Provable Matrix Sensing using Alternating Minimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Alternating minimization has emerged as a popular heuristic for large-scale machine learning problems involving low-rank matrices. However, there have been few (if any) theoretical guarantees on its performance. In this work, we investigate the natural alternating minimization algorithm for the popular matrix sensing problem first formulated in [RFP07]; this problem asks for the recovery of an unknown low-rank matrix from a small number of linear measurements thereof. We show that under suitable RIP conditions, alternating minimization linearly converges to the true matrix. Our result can be extended to matrix completion from randomly sampled entries. Our analysis uses only elementary linear algebra and exploits the fact that, under RIP, alternating minimization can be viewed as a noisy version of orthogonal iteration (which is used to compute the top singular vectors of a matrix).

1 Introduction

Alternating minimization is a popular heuristic for solving non-convex optimization problems in practice [Bra03, Kor08, GB00]. Typically, it involves partitioning the variables into two sets, such that minimizing over either one set is easy when the other one is held fixed. The algorithm then alternates between updating each set in turn, holding the other fixed, until convergence. The most attractive feature of alternating minimization is it's simplicity. In particular, in many problems of interests each of the optimization steps turns out to be simple. For instance, in the case of matrix sensing and matrix completion, each of the optimization steps turns out to be a least squares problem, which can be solved efficiently.

The **low-rank matrix sensing (LRMS) problem**, first proposed by Recht, Fazel and Parrilo [RFP07], seeks to recover a rank-k matrix [1] given a set of linear measurements of the matrix. Formally, let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^d$ be a linear operator, and $b = \mathcal{A}(M) \in \mathbb{R}^d$, be d linear measurements of a rank-k matrix $M \in \mathbb{R}^{m\times n}$. The goal is to recover M given b and $\mathcal{A}$. Now, if $d \geq m \cdot n$, and if $\mathcal{A}$ is invertible then M can be uniquely estimated by solving system of linear equations $b = \mathcal{A}(X)$. However, if $d < m \cdot n$, then $\mathcal{A}$ cannot be invertible and hence infinitely many X satisfy $b = \mathcal{A}(X)$. To alleviate this issue, several measurement schemes ($\mathcal{A}$) have been proposed that ensure a *unique* solution to the following problem:

$$\text{Find } X \quad s.t \quad \mathcal{A}(X) = b,\ \ rank(X) \leq k, \quad X \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^d. \tag{LRMS}$$

While restricting rank of X leads to unique solution, the problem becomes NP-hard in general [MJCD08]. However, several recent results show that by suitably designing measurement matrices ($\mathcal{A}$), (LRMS) can be solved *exactly* in poly(m, n, k) time and hence by uniqueness of the solution to (LRMS), one can recover the underlying matrix M [RFP07, JMD10].

[1] Through out the paper we assume k to be very small compared to matrix size and use rank-k or low-rank matrix interchangeably.

RIP-based Matrix Sensing: One popular method to design measurement matrices ($\mathcal{A}$) is by using random matrices, i.e., each element of A_i (that provides i-th measurement) is sampled i.i.d. from a 0-mean sub-Gaussian distribution. In fact, a more general characterization of such measurement matrices is through Restricted Isometry Property (RIP) [CT05], that requires the transformation $\mathcal{A}$ to act as an (approximate) isometry for all low-rank matrices (see Definition 3.1). Using such design matrices, several methods have been shown to *exactly* recover the underlying rank-k matrix $M \in \mathbb{R}^{m\times n}$ using only $O(kn\log m)$ measurements[2], which is also information theoretically optimal.

There are two existing methods to solve (LRMS) that are known to have provable guarantees on recovery:

1. Convex relaxation via nuclear norm [RFP07]: In this approach, a regularizer term (which is chosen to be the nuclear norm of the matrix) is added to the objective function to promote low rank solutions. The resulting problem is solved via convex optimization techniques.

2. Singular Value Projection [JMD10]: This is a projected gradient descent method where after each gradient descent step, the solution is projected on to the space of low rank matrices.

One issue with both the above algorithms is typically, they require to find SVD at each iteration, hence scales poorly to large matrices. In comparison, alternating minimization solves only a least squares at each step and hence is much more scalable. approaches. Moreover, alternating minimization has been observed to have very good performance as compared to both of the above methods empirically [JMD10].

For the (LRMS) problem, alternating minimization takes advantage of the natural decomposition of a $m \times n$ low rank matrix $X = UV^\dagger$ where $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ and tries to solve the following optimization problem:

$$\begin{aligned}\min_{U,V}\ & \|\mathcal{A}(UV^\dagger) - b\|_2^2,\\ \text{s.}t.\ & U \in \mathbb{R}^{m\times k},\ \ V \in \mathbb{R}^{n\times k},\end{aligned} \tag{1}$$

where $V^\dagger$ denotes the Hermitian conjugate (or transpose) of V. In each step of alternating minimization, the algorithm optimizes over U holding V fixed and then optimizes over V holding U fixed. The algorithm iterates till convergence. Each optimization step in the above procedure turns out to be a least squares problem and can be accomplished efficiently. Though alternating minimization has been observed to perform well empirically, there have been no theoretical guarantees on its performance.

In this work, we prove the first theoretical guarantees on the performance of alternating minimization for the (LRMS) problem. In particular, we show that if the sensing matrices satisfy RIP with a small enough RIP constant (δ_{2k}), then alternating minimization achieves linear convergence to the true low rank matrix M. Finally, we remark that our results can also be extended to the problem of low rank matrix completion from randomly sampled entries under standard incoherence assumptions [CR09]. For matrix completion, we show that the time complexity of alternating minimization method is $\tilde{O}(kmn\log(1/\epsilon))$ while the best known trace-norm minimization based method has time complexity $\tilde{O}(mn^2\frac{1}{\sqrt{\epsilon}})$.

The rest of the paper is organized as follows. We formally present the alternating minimization algorithm for (LRMS) in Section 2, our main result regarding convergence of the alternating minimization in Section 3 and finally conclude with some discussion in Section 4.

2 Alternating Minimization Algorithm

In this section, we present the alternating minimization algorithm for matrix sensing. This algorithm is well known in the literature and we reproduce it here only for the sake of completeness.

[2] Throughout the paper we assume $m < n$.

Algorithm 1 Alternating minimization for matrix sensing

```
Input $b, \mathcal{A}$
$U^0 = SVD(\mathcal{A}^\dagger b, k)$ i.e., top-$k$ left singular vectors of $\mathcal{A}^\dagger b = \sum_i A_i b_i$
for $t = 0, \cdots, T-1$ do
    $V^{t+1} \leftarrow \text{argmin}_{V \in \mathbb{R}^{n \times k}} \; \|\mathcal{A}(U^t V^\dagger) - b\|_2^2$
    $U^{t+1} \leftarrow \text{argmin}_{U \in \mathbb{R}^{m \times k}} \; \|\mathcal{A}(U (V^{t+1})^\dagger) - b\|_2^2$
end for
Return $X = U^T (V^T)^\dagger$
```

Note that in the above we use the fact that without loss of generality, linear measurements $b = \mathcal{A}(M)$ can be represented as:

$$b = \begin{bmatrix} \langle A_1, M \rangle \\ \langle A_2, M \rangle \\ \vdots \\ \langle A_d, M \rangle \end{bmatrix},$$

where $A_i \in \mathbb{R}^{m \times n}, \forall i$ and $\langle A_i, M \rangle = tr(A_i^\dagger M)$ and $A_i^\dagger$ denotes the hermitian conjugate (or transpose) of A_i.

3 Theoretical Guarantees

In this section, we state our main result concerning the convergence of Algorithm 1 for the (LRMS) problem. Due to space constraints, we only give the main intuition and not the entire proof. We will provide a complete proof of the Theorem 3.2 in the full version of this paper.

We start with the definition of Restricted Isometry Property (RIP) of linear measurement operator $\mathcal{A}$.

Definition 3.1. *[RFP07] A linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \to \mathbb{R}^d$ acting on the space of real matrices, $\mathbb{R}^{m \times n}$ is said to satisfy k-RIP with constant δ_k if for every rank-k X, we have the following :*

$$(1 - \delta_k) \|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1 + \delta_k) \|X\|_F^2 . \tag{2}$$

Note that the above definition is a straightforward extension of the RIP assumption proposed by [CT05] for compressive vector sensing. Furthermore, several random matrix ensembles with sufficient measurements (d) satisfy RIP. Formally, if $d \geq \frac{1}{\delta_k^2} kn \log n$ and each entry of A_i is sampled i.i.d. from a 0-mean sub-Gaussian distribution then k-RIP is satisfied with constant δ_k.

Now, we present our result for Algorithm 1 when applied to the RIP-based matrix sensing:

Theorem 3.2. *Let $M = U^* \Sigma^* {V^*}^\dagger$ be a rank-k matrix with non zero singular values $\sigma_1^* \geq \sigma_2^* \geq \cdots \geq \sigma_k^*$. Also, let $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \to \mathbb{R}^d$ be a linear measurement operator that satisfies $2k$-RIP with RIP constant $\delta_{2k} < \frac{\sigma_k}{10k\sigma_1}$. Suppose, AltMin algorithm (Algorithm 1) is supplied inputs $\mathcal{A}$, $b = \mathcal{A}(M)$. Then, the $(t+1)$-th iterates U^{t+1}, V^{t+1} of the AltMin algorithm satisfy:*

$$\begin{aligned} \text{dist}\left(V^{t+1}, V^*\right) &\leq \alpha_V \cdot \text{dist}\left(U^t, U^*\right) \; \textit{and} \\ \text{dist}\left(U^{t+1}, U^*\right) &\leq \alpha_U \cdot \text{dist}\left(V^{t+1}, V^*\right) \end{aligned}$$

where $\alpha_V, \alpha_U < 1$ are some constants depending only on δ_{2k} and $\text{dist}\left(U, W\right)$ denotes the principal angle based distance between subspaces spanned by the columns of $U \in \mathbb{R}^{m \times k}$ and $W \in \mathbb{R}^{m \times k}$ (See Definition 3.3).

Definition 3.3. *[GL96] Given two orthonormal column matrices U and W, the distance between the subspaces spanned by the columns of U and W is given by:*

$$\text{dist}\left(U, W\right) \overset{def}{=} \|U_\perp^\dagger W\|_2 = \|W_\perp^\dagger U\|_2$$

where $U_\perp^\dagger$ and $W_\perp^\dagger$ are orthonormal basis of the spaces Span $(U)^\perp$ and Span $(W)^\perp$ respectively.

The key idea of the proof of Theorem 3.2 is that when $\delta_{2k} = 0$, AltMin is the same as Orthogonal Iteration (a generalization of the power method used to calculate the k largest singular vectors of a matrix; please refer [GL96], Chapter 8.2.4). However, in our case $\delta_{2k} \neq 0$ but is only a small constant. The key observation here is that in this case, AltMin can be viewed as a noisy version of orthogonal iteration with the noise depending on δ_{2k}. For δ_{2k} small enough, the noise decreases in each step leading to a linear convergence of the iterates to the true matrix.

We also prove a similar theorem for the matrix completion problem. That is, assuming incoherence and by observing a "large" enough number of random entries of the underlying matrix, alternating minimization can recover the true matrix. Please refer to the full version of this paper for precise guarantees.

4 Conclusion

In addition to providing theoretical justification to using alternating minimization for matrix sensing, our result suggests many directions for future work. There are many problems in machine learning (for e.g., sparse + low rank completion) where alternating minimization has been observed to perform well. Our methods in this paper may be useful in obtaining theoretical guarantees on the performance of alternating minimization for these problems as well. More generally, our result shows that non-convex optimization techniques may prove much more efficient than convex optimization techniques in solving some problems which motivates us to increase our efforts in understanding various non-convex optimization heuristics.

References

[Bra03] Matthew Brand. *Fast online SVD revisions for lightweight recommender systems*. In *SIAM International Conference on Data Mining*. 2003.

[CR09] Emmanuel J. Candès and Benjamin Recht. *Exact matrix completion via convex optimization*. Foundations of Computational Mathematics, 9(6):717–772, December 2009. `arXiv:0805.4471`.

[CT05] Emmanuel J. Candès and Terence Tao. *Decoding by linear programming*. IEEE Transactions on Information Theory, 51(12):4203–4215, 2005. `doi:10.1109/TIT.2005.858979`.

[GB00] Karolos M. Grigoriadis and Eric B. Beran. *Alternating projection algorithms for linear matrix inequalities problems with rank constraints*. Advances in linear matrix inequality methods in control: advances in design and control, pages 251–267, 2000.

[GL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

[JMD10] Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. *Guaranteed rank minimization via singular value projection*. In *NIPS*. 2010.

[Kor08] Yehuda Koren. *Factorization meets the neighborhood: a multifaceted collaborative filtering model*. In *KDD*, pages 426–434. 2008. `doi:10.1145/1401890.1401944`.

[MJCD08] Raghu Meka, Prateek Jain, Constantine Caramanis, and Inderjit S. Dhillon. *Rank minimization via online learning*. In *ICML*, pages 656–663. 2008. `doi:10.1145/1390156.1390239`.

[RFP07] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, 2007. To appear in SIAM Review. `arXiv:0706.4138`.