherent. By induction, the same is true for  $W_t$ . Indeed, if  $t = 1$ , then  $P_t = W_t$ , and we are done. If  $t \geq 2$ , then we have

$$\sin \theta(W_t, U^{(\leq t)}) \leq \sin \theta(X_{t-1}, U^{(\leq t-1)}) + \sin \theta(P_t, U^{(t)}).$$

Then, the inductive hypothesis (H1) and our conclusion (24) imply that

$$\sin \theta(W_t, U^{(\leq t)}) \leq \frac{1}{k}$$

for suitably large  $C_0, C_1$ . Finally, (25), along with the inductive hypothesis (H3) implies that

$$\begin{aligned} \max_i \|e_i^T W_t\|_2 &\leq \max_i \|e_i^T X_{t-1}\|_2 + \max_i \|e_i^T P_t\|_2 \\ &\leq \sqrt{\frac{k}{n}} \left( \sqrt{\mu_{t-1}} \left( 1 + \frac{C_5}{k} \right) + 16\sqrt{\mu^* \log(n)} \right) \leq \sqrt{\frac{k\mu_t}{n}}. \end{aligned}$$

We remark that this last computation is the only reason we need  $\sin \theta(P_t, U^{(t)}) \lesssim 1/k$ , rather than bounded by  $1/4$ ; eventually, we will iterate and have

$$\sqrt{\mu_T} \leq \sqrt{\mu_0} \left( 1 + \frac{C_5}{k} \right)^T + 16T\sqrt{\mu^* \log(n)} \leq e^{C_5} \sqrt{\mu_0} + 16T\sqrt{\mu^* \log(n)},$$

and we need that  $(1 + \frac{C_5}{k})^T \leq e^{C_5}$  is bounded by a constant (rather than exponential in  $T$ ).

Finally, we have shown that with probability  $1 - 1/n^2$  (that is, in the case that Lemma 6 holds and SUBSIT works), all of the conclusions of Lemma 4 hold as well. This completes the proof of Lemma 4.

## Appendix B. Proof of Lemma 5

In the proof of Lemma 5 we will need an explicit description of the subroutine SMOOTHQR that we include in Algorithm 3.

**Input:** Matrix  $S \in \mathbb{R}^{n \times k}$ , parameters  $\mu, \zeta > 0$ .

```

29  $X \leftarrow \text{QR}(S), H \leftarrow 0$   $\sigma \leftarrow \zeta \|S\|/n$ . while  $\mu(R) > \mu$  and  $\sigma \leq \|S\|$  do
30 |  $R \leftarrow \text{QR}(S + H)$  where  $H \sim \text{N}(0, \sigma^2/n)$   $\sigma \leftarrow 2\sigma$ 
31 end
    
```

**Output:** Matrix  $R$

**Algorithm 3:** SMOOTHQR ( $S, \zeta, \mu$ ) (Smooth Orthonormalization)

To prove Lemma 5, we will induct on the iteration  $\ell$  in S-M-ALTLS (Algorithm 2). Let  $R_\ell$  denote the approximation in iteration  $\ell$ . Thus,  $R_0 = X_{t-1}$ . Above, we are suppressing the dependence of  $R_\ell$  on the epoch number  $t$ , and in general, for this section we will drop the subscripts  $t$  when there is no ambiguity. We'll use the shorthand

$$\Theta_\ell^j = \theta(R_\ell^{(\leq j)}, U^{(\leq j)})$$

and

$$E_\ell^j = (I - R_\ell^{(\leq j)})(R_\ell^{(\leq j)})^T U^{(\leq j)},$$

so that  $\|E_\ell^j\| = \sin(\Theta_\ell^j)$ . Recall the definition (9) that  $\gamma^* = \min\{\gamma, \gamma_k\}$ . Notice that this choice ensures that  $\gamma^* \leq \gamma_{r_j}$  for all choices of  $j$ , including the case of  $j = t$ , in the final epoch of SOFTDEFLATE, when  $r_t = k$ .

We will maintain the following inductive hypothesis:

$$\sigma_{r_j} \tan \Theta_\ell^j \leq \max \left\{ \left( \frac{2e\sigma_{r_t}}{k} \right) \exp(-\gamma^* \ell / 2), \frac{\sigma_{r_{t+1}} + \varepsilon \|M\|}{2eC_0 k^4} \right\} =: \nu_\ell \quad \forall j \leq t. \quad (\text{J1})$$

Above, the tangent of the principal angle obeys

$$\|E_\ell^j\| \leq \tan \Theta_\ell^j = \frac{\|E_{\ell-1}^j\|}{\sqrt{1 - \|E_{\ell-1}^j\|^2}} \leq 2 \|E_{\ell-1}^j\|, \quad (\text{26})$$

whenever  $\|E_{\ell-1}^j\| \leq 1/4$ . We will also maintain the inductive hypothesis

$$\max_i \|e_i^T R_\ell\|_2 \leq \sqrt{\frac{k\mu_t}{n}}. \quad (\text{J2})$$

To establish the base case of (J1) for  $j = t$ , we have

$$\sigma_{r_t} \sin \theta(W_t, U^{(\leq t)}) \leq \frac{\sigma_{r_t}}{k},$$

by conclusion (13) of Lemma 4, and hence by (26),

$$\sigma_{r_t} \tan \theta(W_t, U^{(\leq t)}) \leq \frac{2\sigma_{r_t}}{k}.$$

If  $t = 1$ , then  $W_t = R_0$ , and we are done with the base case for (J1); if  $t \geq 2$ , then for  $j \leq t - 1$ , we have

$$R_0^{(\leq j)} = X_{t-1}^{(\leq j)}.$$

Thus, for  $j \leq t - 1$ , (J1) is implied by (26) again, along with the fact that

$$\sigma_{r_j} \sin \theta(X_{t-1}^{(\leq j)}, U^{(\leq j)}) \leq \frac{1}{k^4} (\sigma_{r_{t-1+1}} + \varepsilon \|M\|) \leq \frac{e\sigma_{r_t} + \varepsilon \|M\|}{k^4} \leq \frac{2e\sigma_{r_t}}{k^4},$$

which is the (outer) inductive hypothesis (H1), followed by the conclusions (11) and (12) from Lemma 4. This establishes the base case for (J1). The base case for (J2) follows from the conclusion (13) of Lemma 4 directly.

**Input:** Number of iterations  $L \in \mathbb{N}$ , symmetric matrix  $A \in \mathbb{R}^{n \times n}$ , initial matrix  $R_0 \in \mathbb{R}^{n \times r}$ .

32 **for**  $\ell = 1, \dots, L$  **do**

33      $S_\ell \leftarrow AR_{\ell-1} + \tilde{G}_\ell$

34      $R_\ell \leftarrow \text{QR}(S_\ell)$

35 **end**

**Output:** Pair of matrices  $(R_L, S_L)$

**Algorithm 4:** NSI( $A, R_0, L$ ) (Noisy Subspace Iteration)

Having established (J1), (J2) for  $\ell = 0$ , we now suppose that they hold for  $\ell - 1$  and consider step  $\ell$ . Notice that, by running SMOOTHQR with parameter  $\mu = \mu_t$ , we automatically ensure (J2) for the next round of induction, and so our next goal is to establish (J1). For this, we need to go deeper into the workings of S-M-ALTLS. The analysis of S-M-ALTLS in Hardt (2013b) is based on an analysis of NSI, given in Algorithm 4. We may view S-M-ALTLS as a special case of NSI. More precisely, let  $H_\ell$  be the noise matrix added from SMOOTHQR in the  $\ell$ 'th iteration of S-M-ALTLS, and define  $G_\ell^{(s)}$  to be

$$G_\ell^{(s)} = \operatorname{argmin}_{S \in \mathbb{R}^{n \times r}} \left\| P_{\Omega_\ell^{(s)}}(A - R_{\ell-1}S^T) \right\|_F^2 - AR_{\ell-1}, \quad (27)$$

and let

$$G_\ell = \operatorname{median}_s(G_\ell^{(s)}).$$

Then we may write  $R_\ell$ , the  $\ell$ 'th iterate in S-M-ALTLS, as

$$R_\ell = \text{SMOOTHQR}(AR_{\ell-1} + G_\ell) = \text{QR}(AR_{\ell-1} + G_\ell + H_\ell) =: \text{QR}(AR_{\ell-1} + \tilde{G}_\ell).$$

That is,  $R_\ell$  is also the  $\ell$ 'th iterate in NSI, when the noise matrices are  $\tilde{G}_\ell = G_\ell + H_\ell$ . We will take this view going forward, and analyze S-M-ALTLS as a special case of NSI. We have the following theorem, which is given in (Hardt, 2013b, Lemma 3.4).

**Theorem 10** *Let  $\tilde{G}_\ell = G_\ell + H_\ell$  be as above. Let  $j \leq t$  and suppose that  $\|E_{\ell-1}^j\| \leq \frac{1}{4}$  and that*

$$\|\tilde{G}_\ell\| \leq \frac{\sigma_{r_j} \gamma_{r_j}}{32}.$$

*Then the next iterate  $R_\ell$  of NSI satisfies*

$$\tan \theta(U^{(\leq j)}, R_{\ell-1}) \leq \max \left\{ \frac{8 \|\tilde{G}_\ell\|}{\sigma_{r_j} \gamma_{r_j}}, \tan \theta(U^{(\leq j)}, R_{\ell-1}) \exp(-\gamma_{r_j}/2) \right\}.$$

To use Theorem 10, we must understand the noise matrices  $\tilde{G}_\ell = G_\ell + H_\ell$ . We begin with  $G_\ell$ .

**Lemma 11 (Noise term  $G_\ell$  in NSI)** *There is a constant  $C$  so that the following holds. Fix  $\ell$  and suppose that (J2) holds for  $\ell - 1$ : that is,  $\mu(R_{\ell-1}) \leq \mu_t$ . Let  $0 < \delta < 1/2$ , and suppose that the samples  $\Omega_t^i$  for S-M-ALTLS are sampled independently with probability*

$$p_t' \geq CL_t s_{\max} \frac{k \mu_t \log(n)}{\delta^2 n},$$



where  $L_t$  is the number of iterations of S-M-ALTLS, and  $s_{\max} \geq C \log(n)$  is the number of trials each iteration of S-M-ALTLS performs before taking a median. Then with probability at least  $1 - 1/n^5$  over the choice of  $\Omega'_t$ , the noise matrix  $G_\ell$  satisfies

$$\|G_\ell\|_F \leq \delta \left( \|N_t\|_F + \sum_{j=1}^n \|E_{\ell-1}^j\| \|M^{(j)}\|_F \right) =: \omega_{\ell-1}$$

and for all  $i \in [n]$ ,

$$\|e_i^T G_\ell\|_2 \leq \delta \left( \|e_i^T N_t\|_2 + \sum_{j=1}^n \|E_{\ell-1}^j\| \|e_i^T M^{(j)}\|_2 \right) =: \omega_{\ell-1}^{(i)}.$$

The proof of Lemma 11 is similar to the analysis in [Hardt \(2013b\)](#). For completeness, we include the proof in Appendix C. Using the inductive hypothesis (J1), and the fact that  $\|M^{(j)}\|_F \leq \sqrt{k}\sigma_{r_j}$ ,

$$\omega_{\ell-1} \leq \delta \left( \sum_j \|E_{\ell-1}^j\| (\sqrt{k}\sigma_{r_j}) + \sqrt{k}\sigma_{r_{t+1}} + \Delta \right) \leq \delta (t\sqrt{k}\nu_{\ell-1} + \sqrt{k}\sigma_{r_{t+1}} + \Delta).$$

We will choose

$$\delta = \frac{\gamma^*}{4eC_0C_3k^4} \min \left\{ \frac{1}{\sqrt{k}}, \frac{\varepsilon \|M\|}{\Delta} \right\}, \quad (28)$$

for a constant  $C_3$  to be chosen sufficiently large. Observe that with this choice of  $\delta$ , the requirement on  $p'_t$  in Lemma 11 is implied by the requirement on  $p'_t$  in the statement in Lemma 5. Then the choice of  $\delta$  implies

$$\|G_\ell\|_F \leq \omega_{\ell-1} \leq \frac{\gamma^*}{4eC_0C_3k^4} (t\nu_{\ell-1} + \sigma_{r_{t+1}} + \varepsilon \|M\|) \leq \frac{\gamma^*4eC_0}{43C_0C_3} \nu_{\ell-1} \leq \frac{\gamma^*}{C_3} \nu_{\ell-1}. \quad (29)$$

Now, we turn to the noise term  $H_\ell$  added by SMOOTHQR. For a matrix  $G \in \mathbb{R}^{n \times k}$  (not necessarily orthonormal), we will define

$$\rho(A) := \frac{n}{k} \max_{i \in [n]} \|e_i^T G\|_2^2.$$

Our analysis of  $H_\ell$  relies on the following lemma from [Hardt \(2013b\)](#).

**Lemma 12 (Lemma 5.4 in [Hardt \(2013b\)](#))** *Let  $\tau > 0$  and suppose that  $r_t = o(n/\log(n))$ . There is an absolute constant  $C$  so that the following claim holds. Let  $G \in \mathbb{R}^{n \times r_t}$ , and let  $R \in \mathbb{R}^{n \times r_t}$  be an orthonormal matrix, and let  $\nu \in \mathbb{R}$  so that  $\nu \geq \max \|G\|, \|N_t R\|$ . Assume that*

$$\mu_t \geq 2\mu(U) + \frac{C}{\tau^2} \left( \frac{\rho(G) + \mu(U) \|(U^{(\leq t)})^T G\|^2 + \rho(N_t R)}{\nu^2} + \log(n) \right).$$

*Then, for every  $\zeta \leq \tau\nu$  satisfying  $\log(n/\zeta) \leq n$ , we have with probability at least  $1 - 1/n^4$  that the algorithm SMOOTHQR ( $AR + G, \zeta, \mu_t$ ) terminates in  $\log(n/\zeta)$  iterations, and the output  $R'$  satisfies  $\mu(R') \leq \mu_t$ . Further, the final noise matrix  $H$  added by SMOOTHQR satisfies  $\|H\| \leq \tau\nu$ .*

We will apply Lemma 12 to our situation.

**Lemma 13 (Noise term  $H_\ell$  in NSI added by SMOOTHQR)** *Suppose that  $k = o(n/\log(n))$ . There is a constant  $C_2$  so that the following holds. Suppose that*

$$\mu_t \geq \frac{C_2}{(\gamma^*)^2} \left( \mu^* \left( k + \left( \frac{k^4 \|N\|_F}{\varepsilon \|M\|} \right)^2 \right) + \log(n) \right).$$

*Suppose that the favorable conclusion of Lemma 11 occurs. Choose  $\zeta = \varepsilon s_0 k^{-5}$ , as in Algorithm 1. Then, with probability at least  $1 - 1/n^4$  over the randomness of SMOOTHQR, the output  $R_\ell$  of SMOOTHQR( $AR_{\ell-1} + G_\ell, \zeta, \mu_t$ ) satisfies*

$$\mu(R_\ell) \leq \mu_t,$$

*and the number of iterations is  $O(\log(n/(\varepsilon \|M\|)))$ . Further, the noise matrix  $H_\ell$  satisfies*

$$\|H_\ell\| \leq \frac{\gamma^* \nu_{\ell-1}}{C_3}.$$

**Proof** We apply Lemma 12 with  $G = G_\ell, R = R_{\ell-1}$ , and  $\nu = \nu_\ell$ , and

$$\tau = \frac{\gamma^*}{C_3}. \tag{30}$$

First, we observe that the choice of  $\zeta = \varepsilon s_0 k^{-5} \leq \varepsilon \|M\| \gamma^* k^{-4} \leq \tau \nu_{\ell-1}$  indeed satisfies the requirements of Lemma 12. Next, we verify that  $\max\{\|G_\ell\|, \|N_t R_{\ell-1}\|\} \leq \nu_{\ell-1}$ . Indeed, from (29),

$$\|G_\ell\| \leq \omega_{\ell-1} \leq \frac{\gamma^*}{C_3} \nu_{\ell-1} \leq \nu_{\ell-1}.$$

Further, we have

$$\|N_t R_{\ell-1}\| \leq \sigma_{r_t} \sin \theta(U^{(\leq t)}, R_{\ell-1}) \leq \nu_{\ell-1}$$

by the inductive hypothesis (J1) for  $j = t$ .

Next, we compute the parameters that show up in Lemma 12. From Lemma 11, we have

$$\rho(G_\ell) \leq \frac{n}{r_t} \max_i \left( \omega_{\ell-1}^{(i)} \right)^2$$

and

$$\mu(U) \left\| U^{(\leq t)} G_\ell \right\|^2 \leq \|G_\ell\|^2 \leq \mu^* \omega_{\ell-1}^2.$$

We also have

$$\begin{aligned}
 \rho(N_t R_{\ell-1}) &= \frac{n}{r_t} \max_i \|e_i^T N_t R_{\ell-1}\|_2^2 \\
 &\leq \frac{n}{r_t} \left( \max_i \|e_i^T U^{(t:k)}\|_2 \sigma_{r_t} \|(U^{(t:k)})^T R_{\ell-1}\|_2 + \max_i \|e_i^T N\|_2 \|R_{\ell-1}\|_2 \right)^2 \\
 &\leq \frac{n}{r_t} \left( \sqrt{\frac{k\mu(U)}{n}} \sigma_{r_t} \|E_{\ell-1}^t\| + \sqrt{\frac{\mu_N \|N\|_F}{n}} \right)^2 \\
 &\leq 2\mu^* \left( \frac{k}{r_t} \sigma_{r_t}^2 \|E_{\ell-1}^t\|^2 + \frac{\|N\|_F^2}{r_t} \right) \\
 &\leq 2\mu^* \left( \frac{k\nu_{\ell-1}^2}{r_t} + \frac{\|N\|_F^2}{r_t} \right),
 \end{aligned}$$

where we have used the inductive hypothesis (J1) in the final line. Then, the requirement of Lemma 12 on  $\mu_t$  reads

$$\mu_t \geq 2\mu^* + \frac{C}{\tau^2} \left( \frac{\frac{n}{r_t} \max_i (\omega_{\ell-1}^{(i)})^2 + \mu^* \omega_{\ell-1}^2 + 2\mu^* \left( \frac{k}{r_t} \nu_{\ell-1}^2 + \frac{\|N\|_F^2}{r_t} \right)}{\nu_{\ell-1}^2} + \log(n) \right).$$

We have, for all  $i$ ,

$$\begin{aligned}
 \frac{\omega_{\ell-1}^{(i)}}{\omega_{\ell-1}} &= \frac{\|e_i^T N_t\|_2 + \sum_{j=1}^t \|E_{\ell-1}^j\| \|e_i^T M^{(j)}\|_2}{\|N_t\|_F + \sum_{j=1}^t \|E_{\ell-1}^j\| \|M^{(j)}\|_F} \\
 &\leq \frac{\sigma_{r_t} \sqrt{\Delta^2 \mu^*/n} + \sum_{j=1}^t \|E_{\ell-1}^j\| \sigma_{r_j} \sqrt{k\mu^*/n}}{\|N_t\|_F + \sum_{j=1}^t \|E_{\ell-1}^j\| \|M^{(j)}\|_F} \\
 &\leq \frac{\|N_t\|_F \sqrt{\Delta^2 \mu^*/n} + \sum_{j=1}^t \|E_{\ell-1}^j\| \|M^{(j)}\|_F \sqrt{k\mu^*/n}}{\|N_t\|_F + \sum_{j=1}^t \|E_{\ell-1}^j\| \|M^{(j)}\|_F} \\
 &= \sqrt{\frac{\mu^*}{n}} (\sqrt{k} + \Delta).
 \end{aligned}$$

We may simplify and bound the requirement on  $\mu_t$  as

$$\begin{aligned}
 & 2\mu^* + \frac{C}{\tau^2} \left( \frac{\frac{n}{r_t} \max_i (\omega_{\ell-1}^{(i)})^2 + \mu^* \omega_{\ell-1}^2 + 2\mu^* \left( \frac{k}{r_t} \nu_{\ell-1}^2 + \frac{\|N\|_F^2}{r_t} \right)}{\nu_{\ell-1}^2} + \log(n) \right) \\
 & \leq 2\mu^* + \frac{C}{\tau^2} \left( \frac{\frac{n}{r_t} \max_i (\omega_{\ell-1}^{(i)})^2 (\gamma^*)^2}{C_3^2 \omega_{\ell-1}^2} + \frac{\mu^* (\gamma^*)^2 \omega_{\ell-1}^2}{C_3^2 \omega_{\ell-1}^2} + \frac{2\mu^* \left( \frac{k}{r_t} \nu_{\ell-1}^2 + \frac{\|N\|_F^2}{r_t} \right)}{\nu_{\ell-1}^2} + \log(n) \right) \\
 & \quad \text{using } \nu_{\ell-1} \geq C_3 \omega_{\ell-1} / \gamma^*, \text{ by (29)} \\
 & \leq 2\mu^* + \frac{C}{\tau^2} \left( \frac{\frac{k+\Delta^2}{r_t} \mu^* (\gamma^*)^2}{C_3^2} + \frac{\mu^* (\gamma^*)^2}{C_3^2} + 2\mu^* \left( \frac{k}{r_t} + \frac{\|N\|_F^2}{r_t \nu_{\ell-1}^2} \right) + \log(n) \right) \\
 & \quad \text{by the bound on } \omega_{\ell-1}^{(i)} / \omega_{\ell-1}, \text{ above} \\
 & \leq \frac{C' \mu^*}{(\gamma^*)^2} \left( \frac{k}{r_t} + \frac{\|N\|_F^2}{\nu_{\ell-1}^2} \right) + \frac{C_3^2 \log(n)}{(\gamma^*)^2} \text{ by the definition of } \tau \text{ and gathering terms} \\
 & \leq \frac{C_2 \mu^*}{(\gamma^*)^2} \left( \frac{k}{r_t} + \frac{k^8 \|N\|_F^2}{\varepsilon^2 \|M\|^2} \right) + \frac{C_3^2 \log(n)}{(\gamma^*)^2} \quad \text{by the fact that } \nu_{\ell-1} \geq \frac{\varepsilon \|M\|}{2eC_0 k^4}.
 \end{aligned}$$

for some constant  $C_2$ , which was the requirement in the statement of the lemma. Thus, as long as the hypotheses of the current lemma hold, Lemma 12 implies that with probability at least  $1 - 1/n^4$ ,

$$\|H_\ell\| \leq \tau \nu_{\ell-1} = \frac{\gamma^* \nu_{\ell-1}}{C_3}.$$

This completes the proof of Lemma 13. ■

Thus, using the inductive hypothesis (J2), Lemmas 11 and 13 imply that as long as the requirements on  $p'_t$  and  $\mu_t$  in the statements of those lemmas are satisfied (which they are, by the choices in Lemma 5), with probability at least  $1 - 2/n^4$  the noise matrices  $\tilde{G}_\ell$  satisfy

$$\|\tilde{G}_\ell\| \leq \|G_\ell\| + \|H_\ell\| \leq \omega_{\ell-1} + \frac{\gamma^* \nu_{\ell-1}}{C_3} \leq \frac{2\gamma^* \nu_{\ell-1}}{C_3},$$

using (29) in the final inequality. Now, we wish to apply Theorem 10. The hypothesis (J1), along with the conclusion (11) from Lemma 4, immediately implies that

$$\|E_{\ell-1}^t\| \leq \frac{1}{k}$$

for all  $j \leq t$ , and so in particular the first requirement of Theorem 10 is satisfied. To satisfy the second requirement of Theorem 10, we must show that

$$\|\tilde{G}_\ell\| \leq \sigma_{r_j} \gamma_{r_j} / 32,$$

for which it suffices to show that

$$\frac{2\gamma^* \nu_{\ell-1}}{C_3} \leq \sigma_{r_j} \gamma_{r_j} / 32. \quad (31)$$

From the definition of  $\nu_{\ell-1}$ , and the fact that  $\gamma^* \leq \gamma_{r_j}$ , we see that (31) is satisfied for a sufficiently large choice of  $C_3$ . Then Theorem 10 implies that with probability at least  $1 - 2/n^4$ , for any fixed  $j$ , we have

$$\begin{aligned} \sigma_{r_j} \tan \Theta_{\ell}^j &\leq \sigma_{r_j} \max \left\{ \frac{8 \|\tilde{G}_{\ell}\|}{\sigma_{r_j} \gamma_{r_j}}, \tan \Theta_{\ell-1}^j \exp(-\gamma_{r_j}/2) \right\} \\ &\leq \max \left\{ \frac{16\nu_{\ell-1}\gamma^*}{C_3\gamma_{r_j}}, \nu_{\ell-1} \exp(-\gamma_{r_j}/2) \right\} \text{ by (J1) and (29)} \\ &\leq \nu_{\ell-1} \exp(-\gamma^*/2) \\ &\leq \nu_{\ell} \end{aligned}$$

provided  $C_3$  is suitably large. A union bound over all  $j$  establishes (J1) for the next iteration of S-M-ALTLS. After another union bound over

$$L_t = \frac{C}{\gamma^*} \log \left( k \cdot \frac{\sigma_{r_t}}{\sigma_{r_{t+1}} + \varepsilon \|M\|} \right)$$

steps of S-M-ALTLS, for some constant  $C$  depending on  $C_0$ , we conclude that with probability at least  $1 - 1/n^2$ , for all  $j$ ,

$$\sigma_{r_j} \sin \theta(R_{\ell-1}^{(\leq j)}, U^{(\leq j)}) \leq \sigma_{r_j} \tan \theta(R_{\ell-1}^{(\leq j)}, U^{(\leq j)}) \leq \frac{\sigma_{r_{t+1}} + \varepsilon \|M\|}{2eC_0k^4}.$$

To establish the second conclusion, we note that we have already conditioned on the event that (29) holds, and so we have

$$\begin{aligned} \left\| M^{(\leq t)} - X_t Y_t^T \right\| &= \left\| \Pi_X N_t + \Pi_{X_{\perp}} M^{(\leq t)} + X_t (A X_t - Y_t)^T \right\| \\ &\leq \left\| \Pi_X N_t \right\| + \left\| \Pi_{X_{\perp}} M^{(\leq t)} \right\| + \left\| X_t (A X_t - Y_t) \right\| \\ &\leq \sigma_{r_{t+1}} \sin \theta(X_t, U^{(\leq t)}) + e \sum_{j=1}^t \sigma_{r_j} \sin \theta(X_t^{(\leq j)}, U^{(\leq j)}) + \|G_L\| \\ &\leq k e \frac{\sigma_{r_{t+1}} + \varepsilon \|M\|}{2eC_0k^4} + \frac{\gamma^*}{2eC_0C_3k^4} (\sigma_{r_{t+1}} + \varepsilon \|M\|) \quad \text{by (29) and the definition of } \nu_L \\ &\leq \frac{\sigma_{r_{t+1}} + \varepsilon \|M\|}{C_0k^3}. \end{aligned}$$

Above, we used the inequality

$$\begin{aligned} \left\| \Pi_{X_{\perp}} M^{(\leq t)} \right\| &= \left\| \sum_{j=1}^t \Pi_{X_{\perp}} M^{(j)} \right\| \leq \sum_{j=1}^t \left\| \Pi_{X_{\perp}} M^{(j)} \right\| \leq \sum_{j=1}^t \sigma_{r_{j-1}+1} \left\| \Pi_{X_{\perp}} U^{(j)} \right\| \\ &\leq \sum_{j=1}^t \sigma_{r_{j-1}+1} \left\| \Pi_{X_{\perp}^{(\leq j)}} U^{(\leq j)} \right\| \leq \sum_{j=1}^t e \sigma_{r_j} \sin \theta(X^{(\leq j)}, U^{(\leq j)}), \end{aligned}$$

using (12) in the final inequality. Finally, the third conclusion, that (H3) holds, follows from the definition of SMOOTHQR.

### Appendix C. Proof of Lemma 11

In this section, we prove Lemma 11, which bounds the noise matrices  $G_\ell^{(s)}$  and which we needed in the proof of Lemma 5. The proof of Lemma 11 is similar to the analysis in [Hardt \(2013b\)](#), Lemmas 4.2 and 4.3. For completeness, we include the details here. As per Section D, we assume that sets  $\Omega_\ell^{(s)}$  are independent random sets, which include each index independently with probability

$$p' := \frac{p'_t}{s_{\max} L_t}.$$

Consider each noise matrix  $G_\ell^{(s)}$ , as in (27). In Lemma 4.2 in [Hardt \(2013b\)](#), an explicit expression for  $G_\ell^{(s)}$  is derived:

**Proposition 14** *Let  $G_\ell^{(s)}$  be as in (27). Then we have*

$$G_\ell^{(s)} = (G_\ell^{(s)})^M + (G_\ell^{(s)})^N,$$

where

$$e_i^T (G_\ell^{(s)})^M = e_i^T M_t (I - R_{\ell-1} R_{\ell-1}^T) P_i^{(s)} R_{\ell-1} (B_i^{(s)})^{-1}.$$

and

$$e_i^T (G_\ell^{(s)})^N = e_i^T \left( N_t P_i^{(s)} R_{\ell-1} (B_i^{(s)})^{-1} - N_t R_{\ell-1} \right).$$

Above,  $P_i^{(s)}$  is the projection onto the coordinates  $j$  so that  $(i, j) \in \Omega_\ell^{(s)}$ , and

$$B_i^{(s)} = R_{\ell-1}^T P_i^{(s)} R_{\ell-1}.$$

We first bound the expression for  $(G_\ell^{(s)})^M$  in terms of the decomposition in Proposition 14. Let

$$D_{\ell-1}^j = (I - R_{\ell-1} R_{\ell-1}^T) U^{(j)}.$$

Thus,  $D_{\ell-1}^j$  is similar to  $E_{\ell-1}^j$ , and more precisely we have

$$\|D_{\ell-1}^j\| \leq \|E_{\ell-1}^j\|. \quad (32)$$

To see (32), observe that (dropping the  $\ell$  subscripts for readability)

$$\begin{aligned} \|E^j\| &= \max_{\|x\|_2=1, \|y\|_2=1} x^T E^j y \\ &= \max_{x, y} x^T \left[ \frac{(R^{(j+1:t)})^T U^{(<j)}}{(R_\perp)^T U^{(<j)}} \mid \frac{(R^{(j+1:t)})^T U^{(j)}}{(R_\perp)^T U^{(j)}} \right] y \\ &\geq \max_{x=(0, x'), y=(0, y')} (x')^T (R_\perp)^T U^{(j)} y' \\ &= \|D^j\| \end{aligned}$$

First, we observe that with very high probability,  $B_i^{(s)}$  is close to the identity.

**Claim 1** *There is a constant  $C$  so that the following holds. Suppose that  $p' \geq Ck\mu_t \log(n)/(n\delta^2)$ . Then*

$$\mathbb{P} \left\{ \lambda_{\min}(B_i^{(s)}) \leq 1 - \delta/2 \text{ or } \lambda_{\max}(B_i^{(s)}) \geq 1 + \delta/2 \right\} \leq 1/n^5.$$

**Proof** We write

$$B_i^{(s)} = R_{\ell-1}^T P_i^{(s)} R_{\ell-1} = \sum_{r=1}^n \frac{1}{p'} \xi_r (R_{\ell-1}^T e_r) (e_r^T R_{\ell-1}),$$

where  $\xi_r$  is 1 with probability  $p'$  and 0 otherwise. We apply the Matrix Chernoff bound (Lemma 22); we have

$$\left\| \frac{1}{p'} \xi_r (R_{\ell-1}^T e_r) (e_r^T R_{\ell-1}) \right\| \leq \frac{\|e_r^T R_{\ell-1}\|_2^2}{p'} \leq \frac{\mu_t k}{np'} \quad \text{almost surely,}$$

and  $\lambda_{\min}(\mathbb{E}B_i^{(s)}) = \lambda_{\max}(\mathbb{E}B_i^{(s)}) = 1$ . Then Lemma 22 implies that

$$\mathbb{P} \left\{ \lambda_{\min}(B_i^{(s)}) \leq 1 - \delta/2 \text{ or } \lambda_{\max}(B_i^{(s)}) \geq 1 + \delta/2 \right\} \leq n \exp(-\delta^2 p' n / (8\mu_t k)) + n \exp(-\delta^2 p' n / (12\mu_t k)).$$

The claim follows from the choice of  $p'$ . ■

Next, we will bound the other part of the expression for  $(G_\ell^{(s)})^M$  in Proposition 14.

**Claim 2** *There is a constant  $C$  so that the following holds. Suppose that  $p' \geq \frac{C\mu_t k}{n\delta^2}$ . Then for each  $s$ ,*

$$\mathbb{P} \left\{ \left\| e_i^T M_t (I - R_{\ell-1} R_{\ell-1}^T) P_i^{(s)} R_{\ell-1} \right\|_2 \geq \frac{\delta}{4} \left( \sum_{j=1}^t \left\| e_i^T M^{(j)} \right\|_2 \left\| E_{\ell-1}^j \right\| \right) \right\} \leq \frac{1}{20}.$$

**Proof** We compute the expectation of  $\left\| e_i^T M_t (I - R_{\ell-1} R_{\ell-1}^T) P_i^{(s)} R_{\ell-1} \right\|_2$  and use Markov's inequality. For the proof of this claim, let  $Y = M_i (I - R_{\ell-1} R_{\ell-1}^T)$ .

$$\begin{aligned} \mathbb{E} \left\| e_i^T Y P_i^{(s)} R_{\ell-1} \right\|_2^2 &= \mathbb{E} e_i^T Y P_i^{(s)} R_{\ell-1} R_{\ell-1}^T P_i^{(s)} Y^T e_i \\ &= e_i^T Y \mathbb{E} \left( P_i^{(s)} R_{\ell-1} R_{\ell-1}^T P_i^{(s)} \right) Y^T e_i \\ &= e_i^T Y \left( R_{\ell-1} R_{\ell-1}^T + \left( \frac{1}{p'} - 1 \right) \text{diag}_r \left( \|e_r^T R_{\ell-1}\|_2^2 \right) \right) Y^T e_i \\ &= \|e_i^T Y R_{\ell-1}\|_2^2 + \left( \frac{1}{p'} - 1 \right) \sum_{r=1}^n \|e_r^T R_{\ell-1}\|_2^2 (Y_{i,r})^2 \\ &= \left( \frac{1}{p'} - 1 \right) \sum_{r=1}^n \|e_r^T R_{\ell-1}\|_2^2 (Y_{i,r})^2 \\ &\leq \|e_i^T Y\|_2^2 \left( \frac{1}{p'} - 1 \right) \left( \frac{\mu_t k}{n} \right) \\ &\leq \frac{\delta^2 \|e_i^T Y\|_2^2}{400}, \end{aligned}$$

using the fact that  $Y R_{\ell-1} = 0$ , and finally our choice of  $p'$  (with an appropriately large constant  $C$ ). Now, using (32),

$$\|e_i^T Y\|_2 = \left\| e_i^T U^{(\leq t)} \Lambda_{(t)} U^{(\leq t)} (I - R_{\ell-1} R_{\ell-1}^T) \right\|_2 \leq \sum_{j=1}^t \left\| e_i^T M^{(j)} \right\|_2 \left\| D_{\ell-1}^j \right\| \leq \sum_{j=1}^t \left\| e_i^T M^{(j)} \right\|_2 \left\| E_{\ell-1}^j \right\|.$$

Along with Markov's inequality, this completes the proof.  $\blacksquare$

Finally, we control the term  $(G_\ell^{(s)})^N$ .

**Claim 3** *There is a constant  $C$  so that the following holds. Suppose that  $p' \geq Ck \log(n) \mu_t / (\delta^2 n)$  for a constant  $C$ . Then for each  $s \leq T$ ,*

$$\mathbb{P} \left\{ \left\| e_i^T (G_\ell^{(s)})^N \right\|_2 \geq \frac{\delta}{4} \|e_i^T N_t\|_2 \right\} \leq \frac{1}{15}.$$

**Proof** Using Proposition 14,

$$\begin{aligned} e_i^T (G_\ell^{(s)})^N &= e_i^T \left( N_t P_i^{(s)} R_{\ell-1} (B_i^{(s)})^{-1} - N_t R_{\ell-1} \right) \\ &= e_i^T \left( N_t P_i^{(s)} R_{\ell-1} - N_t R_{\ell-1} B_i^{(s)} \right) (B_i^{(s)})^{-1} \\ &= \left( e_i^T N_t (P_i^{(s)} - I) R_{\ell-1} + e_i^T N_t R_{\ell-1} (I - B_i^{(s)}) \right) (B_i^{(s)})^{-1} \\ &=: (y_1 + y_2) (B_i^{(s)})^{-1}. \end{aligned}$$

We have already bounded  $\left\| (B_i^{(s)})^{-1} \right\|$  with high probability in Claim 1, when the bound on  $p'$  holds, and so we now bound  $\|y_1\|_2$  and  $\|y_2\|_2$  with decent probability. As we did in Claim 2, we compute the expectation of  $\|y_1\|_2^2$  and use Markov's inequality.

$$\begin{aligned} \mathbb{E} \|y_1\|_2^2 &= \mathbb{E} \left\| e_i^T N_t (P_i^{(s)} - I) R_{\ell-1} \right\|_2^2 \\ &= e_i^T N_t \mathbb{E} \left[ (P_i^{(s)} - I) R_{\ell-1} R_{\ell-1}^T (P_i^{(s)} - I) \right] N_t^T e_i \\ &= e_i^T N_t \left( \frac{1}{p'} - 1 \right) \text{diag}_r (\|e_r^T R_{\ell-1}\|_2^2) N_t^T e_i \\ &= \left( \frac{1}{p'} - 1 \right) \sum_{r=1}^n (N_t)_{ir}^2 \|e_r^T R_{\ell-1}\|_2^2 \\ &\leq \left( \frac{\mu_t k}{np'} \right) \|e_i^T N_t\|_2^2. \end{aligned}$$

Thus, by Markov's inequality, we have

$$\mathbb{P} \left\{ \|y_1\|_2 \geq 20 \sqrt{\frac{\mu_t k}{np'}} \|e_i^T N_t\|_2 \right\} \leq \frac{1}{20}.$$



Next, we turn our attention to the second term  $\|y_2\|_2$ . We have

$$\|y_2\|_2 = \left\| e_i^T (N_t R_{\ell-1}) (I - B_i^{(s)}) \right\|_2 \leq \|e_i^T N_t R_{\ell-1}\|_2 \left\| I - B_i^{(s)} \right\|.$$

By Claim 1, we established that with probability  $1 - 1/n^5$ ,  $\left\| I - (B_i^{(s)}) \right\| \leq \frac{\delta}{2}$ , with our choice of  $p'$ . Thus, with probability at least  $1 - 1/n^5$ ,

$$\|y_2\|_2 \leq \delta \|e_i^T N_t R_{\ell-1}\|_2 \leq \frac{\delta}{2} \|e_i^T N_t\|_2.$$

Altogether, we conclude that with probability at least  $1 - 1/20 - 2/n^5$ , we have

$$\left\| e_i^T (G_\ell^{(s)})^N \right\|_2 \leq (\|y_1\|_2 + \|y_2\|_2) \left\| (B_i^{(s)})^{-1} \right\| \leq \frac{3\delta}{4(1 - \delta/2)} \|e_i^T N_t\|_2 \leq \delta \|e_i^T N_t\|_2$$

as long as  $\delta \leq 1/2$ . This proves the claim.  $\blacksquare$

Putting Claims 1, 2 and 3 together, along with the choice of  $p'_t = L_t s_{\max} p'$ , we conclude that, for each  $s \in [T]$  and for any  $\delta < 1/2$ ,

$$\mathbb{P} \left\{ \left\| e_i^T G_\ell^{(s)} \right\|_2 \geq \frac{\delta}{4(1 - \delta/2)} \left( \|e_i^T N_t\|_2 + \sum_{j=1}^t \left\| e_i^T M^{(j)} \right\|_2 \left\| E_{\ell-1}^j \right\| \right) \right\} \leq \frac{1}{5}. \quad (33)$$

This implies that

$$\|e_i^T G_\ell\|_2 = \left\| e_i^T \text{median}_s G_\ell^{(s)} \right\|_2 = \left\| \text{median}_s (e_i^T G_\ell^{(s)}) \right\|_2$$

is small with exponentially large probability. Indeed, by Lemma 23,

$$\mathbb{P} \left\{ \left\| e_i^T G_\ell \right\|_2 \geq \frac{\delta}{2(1 - \delta/2)} \left( \|e_i^T N_t\|_2 + \sum_{j=1}^t \left\| e_i^T M^{(j)} \right\|_2 \left\| E_{\ell-1}^j \right\| \right) \right\} \leq \exp(-cs_{\max}),$$

for some constant  $c$ . By the choice of  $s_{\max}$ , the failure probability is at most  $1/n^6$ , and a union bound over all  $i$  shows that, with probability at least  $1 - 1/n^5$ ,

$$\|e_i^T G_\ell\|_2 \leq \delta \left( \|e_i^T N_t\|_2 + \sum_{j=1}^t \left\| e_i^T M^{(j)} \right\|_2 \left\| E_{\ell-1}^j \right\| \right) = \omega_{\ell-1}^{(i)}. \quad (34)$$

This was the second claim in Lemma 11. Now, we show that in the favorable case that (34) holds, so does the first claim of Lemma 11, and this will complete the proof of the lemma. Suppose that

(34) holds. Then

$$\begin{aligned}
 \|G_\ell\|_F &= \sqrt{\sum_{i=1}^n \|e_i^T G_\ell\|_2^2} \\
 &\leq \sqrt{\sum_{i=1}^n \delta^2 \left( \|e_i^T N_t\|_2 + \sum_{j=1}^t \|e_i^T M^{(j)}\|_2 \|E_{\ell-1}^j\| \right)^2} \\
 &\leq \delta \sqrt{\sum_{i=1}^n \|e_i^T N_t\|_2^2} + \delta \sqrt{\sum_{i=1}^n \left( \sum_{j=1}^t \|e_i^T M^{(j)}\|_2 \|E_{\ell-1}^j\| \right)^2} \\
 &\leq \delta \|N_t\|_F + \delta \sqrt{\sum_{i=1}^n \left( \sum_{j=1}^t \|e_i^T M^{(j)}\|_2 \|E_{\ell-1}^j\| \right)^2}.
 \end{aligned}$$

Notice that, for any real numbers  $(a_{i,j})$ ,  $i \in [n]$ ,  $j \in [t]$ , and for any real number  $b_j$ ,  $j \in [t]$ , we have

$$\begin{aligned}
 \left( \sum_{i=1}^n \left( \sum_{j=1}^t a_{i,j} b_j \right)^2 \right)^{1/2} &= \|Ab\|_2 = \max_{\|z\|_2=1} z^T Ab = \max_{\|z\|_2=1} \sum_{j=1}^t (z^T A e_j) b_j \\
 &\leq \sum_{j=1}^t \max_{z^{(j)}} ((z^{(j)})^T A e_j) b_j = \sum_{j=1}^t \|A e_j\|_2 b_j = \sum_{j=1}^t \left( \sum_{i=1}^n a_{i,j}^2 \right)^{1/2} b_j.
 \end{aligned}$$

Thus, we may bound the second term above by

$$\begin{aligned}
 \delta \sqrt{\sum_{i=1}^n \left( \sum_{j=1}^t \|e_i^T M^{(j)}\|_2 \|E_{\ell-1}^j\| \right)^2} &\leq \delta \sum_{j=1}^t \left( \sum_{i=1}^n \|e_i^T M^{(j)}\|_2^2 \right)^{1/2} \|E_{\ell-1}^j\| \\
 &= \delta \sum_{j=1}^t \|M^{(j)}\|_F \|E_{\ell-1}^j\|.
 \end{aligned}$$

Altogether, we conclude that, in the favorable case the (34) holds,

$$\|G_\ell\|_F \leq \delta \left( \|N_t\|_F + \sum_{j=1}^t \|M^{(j)}\|_F \|E_{\ell-1}^j\| \right) = \omega_{\ell-1},$$

as desired. This completes the proof of Lemma 11.

## Appendix D. Dividing up $\Omega$

In the Matrix Completion literature, the most common assumption on the distribution of the set  $\Omega$  of observed entries is that each index  $(i, j)$  is included independently with some probability  $p$ . Call

this distribution  $\mathcal{D}(p)$ . In order for our results to be comparable with existing results, this is the model we adopt as well. However, for our analysis, it is much more convenient to imagine that  $\Omega$  is the union of several subsets  $\Omega_t$ , so that the  $\Omega_t$  themselves follow the distribution  $\mathcal{D}(p_t)$  (for some probability  $p_t$ , where  $\sum_t p_t = p$ ), and so that all of the  $\Omega_t$  are independent. Algorithmically, the easiest thing to do to obtain subsets  $\Omega_t$  from  $\Omega$  is to partition  $\Omega$  into random subsets of equal size. However, if we do this, the subsets  $\Omega_t$  will not follow the right distribution; in particular they will not be independent. For theoretical completeness, we show in this section how to split up the set  $\Omega$  in the correct way. More precisely, given  $p_t$  and  $p$  so that  $\sum_t p_t = p$ , we show how to break  $\Omega \sim \mathcal{D}(p)$  into (possibly overlapping) subsets  $\Omega_t$ , so that the  $\Omega_t$  are independent and each  $\Omega_t \sim \mathcal{D}(p_t)$ .

Algorithm 5 contains the details. Observe that the first thing that Algorithm 5 does is *throw away* samples from  $\Omega$ . Thus, while this step is convenient for the analysis, and we include it for theoretical completeness, in practice it may be unnecessary—especially if the assumption on the distribution of  $\Omega$  is an approximation to begin with.

**Input:** Parameters  $p_1, \dots, p_L$ , and a set  $\Omega \subset [n] \times [n]$  so that each index  $(i, j)$  is included in  $\Omega$  independently with probability  $p = \sum_\ell p_\ell$ .

**Output:** Subset  $\Omega_1, \dots, \Omega_L \subset \Omega$  so that each index  $(i, j)$  is included in  $\Omega_\ell$  independently with probability  $p_\ell$ , and so that all of the  $\ell$  are independent.

36 Choose

$$p' = 1 - \prod_{\ell=1}^L (1 - p_\ell).$$

Observe that  $p' \leq p$ .

37 Let  $\Omega'$  be a set that includes each element of  $\Omega$  independently with probability  $p'/p$ .

38 **return** SUBSAMPLE( $p_1, \dots, p_L, [n] \times [n], \Omega'$ )

**Algorithm 5: SPLITUP:** Split a set of indices  $\Omega$  (as in the input to Algorithm 1) into subsets  $\Omega_1, \dots, \Omega_t$  whose distributions are convenient for our analysis.

The correctness of Algorithm 5 follows from the following lemma, about the properties of Algorithm 6.

**Lemma 15** *Pick  $p_1, \dots, p_\ell \in [0, 1]$ , and suppose that  $\Omega \subset \mathcal{U}$  includes each  $u \in \mathcal{U}$  independently with probability  $p1 - \prod_{\ell=1}^L (1 - p_\ell)$ . Then the sets  $\Omega_1, \dots, \Omega_L$  returned by Algorithm 6 are distributed as follows. Each  $\Omega_\ell$  is independent, and includes each  $u \in \mathcal{U}$  independently with probability  $p_\ell$ .*

**Proof** Let  $\mathcal{D}$  denote the distribution we would like to show that that  $\Omega_\ell$  follow; so we want to show that the sets returned by Algorithm 6 are distributed according to  $\mathcal{D}$ . Let  $\mathbb{P}_{\mathcal{A}}\{\cdot\}$  denote the probability of an event occurring in Algorithm 6, and let and  $\mathbb{P}_{\mathcal{D}}\{\cdot\}$  denote the probability of an event occurring under the target distribution  $\mathcal{D}$ . Let  $N_u$  be the random variable that counts the number of times  $u$  occurs between  $\Omega_1, \dots, \Omega_\ell$ . Then observe that by definition,

$$q_r = \mathbb{P}_{\mathcal{D}}\{N_u = r | N_u \geq 1\},$$

and

$$p = \mathbb{P}_{\mathcal{D}}\{N_u \geq 1\}.$$

**Input:** Parameters  $p_1, \dots, p_L$ , a universe  $\mathcal{U}$ , and a set  $\Omega \subset \mathcal{U}$ , so that each element  $u \in \mathcal{U}$  is included independently with probability  $p = 1 - \prod_{\ell=1}^L (1 - p_\ell)$ .

**Output:** Set  $\Omega_1, \dots, \Omega_L \subset \mathcal{U}$ , so that each entry is included in  $\Omega_\ell$  independently with probability  $p_\ell$ , and so that  $\Omega_1, \dots, \Omega_L$  are independent.

39 For  $r \in \{1, \dots, L\}$ , let

$$q_r = \frac{1}{p} \sum_{S \subset \mathcal{U}, |S|=r} \left( \prod_{\ell \in S} p_\ell \right) \left( \prod_{\ell \notin S} (1 - p_\ell) \right).$$

Then  $\sum_{r=1}^L q_r = 1$ .

40 Initialize  $L$  empty sets  $\Omega_1, \dots, \Omega_L$ .

41 **for**  $u \in \Omega$  **do**

42     Draw  $r \in \{1, \dots, L\}$  with probability  $q_r$ .

43     Draw a random set  $T \subset [L]$  of size  $r$ .

44     Add  $u$  to  $\Omega_\ell$  for each  $\ell \in T$ .

45 **end**

46 **return**  $\Omega_1, \dots, \Omega_L$

**Algorithm 6:** SUBSAMPLE: Divide a random set  $\Omega$  into  $L$  subsets  $\Omega_1, \dots, \Omega_L$

We aim to show  $\mathbb{P}_{\mathcal{A}}\{\cdot\} = \mathbb{P}_{\mathcal{D}}\{\cdot\}$ . First, fix  $u \in \mathcal{U}$ , and fix any set  $S \subset [L]$ , and consider the event

$$E(u, S) = (\forall \ell \in S, u \in \Omega_\ell) \wedge (\forall \ell \notin S, u \notin \Omega_\ell).$$

We compute  $\mathbb{P}_{\mathcal{A}}\{E(u, S)\}$ .

$$\begin{aligned} \mathbb{P}_{\mathcal{A}}\{E(u, S)\} &= \mathbb{P}_{\mathcal{A}}\{u \in \Omega\} \sum_{r=1}^L q_r \mathbb{P}_{\mathcal{A}}\{\text{The random set } T \text{ of size } r \text{ is precisely } S\} \\ &= \mathbb{P}_{\mathcal{D}}\{N_u \geq 1\} \sum_{r=1}^L \mathbb{P}_{\mathcal{D}}\{N_u = r | N_u \geq 1\} \mathbb{P}\{\text{A random subset of } [L] \text{ size } r \text{ is precisely } S\} \\ &= \sum_{r=1}^L \mathbb{P}_{\mathcal{D}}\{N_u = r\} \mathbb{P}\{\text{A random subset of } [L] \text{ of size } r \text{ is precisely } S\} \\ &= \sum_{r=1}^L \mathbb{P}_{\mathcal{D}}\{N_u = r\} \mathbb{P}_{\mathcal{D}}\{E(u, S) | N_u = r\} \\ &= \mathbb{P}_{\mathcal{D}}\{E(u, S)\}. \end{aligned}$$

Next, we observe that for any fixed  $S$ , the events  $\{E(u, S)\}_{u \in \mathcal{U}}$  are independent under the distribution induced by Algorithm 6. This follows from the fact that in all of the random steps (including the generation of  $\Omega$  and within Algorithm 6), the  $u \in \mathcal{U}$  are treated independently. Notice that these events are also independent under  $\mathcal{D}$  by definition.

Now, for any instantiation  $\Omega' = (\Omega'_1, \dots, \Omega'_L)$  of the random variables  $(\Omega_1, \dots, \Omega_L)$ , consider the event

$$E(\Omega') = \forall \ell, \Omega_\ell = \Omega'_\ell.$$

We have

$$\begin{aligned}
 \mathbb{P}_{\mathcal{A}} \{E(\Omega')\} &= \mathbb{P}_{\mathcal{A}} \{\forall u, E(u, \{\ell : u \in \Omega'_\ell\})\} \\
 &= \prod_{u \in \mathcal{U}} \mathbb{P}_{\mathcal{A}} \{E(u, \{\ell : u \in \Omega'_\ell\})\} \quad \text{by independence in Alg. 6} \\
 &= \prod_{u \in \mathcal{U}} \mathbb{P}_{\mathcal{D}} \{E(u, \{\ell : u \in \Omega'_\ell\})\} \quad \text{by the above derivation} \\
 &= \mathbb{P}_{\mathcal{D}} \{E(\Omega')\} \quad \text{by independence under } \mathcal{D}
 \end{aligned}$$

Thus the probability of any outcome  $\Omega'$  is the same under  $\mathcal{D}$  and under Algorithm 6, and this completes the proof of the lemma.  $\blacksquare$

## Appendix E. Useful statements

In this appendix, we collect a few useful statements upon which we rely.

### E.1. Coherence bounds

First, we record some consequences of the bound (4) on the coherence of  $A$ . We always have

$$\|A\|_\infty \leq \|M\|_\infty + \|N\|_\infty \leq \max_{i,j} |e_i^T U \Lambda_U U^T e_j| + \|N\|_\infty \leq \sigma_1 \frac{\mu^* k}{n} + \frac{\mu^* \|N\|_F}{n} \leq \frac{\mu^*}{n} (k\sigma_1 + \Delta), \quad (35)$$

and similarly

$$\max_i \|e_i^T A\|_2 \leq \sqrt{\frac{\mu^*}{n}} (\sqrt{k}\sigma_1 + \Delta). \quad (36)$$

It will also be useful to notice that since  $\|e_i^T U^{(>t)}\|_2 \leq \|e_i^T U\|_2$ , (4) implies that for all  $t$ ,

$$\|N_t\|_\infty \leq \|M^{(>t)}\|_\infty + \|N\|_\infty \leq \frac{\mu^*}{n} (k\sigma_{r_t+1} + \Delta). \quad (37)$$

### E.2. Perturbation statements

Next, we will use the following lemma about perturbations of singular values, due to Weyl.

**Lemma 16** *Let  $N, E \in \mathbb{R}^{n \times n}$ , and let  $\tilde{N} = N + E$ . Let  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$  denote the singular values of  $N$ , and similarly let  $\tilde{\sigma}_i$  denote the singular values of  $\tilde{N}$ . Then for all  $i$ ,  $|\sigma_i - \tilde{\sigma}_i| \leq \|E\|$ .*

In order to compare the singular vectors of a matrix  $A$  with those of a perturbed version  $\tilde{A}$ , we will find the following theorem helpful. We recall that for subspaces  $U, V$ ,  $\sin \theta(U, V)$  refers to the sine of the *principal angle* between  $U$  and  $V$ . (See [Stewart and Sun \(1990\)](#) for more on principal angles).

**Theorem 17 (Thm. 4.4 in [Stewart and Sun \(1990\)](#))** *Suppose that  $A$  has the singular value decomposition*

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix},$$

and let  $\tilde{A} = A + E$  be a perturbed matrix with SVD

$$A = \begin{bmatrix} \tilde{U}_1 & \tilde{U}_2 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma}_1 & \\ & \tilde{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{bmatrix}.$$

Let

$$R = A\tilde{V}_1 - \tilde{U}_1\tilde{\Sigma}_1 \quad \text{and} \quad S = A^T\tilde{U}_1 - \tilde{V}_1\tilde{\Sigma}_1.$$

Suppose there are numbers  $\alpha, \delta > 0$  so that  $\sigma_{\min}(\tilde{\Sigma}_1) \geq \alpha + \delta$  and  $\sigma_{\max}(\tilde{\Sigma}_2) \leq \alpha$ . Then,

$$\max\{\sin \Theta(U_1, V_1), \sin \Theta(U_2, V_2)\} \leq \frac{\max\{\|R\|, \|S\|\}}{\delta}.$$

We will also use the fact that if the angle between (the subspaces spanned by) two matrices is small, then there is some unitary transformation so that the two matrices are close.

**Lemma 18** *Let  $U, V \in \mathbb{R}^{n \times k}$  have orthonormal columns, and suppose that  $\sin \theta(U, V) \leq \varepsilon$  for some  $\varepsilon < 1/2$ . Then there is some unitary matrix  $Q \in \mathbb{R}^{k \times k}$  so that  $\|UQ - V\| \leq 2\varepsilon$ .*

**Proof** We have  $V = \Pi_U V + \Pi_{U^\perp} V = U(U^T V) + \Pi_{U^\perp} V$ . Since  $\sin \theta(U, V) \leq \varepsilon$ , we have  $\|\Pi_{U^\perp} V\| \leq \varepsilon$ , and  $\sigma_k(U^T V) = \cos \theta(U, V) \geq \sqrt{1 - \varepsilon^2}$ . Thus, we can write  $U^T V = Q + E$ , where  $\|E\| \leq 1 - \sqrt{1 - \varepsilon^2}$ . The claim follows from the triangle inequality.  $\blacksquare$

### E.3. Subspace Iteration

Our algorithm uses the following standard version of the well-known Subspace Iteration algorithm—also known as Power Method.

**Algorithm 7: SUBSIT** ( $A, k, L$ ) (Subspace Iteration)

**Input:** Matrix  $A$ , target rank  $k$ , number of iterations  $L$

47  $S_0 \in \mathbb{R}^{n \times k} \leftarrow$  random matrix with orthogonal rows **for**  $\ell = 1, \dots, L$  **do**

48 |  $R_\ell \leftarrow AS_{\ell-1}$   $S_\ell \leftarrow \text{QR}(R_\ell)$

49 **end**

50 **for**  $i = 1, \dots, k$  **do**

51 |  $\tilde{\sigma}_i^2 \leftarrow (R_L)_i^T A^T A (R_L)_i$  //  $(R_L)_i$  is the  $i$ -th column of  $R_L$

52 **end**

53 **return**  $(R_L, \tilde{\sigma})$  **Output:** A matrix  $R \in \mathbb{R}^{n \times k}$  approximating the top  $k$  singular vectors of  $A$ , and estimates  $\tilde{\sigma}_1, \dots, \tilde{\sigma}_k$  of the singular values.

We have the following theorem about the convergence of SUBSIT.

**Theorem 19** *Let  $A \in \mathbb{R}^{n \times n}$  be any matrix, with singular values  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$ . Let  $R_L \in \mathbb{R}^{n \times k}$  be the matrix with orthonormal columns returned after  $L$  iterations of SUBSIT (Algorithm 7) with target rank  $k$ . for some suitably small parameter  $\gamma < 1$ . Then the values  $\tilde{\sigma}_i = (R_i)^T A R_i$  satisfy*

$$|\tilde{\sigma}_i - \sigma_i| \leq \sigma_i \left(1 - (1 - \gamma)^k\right) + 2n\sigma_1 (1 - \gamma)^L.$$

In particular, if  $\gamma = o(1/k)$  and if  $L = C \log(n)/\gamma$  then with probability  $1 - 1/\text{poly}(n)$ ,

$$|\tilde{\sigma}_i - \sigma_i| \lesssim \frac{\sigma_1}{n} + \sigma_i k \gamma \lesssim \sigma_1 k \gamma.$$

**Proof** Let  $r_1 \leq r_2 \leq \dots \leq r_t$  be the indices  $r \leq k$  so that  $\sigma_{r+1}/\sigma_r \leq 1 - \gamma$ . Notice that we may assume without loss of generality that  $r_t = k$ . Indeed, the result of running SUBSIT with target rank  $k$  is the same as the result of running SUBSIT with a larger rank and restricting to the first  $k$  columns of  $R_\ell$ . Write  $A = \sum_j U^{(j)} \Sigma_j V^{(j)}$ , where  $\Sigma_j$  contains the singular values  $\sigma_{r_j+1}, \dots, \sigma_{r_{j+1}}$ . Then using (Stewart, 2001, Chapter 6, Thm 1.1) and deviation bounds for the principal angle between a random subspace and fixed subspace, we have

$$\Pr \left\{ \sin \theta \left( U^{(j)}, R_L^{(j)} \right) \leq C n^c (1 - \gamma)^L \right\} \geq 1 - 1/n^{c'}.$$

Here,  $c'$  can be made any constant by increasing  $c$  and  $C$  is an absolute constant. Fix  $i$  and let  $x_i = (R_L)_i$  denote the  $i$ -th column of  $R_L$ . Suppose that  $i \in \{r_j + 1, \dots, r_{j+1}\}$ . Then, the estimates  $\tilde{\sigma}_i$  satisfy

$$\tilde{\sigma}_i = x_i^T A^T A x_i = \left\| A^{(j)} x_i \right\|_2^2 + \sum_{s \neq j} \left\| A^{(s)} x_i \right\|_2^2.$$

The second term satisfies

$$\sum_{s \neq j} \left\| A^{(s)} x_i \right\|_2^2 \leq \sigma_1^2 \sin^2 \theta(U^{(s)}, R_L^{(s)}) \leq \sigma_1^2 n^2 (1 - \gamma)^{2L}.$$

The first term has

$$\left\| A^{(j)} x_i \right\|_2^2 \leq \left\| A^{(j)} \right\|^2 = \sigma_{r_{j+1}}^2$$

and

$$\left\| A^{(j)} x_i \right\|_2^2 \geq \cos^2 \theta \left( U^{(s)}, R_L^{(s)} \right) \cdot \sigma_{\min}(A^{(j)}) \geq (1 - n^2 (1 - \gamma)^{2L}) \cdot \sigma_{r_j}^2.$$

By definition, as there are no significant gaps between  $\sigma_{r_{j+1}}$  and  $\sigma_{r_j}$ , we have

$$\frac{\sigma_{r_{j+1}}}{\sigma_{r_j}} \geq (1 - \gamma)^k,$$

and so this completes the proof after collecting terms. ■

#### E.4. Matrix concentration inequalities

We will repeatedly use the Matrix Bernstein and Matrix Chernoff inequalities; we use the versions from Tropp (2012):

**Lemma 20** [Matrix Bernstein Tropp (2012)] Consider a finite sequence  $\{Z_k\}$  of independent, random,  $d \times d$  matrices. Assume that each matrix satisfies

$$\mathbb{E} X_k = 0, \quad \|X_k\| \leq R \text{ almost surely.}$$

Define

$$\sigma^2 := \max \left\{ \left\| \sum_k \mathbb{E} X_k X_k^T \right\|, \left\| \sum_k \mathbb{E} X_k^T X_k \right\| \right\}.$$

Then, for all  $t \geq 0$ ,

$$\mathbb{P} \left\{ \left\| \sum_k X_k \right\| \geq t \right\} \leq 2d \exp \left( \frac{-t^2/2}{\sigma^2 + R/3} \right).$$

One corollary of Lemma 20 is the following lemma about the concentration of the matrix  $P_\Omega(A)$ .

**Lemma 21** *Suppose that  $A \in \mathbb{R}^{n \times n}$  and let  $\Omega \subset [n] \times [n]$  be a random subset where each entry is included independently with probability  $p$ . Then*

$$\mathbb{P} \{ \|P_\Omega(A) - A\| > u \} \leq 2n \exp \left( \frac{-u^2/2}{\left(\frac{1}{p} - 1\right) \left(\max_i \|e_i^T A\|_2^2 + \frac{u}{3} \|A\|_\infty\right)} \right).$$

**Proof** Let  $\xi_{ij}$  be independent Bernoulli- $p$  random variables, which are 1 if  $(i, j) \in \Omega$  and 0 otherwise.

$$P_\Omega(A) - A = \sum_{i,j} \left( \frac{\xi_{ij}}{p} - 1 \right) A_{i,j} e_i e_j^T,$$

which is a sum of independent random matrices. Using the Matrix Bernstein inequality, Lemma 20, we conclude that

$$\mathbb{P} \{ \|P_\Omega(A) - A\| > u \} \leq 2n \exp \left( \frac{-u^2/2}{\sigma^2 + Ru/3} \right),$$

where

$$\sigma^2 = \left\| \mathbb{E} \sum_{i,j} \left( \frac{\xi_{ij}}{p} - 1 \right)^2 A_{i,j}^2 e_i e_j^T e_j e_i^T \right\| = \left( \frac{1}{p} - 1 \right) \max_i \|A_i\|_2^2$$

and

$$\left\| \left( \frac{\xi_{ij}}{p} - 1 \right) A_{i,j} e_i e_j^T \right\| \leq R = \left( \frac{1}{p} - 1 \right) \|A\|_\infty$$

almost surely. This concludes the proof. ■

Finally, we will use the Matrix Chernoff inequality.

**Lemma 22** [Matrix Chernoff [Tropp \(2012\)](#)] *Consider a finite sequence  $\{X_k\}$  of independent, self-adjoint,  $d \times d$  matrices. Assume that each  $X_k$  satisfies*

$$X_k \succcurlyeq 0, \quad \lambda_{\max}(X_k) \leq R \text{ almost surely.}$$

Define

$$\mu_{\min} := \lambda_{\min} \left( \left\| \sum_k \mathbb{E} X_k \right\| \right), \quad \mu_{\max} := \lambda_{\max} \left( \left\| \sum_k \mathbb{E} X_k \right\| \right).$$

Then for  $\delta \in (0, 1)$ ,

$$\mathbb{P} \left\{ \lambda_{\min} \left( \sum_k X_k \right) \leq (1 - \delta) \mu_{\min} \right\} \leq d \exp(-\delta^2 \mu_{\min} / 2R)$$



and

$$\mathbb{P} \left\{ \lambda_{\max} \left( \sum_k X_k \right) \geq (1 + \delta) \mu_{\max} \right\} \leq d \exp(-\delta^2 \mu_{\max} / 3R).$$

### E.5. Medians of vectors

For  $v \in \mathbb{R}^k$ , let  $\text{median}(v)$  be the entry-wise median.

**Lemma 23** *Suppose that  $v^{(s)}$ , for  $s = 1, \dots, T$  are i.i.d. random vectors, so that for all  $s$ ,*

$$\mathbb{P} \left\{ \|v^{(s)}\|_2^2 > \alpha \right\} \leq 1/5.$$

Then

$$\mathbb{P} \left\{ \left\| \text{median}(v^{(s)}) \right\|_2^2 > 4\alpha \right\} \leq \exp(-\Omega(T)).$$

**Proof** Let  $S \subset [T]$  be the set of  $s$  so that  $\|v^{(s)}\|_2^2 \leq \alpha$ . By a Chernoff bound,

$$\mathbb{P} \left\{ |S| \leq \frac{3T}{4} \right\} = \mathbb{P} \left\{ \sum_{s=1}^T \mathbf{1}_{\|v^{(s)}\|_2^2 > \alpha} > \frac{T}{4} \right\} \leq \exp(-\Omega(T)).$$

Suppose that the likely event occurs, so  $|S| > 3T/4$ . For  $j \in [k]$ , let

$$S_j = \left\{ s \in S : (v_j^{(s)})^2 \geq \text{median}_s((v_j^{(s)})^2) \right\}.$$

Because  $|S| > 3T/4$ , we have  $|S_j| \geq T/4$ . Then

$$\begin{aligned} \left\| \text{median}_s(v_j^{(s)}) \right\|_2^2 &= \sum_{j=1}^n \text{median}_s \left( (v_j^{(s)})^2 \right) \leq \sum_{j=1}^n \frac{1}{|S_j|} \sum_{s \in S_j} (v_j^{(s)})^2 \\ &\leq \sum_{j=1}^n \frac{4}{T} \sum_{s \in S_j} (v_j^{(s)})^2 \leq \sum_{j=1}^n \frac{4}{T} \sum_{s \in S} (v_j^{(s)})^2 \leq \frac{4}{T} \sum_{s \in S} \|v^{(s)}\|_2^2 \leq \frac{4|S|\alpha}{T} \leq 4\alpha. \end{aligned}$$

This completes the proof. ■