Nonconvex Matrix Factorization from Rank-One Measurements

Yuanxin Li CMU
Cong Ma Princeton
Yuxin Chen Princeton
Yuejie Chi CMU

Abstract

We consider the problem of recovering low-rank matrices from random rank-one measurements, which spans numerous applications including phase retrieval, quantum state tomography, and learning shallow neural networks with quadratic activations, among others. Our approach is to directly estimate the low-rank factor by minimizing a nonconvex least-squares loss function via vanilla gradient descent, following a tailored spectral initialization. When the true rank is small, this algorithm is guaranteed to converge to the ground truth (up to global ambiguity) with near-optimal sample and computational complexities with respect to the problem size. To the best of our knowledge, this is the first theoretical guarantee that achieves near optimality in both metrics. In particular, the key enabler of near-optimal computational guarantees is an implicit regularization phenomenon: without explicit regularization, both spectral initialization and the gradient descent iterates automatically stay within a region incoherent with the measurement vectors. This feature allows one to employ much more aggressive step sizes compared with the ones suggested in prior literature, without the need of sample splitting.

1 Introduction

This paper is concerned with estimating a low-rank positive semidefinite matrix $\boldsymbol{M}^\natural \in \mathbb{R}^{n\times n}$ from a few *rank-one* measurements. Specifically, suppose that the matrix of interest can be factorized as $\boldsymbol{M}^\natural = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top} \in \mathbb{R}^{n\times n}$, where $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ ($r \ll n$) denotes the low-rank factor. We collect m measurements $\{y_i\}_{i=1}^m$ of $\boldsymbol{M}^\natural$ taking the form

$$y_i = \boldsymbol{a}_i^\top \boldsymbol{M}^\natural \boldsymbol{a}_i = \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2, \quad i = 1, \cdots, m, \qquad (1)$$

where $\boldsymbol{a}_i \in \mathbb{R}^n$, $1 \le i \le m$, represent the measurement vectors known *a priori*. One can think of $\{\boldsymbol{a}_i\boldsymbol{a}_i^\top\}_{i=1}^m$ as a set of rank-one linear sensing matrices such that $y_i = \langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{M}^\natural\rangle$. The goal is to recover $\boldsymbol{M}^\natural$, or equivalently, the low-rank factor $\boldsymbol{X}^\natural$, from a limited number of rank-one measurements. This problem spans a variety of important practical applications. We list a few examples below.

- **Phase retrieval and mixed linear regression.** The above problem subsumes as a special case the phase retrieval problem (Candès et al., 2015), which aims to estimate an unknown signal $\boldsymbol{x}^\natural \in \mathbb{R}^n$ from intensity measurements (which can often be modeled or approximated by quadratic measurements of the form $y_i = (\boldsymbol{a}_i^\top \boldsymbol{x}^\natural)^2$). The phase retrieval problem has found numerous applications in X-ray crystallography, optical imaging, astronomy, etc. Another related task in machine learning is mixed linear regression with two components, where the data one collects are generated from one of two unknown regressors; see Chen et al. (2014) for a precise formulation.

- **Quantum state tomography.** Estimating the density operator of a quantum system can be formulated as recovering a low-rank positive semidefinite matrix from rank-one measurements, when the density operator is *almost pure* (Kueng et al., 2017). A problem of similar mathematical formulation occurs in phase space tomography (Tian et al., 2012) in optics.

- **Learning shallow polynomial neural networks.** Treating $\{\boldsymbol{a}_i, y_i\}_{i=1}^m$ as training data, our problem is equivalent to learning a one-hidden-layer, fully-connected neural network with quadratic activation functions (Livni et al., 2014; Soltanolkotabi et al., 2017; Soltani and Hegde, 2018; Du and Lee, 2018). Here, the output of the network is expressed as $y_i = \sum_{j=1}^r \sigma(\boldsymbol{a}_i^\top \boldsymbol{x}_j^\natural)$ with the weight matrix $\boldsymbol{X}^\natural = [\boldsymbol{x}_1^\natural, \boldsymbol{x}_2^\natural, \cdots, \boldsymbol{x}_r^\natural] \in \mathbb{R}^{n\times r}$ and the quadratic activation function $\sigma(z) = z^2$.

- **Covariance sketching.** Consider a zero-mean data stream $\{\boldsymbol{x}_t\}_{t\in\mathcal{T}}$, whose covariance matrix $\boldsymbol{M}^\natural := \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]$ is (approximately) low-rank. To estimate the covariance matrix, one can collect m aggregated quadratic sketches of the form

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

$y_i = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2$, which converges to $\mathbb{E}[(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2] = \boldsymbol{a}_i^\top \boldsymbol{M}^\natural \boldsymbol{a}_i$ as the number of data instances grows. This quadratic covariance sketching scheme can be performed under minimal storage requirement and low sketching cost. See Chen et al. (2015) for detailed descriptions.

1.1 Main Contributions

To recover $\boldsymbol{X}^\natural$, we consider the following natural least-squares empirical risk minimization problem

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times r}}{\text{minimize}}\ f(\boldsymbol{X}) := \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - \left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2\right)^2. \tag{2}$$

Due to the quadratic nature of the measurements, the loss function (2) is highly nonconvex and in general challenging to solve. The problem, however, becomes tractable under certain random designs, and may even be efficiently solved using simple methods like gradient descent. Our main finding is the following: *under i.i.d. Gaussian design (i.e. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$), vanilla gradient descent with spectral initialization achieves appealing performance guarantees both statistically and computationally*.

- Statistically, we show that gradient descent converges exactly to the true factor $\boldsymbol{X}^\natural$ (modulo unrecoverable global ambiguity), as soon as the number of measurements exceeds the order of $O(nr^4\log n)$. When r is fixed and independent of n, this sample complexity is near-optimal up to some logarithmic factor.
- Computationally, to achieve relative ϵ-accuracy, gradient descent requires an iteration complexity of $O(r^2\log(1/\epsilon))$ (up to logarithmic factors), with a per-iteration cost of $O(mnr)$. When r is fixed and independent of n, the computational complexity scales linearly with mn, which is proportional to the time taken to read all data.

These findings significantly improve upon existing results that require either resampling (Zhong et al., 2015; Lin and Ye, 2016; Soltani and Hegde, 2018)[1], or high iteration / computational complexity (Sanghavi et al., 2017). In particular, our work is most related to Sanghavi et al. (2017) which also studied the effectiveness of gradient descent. The results in Sanghavi et al. (2017) require a sample complexity on the order of $O(nr^6\log^2 n)$, as well as an iteration complexity of $O(n^4r^2\log(1/\epsilon))$ (up to logarithmic factors) to attain ϵ-accuracy. In comparison, our theory improves the sample complexity to $O(nr^4\log n)$ and, perhaps more importantly, establishes a much lower iteration complexity of $O(r^2\log(1/\epsilon))$ (up to logarithmic factor). To the best of our knowledge, this work is the first algorithm (without resampling) that achieves both near-optimal statistical and computational guarantees with respect to n, for the problem of low-rank matrix recovery from rank-one measurements.

[1] Algorithms with resampling are easier for theoretical analysis but are not sample-efficient and rarely adopted in practice if at all.

1.2 Effectiveness of Gradient Descent

Recently, gradient descent has been widely employed to address various nonconvex optimization problems, due to its appealing statistical and computational efficiency. Despite the nonconvexity of (2), Sanghavi et al. (2017) showed that within a local neighborhood of $\boldsymbol{X}^\natural$:

$$\left\{\boldsymbol{X}\ \middle|\ \left\|\boldsymbol{X}-\boldsymbol{X}^\natural\right\|_{\mathsf{F}} \le \frac{1}{24}\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}\right\}, \tag{3}$$

$f(\boldsymbol{X})$ behaves like a strongly convex function, at least along certain descent directions. Here, $\sigma_i(\boldsymbol{X}^\natural)$ denotes the ith largest singular value of $\boldsymbol{X}^\natural$. However, strong convexity alone is not enough to guarantee computational efficiency. One still needs to take smoothness into account. In fact, the smoothness parameter derived in Sanghavi et al. (2017) is as large as n^2 (even ignoring additional polynomial factors in r), thus leading to a step size as small as $O(1/n^4)$ and an iteration complexity of $O(n^4\log(1/\epsilon))$. This step-size choice is fairly conservative and leads to prohibitive computation burdens when n is large.

One way to improve the computational guarantee is to employ appropriately designed regularization operations — such as truncation (Chen and Candès, 2017) and projection (Chen and Wainwright, 2015). These explicit regularization operations are capable of stabilizing the search directions, and ensure the whole trajectory lies in a neighborhood surrounding the ground truth with well controlled strong convexity and smoothness properties. However, such explicit regularizations complicate algorithm implementations, as they introduce more (and often unnecessary) tuning parameters.

In this paper, we demonstrate that vanilla gradient descent is almost as effective as its regularized counterparts. In fact, even without explicit regularization, the iterates always follow a path within some region with nice geometric structures, which enables fast convergence. To be more precise, we first specify the region which enjoys the desired geometric properties. Consider a local region around $\boldsymbol{X}^\natural$ where $\boldsymbol{X}$ is "incoherent"[2] with all sensing vectors in the sense that

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2 \le \frac{\sqrt{\log n}}{24}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}. \tag{4}$$

We term the intersection of (3) and (4) the *Region of Incoherence and Contraction* (RIC). The nice feature of the

[2] This is called incoherent because if $\boldsymbol{X}$ is aligned (and hence coherent) with the sensing vectors, $\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}-\boldsymbol{X}^\natural\right)\right\|_2$ can be $O(\sqrt{n})$ times larger than the right-hand side of (4).

RIC is that: in addition to the (restricted) strong convexity, the loss function $f(\cdot)$ in this region enjoys a smoothness parameter that scales as $O(\max\{r, \log n\})$ (namely, $\|\nabla^2 f(\boldsymbol{X})\| \lesssim \max\{r, \log n\}$), which is much smaller than $O(n^2)$ provided in Sanghavi et al. (2017). This benign geometric property of RIC enables gradient descent to linearly converge to the ground truth, as long as the iterates stay within RIC.

A key contribution of our work is to demonstrate that the trajectory of vanilla gradient descent never leaves RIC as if it is "implicitly regularized". Such a statement is, unfortunately, not guaranteed by standard optimization theory, which only ensures contraction of the Euclidean error. Rather, we need to exploit the statistical model of data generation, taking into consideration of the "homogeneity" of the samples together with the finite-sum form of the loss function. Specifically, we resort to the leave-one-out trick (Ma et al., 2017; Zhong and Boumal, 2017; Chen et al., 2017) that produces auxiliary trajectories of gradient descent that use all but one sample as a proof strategy. This allows us to establish the incoherence condition (4) by leveraging the statistical independence of the leave-one-out trajectory w.r.t. the corresponding sensing vector that has been left out. Our theory refines existing leave-one-out arguments in Ma et al. (2017) and further establishes linear contraction in terms of the entry-wise prediction error. In sum, our work highlights the substantial gain of jointly considering optimization and statistics in understanding learning algorithms.

2 Algorithms and Main Results

To begin with, we present the formal problem setup. Suppose we are given a set of m rank-one measurements as given in (1), where $\boldsymbol{a}_i \in \mathbb{R}^n$ is the ith sensing vector composed of i.i.d. standard Gaussian entries, i.e. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, for $i = 1, \cdots, m$. The underlying ground truth $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ is assumed to have full column rank but not necessarily orthogonal columns. Define the condition number of $\boldsymbol{M}^\natural = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ as $\kappa = \sigma_1^2(\boldsymbol{X}^\natural)/\sigma_r^2(\boldsymbol{X}^\natural)$. Throughout this paper, we assume the condition number is bounded by some constant independent of n and r, i.e. $\kappa = O(1)$. Our goal is to recover $\boldsymbol{X}^\natural$, up to (unrecoverable) orthonormal transformations, from the measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$ in a statistically and computationally efficient manner.

2.1 Vanilla Gradient Descent with Spectral Initialization

The algorithm studied herein is a combination of vanilla gradient descent and a carefully-designed spectral initialization. Specifically, we attempt to minimize the nonconvex loss function (2) iteratively via gradient descent

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \mu_t \nabla f(\boldsymbol{X}_t), \qquad t = 0, 1, \cdots, \tag{5}$$

Algorithm 1: Gradient Descent with Spectral Initialization

Input: Measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$, and sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$.

Parameters: Step size μ_t, rank r, and number of iterations T.

Initialization: Set $\boldsymbol{X}_0 = \boldsymbol{Z}_0 \boldsymbol{\Lambda}_0^{1/2}$, where the columns of $\boldsymbol{Z}_0 \in \mathbb{R}^{n\times r}$ contain the normalized eigenvectors corresponding to the r largest eigenvalues of the matrix

$$\boldsymbol{Y} = \frac{1}{2m}\sum_{i=1}^m y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top, \tag{7}$$

and $\boldsymbol{\Lambda}_0$ is an $r \times r$ diagonal matrix, with the entries on the diagonal given as

$$[\boldsymbol{\Lambda}_0]_i = \lambda_i\big(\boldsymbol{Y}\big) - \lambda, \quad i = 1, \cdots, r, \tag{8}$$

where $\lambda = \frac{1}{2m}\sum_{i=1}^m y_i$ and $\lambda_i(\boldsymbol{Y})$ is the ith largest eigenvalue of $\boldsymbol{Y}$.

Gradient descent: For $t = 0 : 1 : T-1$, do

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \mu_t \cdot \frac{1}{m}\sum_{i=1}^m \Big(\big\|\boldsymbol{a}_i^\top \boldsymbol{X}_t\big\|_2^2 - y_i\Big)\boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{X}_t. \tag{9}$$

Output: $\boldsymbol{X}_T$.

where $\boldsymbol{X}_t$ denotes the tth iterate, μ_t is the step size / learning rate, and the gradient $\nabla f(\boldsymbol{X})$ is given by

$$\nabla f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^m \big(\big\|\boldsymbol{a}_i^\top \boldsymbol{X}\big\|_2^2 - y_i\big)\boldsymbol{a}_i \boldsymbol{a}_i^\top \boldsymbol{X}. \tag{6}$$

For initialization, we apply the spectral method, which sets the columns of $\boldsymbol{X}_0$ as the top-r eigenvectors — properly scaled — of a matrix $\boldsymbol{Y}$ as defined in (7). The rationale is this: the mean of $\boldsymbol{Y}$ is given by $\mathbb{E}[\boldsymbol{Y}] = \frac{1}{2}\big\|\boldsymbol{X}^\natural\big\|_{\mathsf{F}}^2 \boldsymbol{I}_n + \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$, and hence the principal components of $\boldsymbol{Y}$ form a reasonable estimate of $\boldsymbol{X}^\natural$, provided that there are sufficiently many samples. The full algorithm is described in Algorithm 1.

2.2 Performance Guarantees

Before proceeding to our main results, we pause here to introduce the metric used to assess the estimation error of the running iterates. Since $(\boldsymbol{X}^\natural \boldsymbol{P})(\boldsymbol{X}^\natural \boldsymbol{P})^\top = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ for any orthonormal matrix $\boldsymbol{P} \in \mathbb{R}^{r\times r}$, $\boldsymbol{X}^\natural$ is recoverable only up to orthonormal transforms. Hence, we define the error of the tth iterate $\boldsymbol{X}_t$ as

$$\mathrm{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big) = \big\|\boldsymbol{X}_t \boldsymbol{Q}_t - \boldsymbol{X}^\natural\big\|_{\mathsf{F}}, \tag{10}$$

where $\boldsymbol{Q}_t$ is the best orthonormal transformation, i.e.

$$\boldsymbol{Q}_t := \mathrm{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}} \big\|\boldsymbol{X}_t \boldsymbol{P} - \boldsymbol{X}^\natural\big\|_{\mathsf{F}}. \tag{11}$$

Here $\mathcal{O}^{r\times r}$ denotes the set of all $r\times r$ orthonormal matrices. Accordingly, we have the following theoretical performance guarantees of Algorithm 1.

Theorem 1. *Suppose that $m \geq Cnr^3(r+\sqrt{\kappa})\kappa^3\log n$ for some large enough constant $C>0$, and that the step size obeys $0<\mu_t\equiv\mu=\frac{c_4}{(r\kappa+\log n)^2\sigma_r^2(\boldsymbol{X}^\natural)}$ for some sufficiently small constant $c_4>0$. Then with probability at least $1-O(mn^{-7})$, the iterates satisfy*

$$\operatorname{dist}\big(\boldsymbol{X}_t,\boldsymbol{X}^\natural\big)\leq c_1\left(1-0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \quad (12)$$

for all $t\geq 0$. In addition,

$$\max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\big(\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}^\natural\big)\right\|_2 \leq c_2\left(1-0.5\mu\sigma_r^2(\boldsymbol{X}^\natural)\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \quad (13)$$

for all $0\leq t\leq c_3n^5$. Here, c_1,c_2,c_3 are some universal positive constants.

A few remarks regarding Theorem 1 are in order.

- *Near-optimal sample complexity when r is fixed:* Theorem 1 suggests that spectrally-initialized vanilla gradient descent succeeds as soon as $m=O(nr^4\log n)$. When $r=O(1)$, this leads to near-optimal sample complexity up to a logarithmic factor. To the best of our knowledge, this outperforms all performance guarantees in the literature for any nonconvex method without requiring *resampling*.
- *Near-optimal computational complexity:* In order to achieve ϵ-accuracy, i.e. $\operatorname{dist}(\boldsymbol{X}_t,\boldsymbol{X}^\natural)\leq\epsilon\|\boldsymbol{X}^\natural\|_{\mathsf{F}}$, it suffices to run gradient descent for $T=O\left(r^2\mathrm{polylog}(n)\log(1/\epsilon)\right)$ iterations. This results in a total computational complexity of $O\left(mnr^3\mathrm{polylog}(n)\log(1/\epsilon)\right)$, which is proportional to the time taken to read all data when $r=O(1)$.
- *Implicit regularization:* Theorem 1 demonstrates that both the spectral initialization and the gradient descent updates provably control the sample-wise error $\max_{1\leq l\leq m}\|\boldsymbol{a}_l^\top\big(\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}^\natural\big)\|_2$, and the iterates remain incoherent with respect to all the sensing vectors. In fact, the sample-wise error decreases linearly as well, which is not characterized in earlier work for phase retrieval (Ma et al., 2017).

3 Related Work

Instead of directly estimating $\boldsymbol{X}^\natural$, the problem of interest can be also solved by estimating $\boldsymbol{M}^\natural=\boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}$ in higher dimension via nuclear norm minimization, which requires $O(nr)$ measurements for exact recovery (Chen et al., 2015; Cai and Zhang, 2015; Kueng et al., 2017; Li et al., 2017). See also Candès et al. (2013); Candès and Li (2014); Demanet and Hand (2014); Waldspurger et al. (2015) for the phase retrieval problem. However, nuclear norm minimization, often cast as the semidefinite programming, is in general computationally expensive when dealing with large-scale data.

On the other hand, nonconvex approaches have drawn intense attention in the past decade due to their ability to achieve computational and statistical efficiency all at once Chi et al. (2018). Specifically, for the phase retrieval problem, Wirtinger Flow (WF) and its variants (Candès et al., 2015; Chen and Candès, 2017; Cai et al., 2016; Ma et al., 2017; Zhang et al., 2017; Soltanolkotabi, 2017; Wang et al., 2017) have been proposed. As a two-stage algorithm, it consists of spectral initialization and iterative gradient updates. This strategy has found enormous success in solving other problems such as low-rank matrix recovery and completion (Chen and Wainwright, 2015; Tu et al., 2016), blind deconvolution (Li et al., 2016), and spectral compressed sensing (Cai et al., 2017). We follow a similar route but analyze a more general problem that includes phase retrieval as a special case.

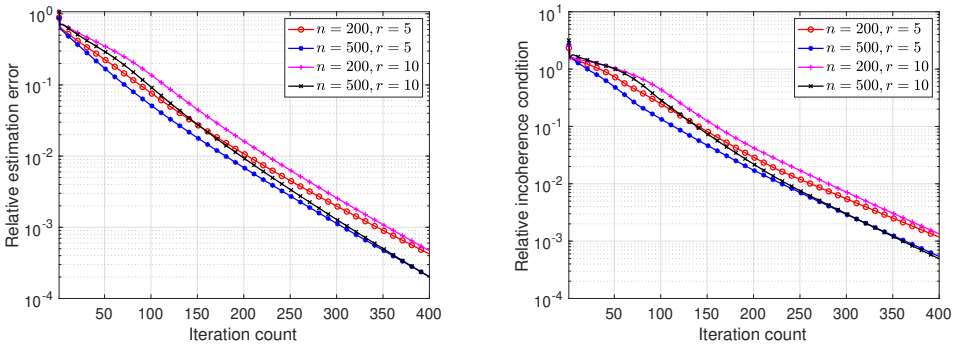
The paper by Sanghavi et al. (2017) is most close to our work, which studied the local convexity of the same loss function and developed performance guarantees for gradient descent using a similar, but different spectral initialization scheme. As discussed earlier, due to the pessimistic estimate of the smoothness parameter, they only allow a diminishing learning rate (or step size) of $O(1/n^4)$, leading to a high iteration complexity. We not only provide stronger computational guarantees, but also improve the sample complexity, compared with Sanghavi et al. (2017).

Several other existing works have suggested different approaches for low-rank matrix factorization from rank-one measurements, of which the statistical and computational guarantees to reach ϵ-accuracy are summarized in Table 1. We note our guarantee is the only one that achieves simultaneous near-optimal sample complexity and computational complexity. Iterative algorithms based on alternating minimization or noisy power iterations (Zhong et al., 2015; Lin and Ye, 2016; Soltani and Hegde, 2018) require a *fresh* set of samples at every iteration, which is never executed in practice and usually leads to much easier analysis due to statistical independence, and the sample complexity grows unbounded for *exact* recovery.

Our model is also related to learning shallow neural networks. Zhong et al. (2017) studied the performance of gradient descent with resampling and an initialization provided by the tensor method for various activation functions, however their analysis did not cover quadratic activations. For quadratic activations, Livni et al. (2014) adopts a greedy learning strategy, and can only guarantee sublinear convergence rate. Moreover, Soltanolkotabi et al.

Algorithms with resampling	Sample complexity	Computational complexity
AltMin-LRROM (Zhong et al., 2015)	$O(nr^4 \log^2 n \log(\frac{1}{\epsilon}))$	$O(mnr \log(\frac{1}{\epsilon}))$
gFM (Lin and Ye, 2016)	$O(nr^3 \log(\frac{1}{\epsilon}))$	$O(mnr \log(\frac{1}{\epsilon}))$
EP-ROM (Soltani and Hegde, 2018)	$O(nr^2 \log^4 n \log(\frac{1}{\epsilon}))$	$O(mn^2 \log(\frac{1}{\epsilon}))$
AP-ROM (Soltani and Hegde, 2018)	$O(nr^3 \log^4 n \log(\frac{1}{\epsilon}))$	$O(mnr \log n \log(\frac{1}{\epsilon}))$
Algorithms without resampling	**Sample complexity**	**Computational complexity**
Convex (Chen et al., 2015)	$O(nr)$	$O(mn^2 \frac{1}{\sqrt{\epsilon}})$
GD (Sanghavi et al., 2017)	$O(nr^6 \log^2 n)$	$O(mn^5 r^3 \log^4 n \log(\frac{1}{\epsilon}))$
GD (Algorithm 1, Ours)	$O(nr^4 \log n)$	$O(mnr \max\{\log^2 n, r^2\} \log(\frac{1}{\epsilon}))$

Table 1: Comparisons with existing results in terms of sample complexity and computational complexity to reach ϵ-accuracy. The top half of the table is concerned with algorithms that require resampling, while the bottom half of the table covers algorithms without resampling.



(a) Relative estimation error (b) Relative incoherence condition

Figure 1: Performance of the proposed algorithm in regard to (a) relative estimation error, and (b) relative incoherence condition with respect to the iteration count using different problem sizes, when $m = 5nr$.

(2017); Du and Lee (2018) studied the optimization landscape for an over-parameterized shallow neural network with quadratic activation, where r is larger than n.

4 Numerical Experiments

In this section, we provide several numerical experiments to validate the effective and efficient performance of the proposed algorithm. During each experiment, given a pair of (n, r), the ground truth $\boldsymbol{X}^\natural \in \mathbb{R}^{n\times r}$ is generated with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries. We first examine the relative estimation error $\mathrm{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big)/\|\boldsymbol{X}^\natural\|_{\mathsf{F}}$ and the relative incoherence condition $\max_{1\le l\le m} \|\boldsymbol{a}_l^\top \big(\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\big)\|_2/\|\boldsymbol{X}^\natural\|_{\mathsf{F}}$ with respect to the iteration count using a constant step size $\mu_t = 0.03$, where the number of measurements is set as $m = 5nr$. The convergence rates in Figure 1 are approximately linear, validating our theory.

We then examine the phase transitions of the proposed algorithm with respect to the number of measurements. Multiple Monte Carlo trials are conducted, and each trial is deemed a success if the relative estimation error is less than 10^{-6} within $T = 1000$ iterations. Figure 2 depicts the success rate over 20 trials, where the proposed algorithm successfully recovers the ground truth as soon as the number of measurements is about 4 times above the degrees of freedom nr. These results suggest that the required sample complexity scales linearly with the degrees of freedom, and our theoretical guarantees are near-optimal up to logarithmic factors.

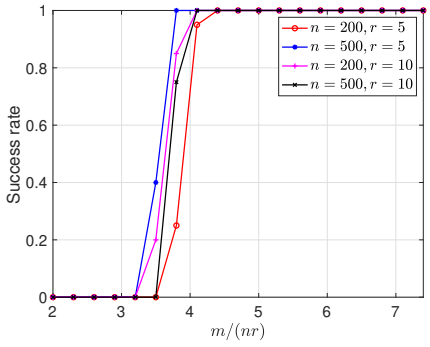


Figure 2: The success rate of the proposed algorithm with respect to the number of measurements $m/(nr)$ using different problem sizes.

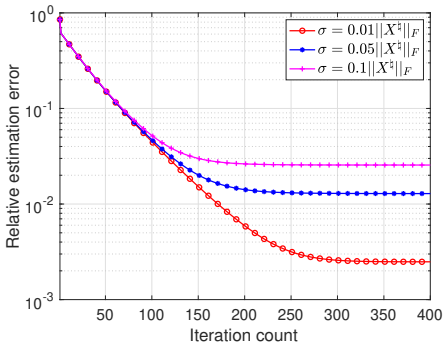


Figure 3: Relative estimation error with respect to iteration count in different noisy levels, when $n = 500$, $r = 5$ and $m = 5nr$.

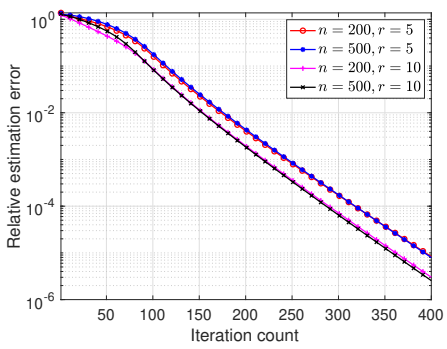


Figure 4: Relative estimation error with respect to the iteration count using different problem sizes when the sensing vectors are generated from sub-Gaussian distributions and a random initialization is employed, when $m = 5nr$.

Next, we numerically verify the stability of the proposed algorithm against additive noise, where each measurement is given as $y_i = \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 + \epsilon_i$, where the noise ϵ_i is generated i.i.d. from $\mathcal{N}(0, \sigma^2)$. Figure 3 shows the estimation error with respect to the iteration count at different noise levels when $n = 500$, $r = 5$ and $m = 5nr$. As the noise variance σ^2 increases, the performance of the proposed algorithm degenerates smoothly.

Finally, we test the performance of the proposed algorithm when the measurement vectors $\boldsymbol{a}_i$'s are i.i.d. generated from a sub-Gaussian distribution under random initialization. Specifically, we consider a case where each entry in $\boldsymbol{a}_i$ is drawn i.i.d. from a uniform distribution $\mathcal{U}[-1, 1]$. We then implement gradient descent with a constant step size $\mu_t = 0.5$ starting from a random initialization, whose entries are generated i.i.d. following $\mathcal{N}(0, \frac{1}{n})$. Figure 4 shows the appealing convergence performance of the proposed algorithm.

5 Outline of Theoretical Analysis

This section provides the proof sketch of the main results. Our theoretical analysis is inspired by the work of Ma et al. (2017) for phase retrieval and follows the general recipe outlined therein. However, significant changes and efforts are needed when dealing with rank-r matrices, compared with rank-one matrices in Ma et al. (2017). In addition, we refine the analysis to show that both the signal reconstruction error (see (12)) and the entry-wise error (see (13)) contract linearly, where the latter is not revealed by Ma et al. (2017). In below, we first characterize a region of incoherence and contraction that enjoys both strong convexity and smoothness along certain directions. We then demonstrate — via an induction argument — that the iterates always stay within this nice region. Finally, the proof is complete by validating the desired properties of spectral initialization.

5.1 Local Geometry and Error Contraction

We start with characterizing a local region around $\boldsymbol{X}^\natural$, within which the loss function enjoys desired restricted strong convexity and smoothness properties. This requires exploring the property of the Hessian of $f(\boldsymbol{X})$, which is given by

$$\nabla^2 f(\boldsymbol{X}) = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\left\|\boldsymbol{a}_i^\top \boldsymbol{X}\right\|_2^2 - y_i\right)\boldsymbol{I}_r + 2\boldsymbol{X}^\top \boldsymbol{a}_i\boldsymbol{a}_i^\top \boldsymbol{X}\right] \otimes \left(\boldsymbol{a}_i\boldsymbol{a}_i^\top\right). \tag{14}$$

Here, we use $\otimes$ to denote the Kronecker product and hence $\nabla^2 f(\boldsymbol{X}) \in \mathbb{R}^{nr\times nr}$. Now we are ready to state the following lemma regarding the properties of the Hessian in a local region around $\boldsymbol{X}^\natural$, which will be referred to as the region of incoherence and contraction (RIC) throughout this paper.

Lemma 1. *Suppose the sample size obeys* $m \geq c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)} nr\log(n\kappa)$ *for some sufficiently large constant* $c > 0$. *Then with probability at least* $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$, *we have*

$$\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f(\boldsymbol{X})\mathrm{vec}\left(\boldsymbol{V}\right) \geq 1.026\sigma_r^2(\boldsymbol{X}^\natural)\left\|\boldsymbol{V}\right\|_{\mathsf{F}}^2, \tag{15}$$

and

$$\left\|\nabla^2 f(\boldsymbol{X})\right\| \leq 1.5\sigma_r^2(\boldsymbol{X}^\natural)\log n + 6\left\|\boldsymbol{X}^\natural\right\|_{\mathsf{F}}^2 \tag{16}$$

hold simultaneously for all matrices $\boldsymbol{X}$ *and* $\boldsymbol{V}$ *satisfying the following constraints:*

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_{\mathsf{F}} \leq \frac{1}{24}\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{17a}$$

$$\max_{1\leq l\leq m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X} - \boldsymbol{X}^\natural\right)\right\|_2 \leq \frac{1}{24}\sqrt{\log n}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{17b}$$

and $\boldsymbol{V} = \boldsymbol{T}_1\boldsymbol{Q}_{\boldsymbol{T}} - \boldsymbol{T}_2$ *satisfying* $\|\boldsymbol{T}_2 - \boldsymbol{X}^\natural\| \le \frac{1}{24}\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|}$, *where* $\boldsymbol{Q}_{\boldsymbol{T}} := \operatorname{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}} \|\boldsymbol{T}_1\boldsymbol{P} - \boldsymbol{T}_2\|_{\mathsf{F}}$. *Here,* c_1 *is some absolute positive constant.*

The condition (17) on $\boldsymbol{X}$ formally characterizes the RIC, which enjoys the claimed restricted strong convexity (see (15)) and smoothness (see (16)). With Lemma 1 in mind, it is easy to see that if $\boldsymbol{X}_t$ lies within the RIC, the estimation error shrinks in the presence of a properly chosen step size. This is given in the lemma below.

Lemma 2. *Suppose the sample size obeys* $m \ge c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)} nr\log(n\kappa)$ *for some sufficiently large constant* $c > 0$. *Then with probability at least* $1 - c_1 n^{-12} - me^{-1.5n} - mn^{-12}$, *if* $\boldsymbol{X}_t$ *falls within the RIC as described in* (17)*, we have*

$$\operatorname{dist}\big(\boldsymbol{X}_{t+1}, \boldsymbol{X}^\natural\big) \le \big(1 - 0.513\mu\sigma_r^2(\boldsymbol{X}^\natural)\big)\operatorname{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big),$$

provided that the step size obeys $0 < \mu_t \equiv \mu \le \frac{1.026\sigma_r^2(\boldsymbol{X}^\natural)}{\left(1.5\sigma_r^2(\boldsymbol{X}^\natural)\log n + 6\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\right)^2}$. *Here,* $c_1 > 0$ *is some universal constant.*

Assuming that the iterates $\{\boldsymbol{X}_t\}$, stay within the RIC (see (17)) for the first T_c iterations, according to Lemma 2, we have, by induction, that

$$\begin{aligned}
&\operatorname{dist}\big(\boldsymbol{X}_{T_c+1}, \boldsymbol{X}^\natural\big) \\
&\le \Big(1 - 0.513\mu\sigma_r^2(\boldsymbol{X}^\natural)\Big)^{T_c+1} \operatorname{dist}\big(\boldsymbol{X}_0, \boldsymbol{X}^\natural\big) \\
&\le \frac{1}{24\sqrt{6}} \cdot \frac{\sqrt{\log n}}{\sqrt{n}} \cdot \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}
\end{aligned}$$

as soon as

$$T_c \ge c\max\left\{\log^2 n, \frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)}\right\}\log n, \tag{18}$$

for some large enough constant c. The iterates after $t \ge T_c$ are easier to deal with; in fact, it is easily seen that $\boldsymbol{X}_{t+1}$ stays in the RIC since

$$\begin{aligned}
&\max_{1\le l\le m} \Big\|\boldsymbol{a}_l^\top\big(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\big)\Big\|_2 \\
&\le \max_{1\le l\le m} \big\|\boldsymbol{a}_l\big\|_2 \|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{X}^\natural\| \\
&\le \sqrt{6n} \cdot \frac{1}{24\sqrt{6}} \cdot \frac{\sqrt{\log n}}{\sqrt{n}} \cdot \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}} \\
&= \frac{1}{24}\sqrt{\log n} \cdot \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}},
\end{aligned}$$

where the second line follows from the Gaussian concentration inequalities. Consequently, contraction of the estimation error $\operatorname{dist}\big(\boldsymbol{X}_t, \boldsymbol{X}^\natural\big)$ can be guaranteed by Lemma 1 for all $t \ge T_c$.

Algorithm 2: Leave-One-Out Versions

Input: Measurements $\{y_i\}_{i:i\ne l}$, and sensing vectors $\{\boldsymbol{a}_i\}_{i:i\ne l}$.

Parameters: Step size μ_t, rank r, and number of iterations T.

Initialization: $\boldsymbol{X}_0^{(l)} = \boldsymbol{Z}_0^{(l)}\boldsymbol{\Lambda}_0^{(l)1/2}$, where the columns of $\boldsymbol{Z}_0^{(l)} \in \mathbb{R}^{n\times r}$ contain the normalized eigenvectors corresponding to the r largest eigenvalues of the matrix $\boldsymbol{Y}^{(l)} = \frac{1}{2m}\sum_{i:i\ne l} y_i\boldsymbol{a}_i\boldsymbol{a}_i^\top$, and $\boldsymbol{\Lambda}_0^{(l)}$ is an $r\times r$ diagonal matrix, with diagonal entries given as $\Big[\boldsymbol{\Lambda}_0^{(l)}\Big]_i = \lambda_i\big(\boldsymbol{Y}^{(l)}\big) - \lambda^{(l)}$, for $i = 1, \cdots, r$, where $\lambda^{(l)} = \frac{1}{2m}\sum_{i:i\ne l} y_i$ and $\lambda_i\big(\boldsymbol{Y}^{(l)}\big)$ is the ith largest eigenvalue of $\boldsymbol{Y}^{(l)}$.

Gradient descent: For $t = 0 : 1 : T-1$, do

$$\boldsymbol{X}_{t+1}^{(l)} = \boldsymbol{X}_t^{(l)} - \mu_t\cdot\frac{1}{m}\sum_{i:i\ne l}\Big(\big\|\boldsymbol{a}_i^\top\boldsymbol{X}_t^{(l)}\big\|_2^2 - y_i\Big)\boldsymbol{a}_i\boldsymbol{a}_i^\top\boldsymbol{X}_t^{(l)}. \tag{20}$$

Output: $\boldsymbol{X}_T^{(l)}$.

5.2 Introducing Leave-One-Out Sequences

It now becomes clear that the key remaining step is to verify that the iterates $\{\boldsymbol{X}_t\}$ satisfy (17) for the first T_c iterations, where T_c is on the order of (18). Verifying (17b) is conceptually hard since the iterates $\{\boldsymbol{X}_t\}$ are statistically dependent with all the sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$. To tackle this problem, for each $1 \le l \le m$, we introduce an auxiliary leave-one-out sequence $\{\boldsymbol{X}_t^{(l)}\}$, which discards a single measurement from consideration. Specifically, the sequence $\{\boldsymbol{X}_t^{(l)}\}$ is the gradient iterates operating on the following leave-one-out function

$$f^{(l)}(\boldsymbol{X}) := \frac{1}{4m}\sum_{i:i\ne l}\Big(y_i - \big\|\boldsymbol{a}_i^\top\boldsymbol{X}\big\|_2^2\Big)^2. \tag{19}$$

See Algorithm 2 for a formal definition of the leave-one-out sequences. Again, we want to emphasize that Algorithm 2 is just an auxiliary procedure useful for the theoretical analysis, and it does not need to be implemented in practice.

5.3 Establishing Incoherence via Induction

Our proof is inductive in nature with the following induction hypotheses:

$$\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}^\natural\|_{\mathsf{F}} \le C_1\Big(1 - 0.5\sigma_r^2(\boldsymbol{X}^\natural)\mu\Big)^t\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{21a}$$

$$\begin{aligned}
&\max_{1\le l\le m}\Big\|\boldsymbol{X}_t\boldsymbol{Q}_t - \boldsymbol{X}_t^{(l)}\boldsymbol{R}_t^{(l)}\Big\|_{\mathsf{F}} \\
&\le C_3\Big(1 - 0.5\sigma_r^2(\boldsymbol{X}^\natural)\mu\Big)^t\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\kappa\|\boldsymbol{X}^\natural\|_{\mathsf{F}}},
\end{aligned} \tag{21b}$$

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l^\top\left(\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}^\natural\right)\right\|_2 \le C_2\left(1-0.5\sigma_r^2(\boldsymbol{X}^\natural)\mu\right)^t\sqrt{\log n}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}, \tag{21c}$$

where $\boldsymbol{R}_t^{(l)}=\operatorname{argmin}_{\boldsymbol{P}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}_t\boldsymbol{Q}_t-\boldsymbol{X}_t^{(l)}\boldsymbol{P}\right\|_{\mathsf{F}}$, and the positive constants C_1, C_2 and C_3 satisfy

$$C_1+C_3\le\frac{1}{24},\quad C_2+\sqrt{6}C_3\le\frac{1}{24}, \tag{22a}$$

$$5.86C_1+29.3C_3+5\sqrt{6}C_3\le C_2. \tag{22b}$$

Furthermore, the step size μ is chosen as

$$\mu=\frac{c_0\sigma_r^2(\boldsymbol{X}^\natural)}{\left(\sigma_r^2(\boldsymbol{X}^\natural)\log n+\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^2\right)^2} \tag{23}$$

with appropriate universal constant $c_0>0$.

Our goal is to show that if the tth iteration $\boldsymbol{X}_t$ satisfies the induction hypotheses (21), then the $(t+1)$th iteration $\boldsymbol{X}_{t+1}$ also satisfies (21). It is straightforward to see that the hypothesis (21a) has already been established by Lemma 2, and we are left with (21b) and (21c). We first establish (21b) in the following lemma, which measures the proximity between $\boldsymbol{X}_t$ and the leave-one-out versions $\boldsymbol{X}_t^{(l)}$.

Lemma 3. *Suppose the sample size obeys $m\ge c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)}nr\log(n\kappa)$ for some sufficiently large constant $c>0$. If the induction hypotheses* (21) *hold for the tth iteration, with probability at least $1-c_1n^{-12}-me^{-1.5n}-mn^{-12}$, we have*

$$\max_{1\le l\le m}\left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_{t+1}^{(l)}\boldsymbol{R}_{t+1}^{(l)}\right\|_{\mathsf{F}} \le C_3\left(1-0.5\sigma_r^2(\boldsymbol{X}^\natural)\mu\right)^{t+1}\sqrt{\frac{\log n}{n}}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\kappa\|\boldsymbol{X}^\natural\|_{\mathsf{F}}},$$

as long as the step size obeys (23)*. Here, $c_1>0$ is some absolute constant.*

In addition, the incoherence property of $\boldsymbol{X}_{t+1}^{(l)}$ with respect to the lth sensing vector $\boldsymbol{a}_l$ is relatively easier to establish, due to their statistical independence. Combined with the proximity bound from Lemma 3, this allows us to justify the incoherence property of the original iterates $\boldsymbol{X}_{t+1}$, as summarized in the lemma below.

Lemma 4. *Suppose the sample size obeys $m\ge c\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^4}{\sigma_r^4(\boldsymbol{X}^\natural)}nr\log(n\kappa)$ for some sufficiently large constant $c>0$. If the induction hypotheses* (21) *hold for the tth iteration, with probability exceeding $1-c_1n^{-12}-me^{-1.5n}-2mn^{-12}$,*

$$\max_{1\le l\le m}\|\boldsymbol{a}_l^\top(\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}^\natural)\|_2 \le C_2\left(1-0.5\sigma_r^2(\boldsymbol{X}^\natural)\mu\right)^{t+1}\sqrt{\log n}\cdot\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}$$

holds as long as the step size satisfies (23)*. Here, $c_1>0$ is some universal constant.*

5.4 Spectral Initialization

Finally, it remains to verify that the induction hypotheses hold for the initialization, i.e. the base case when $t=0$. This is supplied by the following lemma.

Lemma 5. *Suppose that the sample size exceeds $m\ge c\max\{\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}}{\sigma_r(\boldsymbol{X}^\natural)}\sqrt{r},\kappa\}\frac{\|\boldsymbol{X}^\natural\|_{\mathsf{F}}^5}{\sigma_r^5(\boldsymbol{X}^\natural)}n\sqrt{r}\log n$ for some sufficiently large constant $c>0$. Then $\boldsymbol{X}_0$ satisfies* (21) *with probability at least $1-c_1n^{-12}-me^{-1.5n}-3mn^{-12}$, where c_1 is some absolute positive constant.*

6 Conclusions

In this paper, we show that low-rank positive semidefinite matrices can be recovered from a near-minimal number of random rank-one measurements, via the vanilla gradient descent algorithm following spectral initialization. Our results significantly improve upon existing ones in terms of both computational and statistical complexities. In particular, our algorithm does not require resampling at every iteration (and hence requires fewer samples). The gradient iteration can provably employ a much more aggressive step size than what was suggested in prior literature (e.g. Sanghavi et al. (2017)), thus resulting in a much smaller iteration complexity and hence lower computational cost. All of these are enabled by establishing the implicit regularization feature of gradient descent for nonconvex statistical estimation, where the iterates remain incoherent with the sensing vectors throughout the execution of the whole algorithm.

There are several problems that are worth exploring in the future. For example, our theory reveals the typical size of the fitting error of $\boldsymbol{X}_t$ (i.e. $y_i-\|\boldsymbol{a}_i^\top\boldsymbol{X}_t\|_2$) in the absence of noise, which would serve as a helpful benchmark when separating sparse outliers in the more realistic scenario. Another direction is to explore whether implicit regularization remains valid for learning shallow neural networks (Zhong et al., 2017). Since the current work can be viewed as learning a one-hidden-layer fully-connected network with quadratic activation functions, it would be of great interest to study if the techniques utilized herein can be used to develop strong guarantees when the activation function takes other forms. Finally, it will be worthwhile to investigate the performance of vanilla gradient descent under random initializations (Chen et al., 2018).

Acknowledgements

The work of Y. Li and Y. Chi is supported in part by AFOSR under the grant FA9550-15-1-0205, by ONR under the grant N00014-18-1-2142, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571 and CCF-1704245. The work of Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ARO grant W911NF-18-1-0303, and by the Princeton SEAS innovation award.

References

J.-F. Cai, T. Wang, and K. Wei. Spectral compressed sensing via projected gradient descent. *arXiv preprint arXiv:1707.09726*, 2017.

T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

T. T. Cai, X. Li, and Z. Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

E. J. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, 2017.

Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604, 2014.

Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-k ranking. *arXiv preprint arXiv:1707.09971*, 2017.

Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *arXiv preprint arXiv:1803.07726*, 2018.

Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

L. Demanet and P. Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2014.

S. S. Du and J. D. Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *CoRR*, abs/1606.04933, 2016. URL `http://arxiv.org/abs/1606.04933`.

Y. Li, Y. Sun, and Y. Chi. Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Transactions on Signal Processing*, 65(2):397–408, 2017.

M. Lin and J. Ye. A non-convex one-pass framework for generalized factorization machine and rank-one matrix sensing. In *Advances in Neural Information Processing Systems*, pages 1633–1641, 2016.

R. Livni, S. Shalev-Shwartz, and O. Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

S. Sanghavi, R. Ward, and C. D. White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71(3-4):569–608, 2017.

M. Soltani and C. Hegde. Towards provable learning of polynomial neural networks using low-rank matrix estimation. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018.

M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

L. Tian, J. Lee, S. B. Oh, and G. Barbastathis. Experimental compressive phase space tomography. *Optics express*, 20 (8):8296–8308, 2012.

S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 964–973. JMLR. org, 2016.

I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 2017.

H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 2017.

K. Zhong, P. Jain, and I. S. Dhillon. Efficient matrix sensing using rank-1 Gaussian measurements. In *International Conference on Algorithmic Learning Theory*, pages 3–18. Springer, 2015.

K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 4140–4149, 2017.

Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. *arXiv preprint arXiv:1703.06605*, 2017.