On the Provable Convergence of Alternating Minimization for Matrix Completion

Moritz Hardt
IBM Research Almaden
San Jose, CA 95120, USA
Email: mhardt@us.ibm.com

Abstract—Alternating Minimization is a widely used and empirically successful framework for Matrix Completion and related low-rank optimization problems. We present a new algorithm based on Alternating Minimization that provably recovers an unknown low-rank matrix from a random subsample of its entries under a standard incoherence assumption while achieving a linear convergence rate. Compared to previous work, our results reduce the provable sample complexity requirements of the Alternating Minimization approach by at least a quartic factor in the rank and the condition number of the unknown matrix. These improvements apply when the matrix is exactly low-rank and when it is only close to low-rank in the Frobenius norm.

Underlying our work is a new robust convergence analysis of the well-known Subspace Iteration algorithm for computing the dominant singular vectors of a matrix also known as the Power Method. This viewpoint leads to a conceptually simple understanding of Alternating Minimization that we exploit. Additionally, we contribute a new technique for controlling the coherence of intermediate solutions arising in iterative algorithms. These techniques may be of interest beyond their application here.

I. INTRODUCTION

Alternating Minimization is an empirically successful heuristic for the Matrix Completion problem in which the goal is to recover an unknown low-rank matrix from a subsample of its entries. Matrix Completion has received a tremendous amount of attention over the past few years not least due to its applicability in collaborative filtering and recommender systems. Alternating Minimization was first used in this context by [1], [7] and continues to play an important role in practical approaches to the Matrix Completion problem. As a case in point, the approach formed an important component in the winning submission for the Netflix Prize [17].

Given a subsample \(\Omega\) of entries drawn from an unknown matrix \(A\), Alternating Minimization starts from a poor approximation \(X_0, Y_0\) to the target matrix and gradually improves 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

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

This optimization step is then repeated with \(X_\ell\) fixed in order to determine \(Y_\ell\) as

\[
Y_\ell = \arg\min_Y \sum_{(i,j) \in \Omega} \left| A_{ij} - (X Y^\ell)_{ij} \right|^2.
\]

This basic update step is usually combined with certain initialization procedures for finding \(X_0, Y_0\), as well as methods for modifying intermediate iterates, e.g., truncating large entries. More than a specific algorithm, we therefore think of Alternating Minimization as a framework for solving a non-convex low-rank optimization problem.

A major advantage of Alternating Minimization over alternatives is that each update is computationally cheap and has a small memory footprint as we only need to keep track of \(2k\) vectors. In contrast, the popular “nuclear norm” approach to Matrix Completion [2], [18], [3] requires solving a semidefinite program. The advantage of the nuclear norm approach is that it comes with strong provable guarantees under certain assumptions on the unknown matrix and the subsample of its entries. There are two common assumptions which together imply that Nuclear Norm minimization succeeds. The first is that the subsample \(\Omega\) includes each entry of \(A\) uniformly at random with probability \(p\). The second assumption is that the first \(k\) singular vectors of \(A\) span an incoherent subspace. Informally coherence measures the correlation of the subspace with any standard basis vector. More formally, the coherence of a \(k\)-dimensional subspace of \(\mathbb{R}^n\) is at most \(\mu\) if the projection of each standard basis vector has norm at most \(\sqrt{\mu k}/n\). The dominant singular space of various random matrix models typically satisfies this property with small \(\mu\). But even real-world matrices tend to exhibit this property.

The broad goal of this work is to obtain provable bounds on the sample complexity of Alternating Minimization under the same assumptions.

a) Rigorous Convergence Bounds for Alternating Minimization.: Our results build on two remarkable recent works by Jain, Netrapalli and Sanghavi [13] and Keshavan [14] who gave the first rigorous sample complexity bounds for Alternating Minimization that we are aware of. Informally, the bounds of Jain et al. guarantee convergence up to small error provided that each entry is revealed independently with probability at least

\[
p \geq \frac{k^7 (\sigma_1/\sigma_k)^6 \mu(U)^2}{n}.
\]
Here, \( k \) is the rank of the unknown matrix and \( \mu(U) \) is the coherence of the range of \( A \). We ignore logarithmic factors in \( n \). Keshavan obtained the incomparable bound \( p \gtrsim k(\sigma_1/\sigma_k)^3 \mu(U)/n \) that is superior when the matrix has small condition number \( \sigma_1/\sigma_k \).

An important observation of [13] is that the update rule in Alternating Minimization can be analyzed as a noisy update step of the well known Subspace Iteration Algorithm—often called Power Method, Orthogonal Iteration, or Simultaneous Iteration. The noise term that arises depends on the sampling error induced by the subsample of the entries. This situation suggests that in order to understand Alternating Minimization, it makes sense to first understand the robustness properties of Subspace Iteration. This is in itself an interesting question.

b) Robustness of Subspace Iteration. Consider a model of numerical linear algebra in which an input matrix \( A \) can only be accessed through noisy matrix vector products of the form \( Ax + g \), where \( x \) is a chosen vector and \( g \) is a possibly adversarial noise term. Our goal is to compute the dominant singular vectors \( u_1, \ldots, u_k \) of the matrix \( A \) via Subspace Iteration. Subspace Iteration starts with an initial guess, an orthonormal matrix \( X_0 \in \mathbb{R}^{n \times k} \) typically chosen at random. The algorithm then repeatedly refines its guess by virtue of the operation \( Y_q = AX_{q-1} + G_q \), followed by an orthonormalization step in order to obtain \( X_q \) from \( Y_q \). Here, \( G_q \) is the noise variable added to the computation.

We present convergence bounds characterizing how quickly subspace iteration converges from a random initial subspace to an approximation of the space \( U \) spanned by the vectors \( u_1, \ldots, u_k \). The quality of approximation is measured in terms of the tangent of the largest principal angle between \( X_q \) and \( U \). Principal angles are a useful tool in numerical analysis that makes it convenient to talk about the similarity between subspaces.

Our understanding of subspace iteration leads to a conceptually simple view of Alternating Minimization which is crucial in obtaining our improvements. Moreover, it gives the convergence behavior an appealing geometric interpretation in terms of principal angles.

A. Our results

Our first result pertains to the exact low-rank problem where the goal is to recover an unknown rank \( k \) matrix \( M \) from a subsample \( \Omega \) of its entries where each entry is included independently with probability \( p \). Here and in the following we will always assume that \( M = U\Lambda U^\dagger \) is a symmetric \( n \times n \) matrix with singular values \( \sigma_1 \geq \cdots \geq \sigma_k \). Our result generalizes straightforwardly to rectangular matrices as we will see. We express our results in terms of an upper bound \( \epsilon \) on the largest principal angle between the range of \( M \) and the output subspace of our algorithm represented by an orthonormal matrix \( X \in \mathbb{R}^{n \times k} \). We omit a formal definition at this point. Suffice it to say for the moment that a bound on the principal angle is a strong guarantee that immediately implies bounds on other measures such as reconstruction error in terms of spectral or Frobenius norm. Informally speaking, we prove that our algorithm achieves principal angle \( \epsilon \) assuming that

\[
p \gtrsim \frac{k^2 \mu(U)(k + \mu(U))(\sigma_1/\sigma_k)^2 \log(1/\epsilon)^2}{n}.
\]

As before, we ignore logarithmic factors in \( n \). Assuming the natural regime in which \( \mu(U) = \Omega(k) \) our bound simplifies to

\[
p \gtrsim \frac{k^2 \mu(U)(\sigma_1/\sigma_k)^2 \log(1/\epsilon)^2}{n}.
\]

This result improves upon [13] by a factor of \( k^5(\sigma_1/\sigma_k)^4 \) and improves on [14] as soon as \( \sigma_1/\sigma_k \gg k^{1/6} \mu(U)^{1/3} \).

We remark that the number of least squares update steps is bounded by \( O(\log(n/\epsilon)) \). The cost of performing these update steps is up to a logarithmic factor what dominates the worst-case running time of our algorithm. It can be seen that the least squares problem can be solved in time \( O(\Omega \cdot k^2) \) which is is linear in \( n \) and polynomial in \( k \). The number of update steps enters the sample complexity since we assume (as in previous work) that fresh samples are used in each step. However, the logarithmic dependence on \( 1/\epsilon \) guarantees a linear convergence rate and allows us to obtain any inverse polynomial error with only a constant factor overhead in sample complexity.

c) Noisy Matrix Completion. Noisy Matrix Completion refers to the Matrix Completion problem where the unknown matrix is only close to low-rank, typically in Frobenius norm. The approach of Jain et al. was adapted to the noisy setting by Gunasekar et al. [6] showing that the same sample complexity that we saw in Equation 1 also holds in the noisy setting under suitable assumptions.

Our results apply to any matrix of the form \( A = M + N \), where \( M = U\Lambda U^\dagger \) is a matrix of rank \( k \) as before and \( N = (I - UU^\dagger)A \) is the part of \( A \) not captured by the dominant singular vectors. Here, \( N \) can be an arbitrary deterministic matrix. The assumption that we will make is that \( N \) satisfies the following norm constraints:

\[
\max_{i \in [n]} \|e_i N\|^2 \leq \frac{\mu N}{n} \cdot \sigma_k^2
\]

\[
\max_{j \in [n]} \|N_{ij}\|^2 \leq \frac{\mu N}{n} \cdot \|e_i N\|^2.
\]

Here, \( e_i \) denotes the \( i \)-th standard basis vector so that \( \|e_i N\| \) 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 \( \sigma_k \).

From here on we let \( \mu^* = \max\{\mu(U), \mu_N\} \). Our results show that in order to achieve principal angle \( \epsilon \) between \( U \) and the subspace found by our algorithm, it suffices to have sample complexity

\[
p \gtrsim \frac{k^2 \mu^*(k + \mu^*)(\sigma_1/\sigma_k)^2 + (\|N\|/\epsilon \sigma_k)^2 \log(1/\epsilon)^2}{\gamma^5 n}.
\]

Here, \( \gamma = 1 - \sigma_{k+1}/\sigma_k \) 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 $N = 0$ in which case $\gamma = 1$. Compared to our noise-free bound, there are two new parameters that enter the sample complexity. The first one is $\gamma$. The second is the term $\|N\|_F / \epsilon$. To interpret this quantity, suppose that $A$ has a good low-rank approximation in Frobenius norm, formally, $\|N\|_F \leq \epsilon \|A\|_F$ for $\epsilon \leq 1/2$, then it must also be the case that $\|N\|_F / \epsilon \leq 2 \|M\|_F$. Our algorithm then finds a good rank $k$ approximation with at most $O(k^2 (\sigma_1 / \sigma_k)^2 (\mu^*)^2 n)$ samples assuming $\gamma = \Omega(1)$ and $\mu^* = \Omega(k)$. Hence, assuming that $A$ has a good rank $k$ approximation in Frobenius norm and that $\sigma_k$ and $\sigma_{k+1}$ are well-separated, our bound recovers the noise-free bound up to a constant factor. We can also show that assuming $6$, our algorithm outputs matrices $X, Y \in \mathbb{R}^{n \times k}$ such that we have the following reconstruction error bounds:

$$
\|A - XY^\dagger\|_2^2 \leq \epsilon \|M\|_2^2 + \|N\|_2^2 \\
\|A - XY^\dagger\|_F^2 \leq \epsilon \|M\|_F + \|N\|_F.
$$

Compared to previous work, we achieve the same improvements over [6] as we did compared to [13] in the noise-free case. Further our assumptions in 4 are weaker than the assumption of [6]. The latter work required the largest entry of $N$ in absolute value to be bounded by $O(\sigma_k / \sqrt{n})$. This directly implies that each row of $N$ has norm at most $O(\sigma_k / \sqrt{n} \sqrt{k})$ and that the $\|N\|_F \leq O(\sigma_k / \sqrt{k})$. These assumptions roughly correspond to setting $\mu_N = 1 / \sqrt{k}$. Moreover, under this assumption we would have $\gamma \geq 1 - o(1)$. Keshavan’s result [14] also applies to the noisy setting, but it requires $\|N\|_2 \leq (\sigma_k / \sigma_1)^3$ and $\max_i \epsilon_i^* N_i \leq \sqrt{\mu(U) k / n} \|N\|_2$. Taken together this implies $\|N\|_F \leq \sqrt{k} (\sigma_k / \sigma_1)^3$ which can be a substantially stronger requirement on $\|N\|_F$ than what is needed in our work.

II. PROOF OVERVIEW AND TECHNIQUES

We briefly outline various components of the analysis below.

d) Robust Convergence of Subspace Iteration.: We begin with a convergence analysis of the subspace iteration algorithm where an arbitrary perturbation $G\ell$ is added at each step. A crucial component of our analysis is the choice of a suitable potential function that decreases at each step. Here we make use of the tangent of the largest principal angle between the subspace $U$ spanned by the first $k$ singular vectors of the input matrix and the $k$-dimensional space spanned by the columns of the iterate $X\ell$. Principal angles are a very useful tool from numerical analysis in analyzing iterative methods. Our analysis shows that the algorithm essentially converges at the rate of $(\sigma_k + \Delta) / (\sigma_k - \Delta)$ for some $\Delta \ll \sigma_k$ under suitable conditions on the noise matrix $G\ell$. In measuring the "magnitude" of the noise term it is important to distinguish between the projection of $G\ell$ onto the space $U$ and the projection onto its orthogonal complement $V = U^\perp$. Intuitively speaking, the norm $\|U^\dagger G\ell\|$ controls how much the cosine of the largest principal angle is perturbed, while the norm $\|V^\dagger G\ell\|$ controls the perturbation of the sine of the principle angle. Keeping track of both quantities controls the tangent. In order to obtain strong bounds, it is later on essential to leverage the fact that $\|U^\dagger G\ell\|$ is the norm of a $k$-dimensional projection of $G\ell$ which we expect to be much smaller than $V^\dagger G\ell$ when $k$ is small.

e) Alternating Least Squares.: In the following we will always assume that $A$ is a symmetric $n \times n$ matrix. This is without loss of generality as can be shown using standard argument.

We analyze two methods based on Alternating Least Squares. In each case the update step can be expressed as an instance of NSI. We obtain convergence bounds directly after analyzing the error terms that arise in each case. We call our first method BasicLS. An update step takes the form

$$
Y\ell = \arg \min_Y \|P_\Omega(A) - X\ell_{-1} Y^\dagger\|^2_2. \quad (7)
$$

Here $P_\Omega$ is the linear operator on matrices that zeros out all entries except those in $\Omega$. It is not hard to see that the optimizer in Equation 7 satisfies $Y\ell = P_\Omega(A) X\ell_{-1}$. In other words the algorithm simply performs subspace iteration on the subsampled matrix $P_\Omega(A)$. The range of $Y\ell$ is unaffected by scaling so we can express $Y\ell = A X\ell_{-1} + G\ell$ where $G\ell = (p^{-1} P_\Omega(A) - A) X\ell_{-1}$. Appealing to the Matrix Bernstein concentration inequality, we obtain suitable bounds on the quantities $\|G\ell\|$ and $\|U^\dagger G\ell\|$. These bounds then directly lead to a convergence result for BasicLS using our analysis of NSI. The advantage of BasicLS is that it is feasible to argue convergence from a random initial factorization since we can control $\|U^\dagger G\ell\|$. Moreover, each update is very efficient. The main disadvantage of BasicLS is that it has a quadratic dependence on the error parameter $\epsilon$ even in the case where $A$ is exactly rank $k$.

This behavior is remedied by the well-known Alternating Least Squares update rule that we call AltLS:

$$
Y\ell = \arg \min_Y \|P_\Omega(A) - X\ell_{-1} Y^\dagger\|_2^2. \quad (8)
$$

We can again express the optimal $Y\ell$ as $Y\ell = A X\ell_{-1} + G\ell$ using gradient information about the least squares objective. The error term $G\ell$ has an intriguing property. Its spectral norm $\|G\ell\|$ depends on the sine of the principal angle between $U$ and $X\ell_{-1}$, i.e., $\sin \theta(U, X\ell_{-1})$. As the algorithm begins to converge the norm of the error term starts to diminish. Near exact recovery is now possible (assuming the matrix has rank at most $k$). The main disadvantage of AltLS is that it is more difficult to control the quantity $\|U^\dagger G\ell\|$ which is required in order to argue about convergence from a random subspace. Moreover, the algorithm is somewhat less efficient than BasicLS.

For these reasons we combine our two algorithms. We first run BasicLS starting from a random factorization and then switch to AltLS once there is non-trivial correlation between $U$ and $X\ell_{-1}$. For example, $\sin \theta(U, X\ell_{-1}) \leq 1/4$ is sufficient.

f) Smooth Orthonormalization.: A novelty in our approach is the way we argue about the coherence of each iterate $X\ell$. Ideally, we would like to argue that $\mu(X\ell) = O(\mu^*)$. One approach is syntactic in nature and tries to reason about how $X\ell$ was obtained from $Y\ell$ as a result of the Gram-Schmidt
algorithm. This approach quickly runs into some overhead that depends on the perturbation stability of the QR-factorization which in turn depends on the condition number of the matrix.\footnote{Nevertheless this direct approach could be plugged into our analysis and it still results in some improvements over [13].} We take a more geometric approach. We will certify that the coherence of the range of $Y_t$ is small by exhibiting a subspace $S$ of dimension slightly larger than $k$ which contains the range of $Y_t$ and satisfies $\mu(S) = O(\mu^*)$. It is not hard to show that this implies that $\mu(X_t) = O(\mu^*)$. The subspace $S$ that we construct is obtained from the space spanned by the top $k$ singular vectors $U$ of $A$ as well as the columns of $G_t$. Assuming for now that $A$ has rank $k$, it is easy to see that these vectors span the range of $Y_t$. What’s important is that we can control the row norms of the matrices $U$ and $G_t$. In the case of $U$ this is just given by the coherence of $A$ and in the case of $G_t$ it can be shown using simple concentration bounds. However, $G_t$ is not an orthonormal matrix and to reason about the coherence of $S$ we need to find an orthonormal basis for it. The natural way of doing this is to start with the top $k$ singular vectors $u_1, \ldots, u_k$ and use the Gram-Schmidt algorithm to orthonormalize $G_t$ against $u_1, \ldots, u_k$. An obstruction arises when two columns of $G_t$ might be nearly collinear so that they could span a vector of large $\ell_\infty$-norm even though neither of the columns has large $\ell_\infty$-norm (relative to its Euclidean norm). To avoid this pathological situation, the key idea is to inject an additional Gaussian noise term $H_t$ after the update rule that generated $Y_t$. This additional noise term guarantees that the matrix $G_t + H_t$ is well-conditioned enough so that we can find a suitable basis for its range whose coherence is at most $O(\mu^*)$. Moreover, if we can guarantee that the magnitude of $H_t$ is not larger than that of $G_t$, then our previous convergence analysis will continue to go through. We call the resulting procedure SmoothGS.

\textbf{g) Full version.:} A full version of this short abstract is available [8].

A. Further Discussion of Related Work

\textbf{h) Matrix Completion and Alternating Minimization.:} There is a vast literature on other approaches that we cannot survey here. Most closely related is the work of Jain et al. [13] that suggested the idea of thinking of Alternating Least Squares as a noisy update step in the Power Method. The approach was adapted to the noisy setting by [6] following the approach of Jain et al. Our approach takes inspiration from Jain et al. by analyzing Alternating Least Squares as a noisy Power method. However, our analysis is substantially different in both how convergence and low coherence is argued. The approach of Keshavan [14] uses a rather different argument. Our result in the noisy setting is incomparable to the work of Keshavan, Montanari and Oh [15], [16]. Their algorithm, called OptSPACE, requires a stronger incoherence assumption and has larger sample complexity in terms of the condition number. However, the requirement on $\max_{ij} |N_{ij}|$ is weaker in certain cases.

\textbf{i) Privacy-preserving spectral analysis.:} Our work is also closely related to a line of work on differentially private singular vector computation [10], [11], [9]. These papers each consider algorithms based on the power method where noise is injected to achieve the privacy guarantee known as Differential Privacy [5]. Hardt and Roth [10], [11], [9] observed that incoherence could be used to obtain improved guarantees. This, however, required controlling the coherence of the iterates produced by the noisy power method which leads to similar problems as the ones faced here. What’s simpler in the privacy setting is that the noise term is typically Gaussian leading to a cleaner analysis. Our work uses a similar convergence analysis for noisy subspace iteration that was used in a concurrent work by the author [11].

\textbf{j) Numerical Analysis.:} One might expect our analysis of the Power Method to have appeared in the numerical analysis literature. However, we are not aware of a reference and there are a number of points to consider. First, our noise model is adaptive thus setting it apart from the classical perturbation theory of the singular vector decomposition [4]. Second, we think of the perturbation at each step as large making it conceptually different from floating point errors. Third, research in numerical analysis over the past decades has largely focused on faster Krylov subspace methods. There is some theory of \textit{inexact Krylov methods} [19], [20] that captures the effect of noisy matrix-vector products in this context. Related to our work are also results on the perturbation stability of the QR-factorization since those could be used to obtain convergence bounds for subspace iteration. Such bounds, however, must depend on the condition number of the matrix that the QR-factorization is applied to. See Chapter 19.9 in [12] and the references therein for background. Our proof strategy avoids this particular dependence on the condition number.

ACKNOWLEDGMENT

Part of this research was done while the author visited the Simons Institute for Theoretical Computer Science at UC Berkeley.

REFERENCES